# Contents

### 0.0.1 Why I Wrote "The Crucible," Arthur Miller (*New Yorker*), October 21, 1996

October 21, 1996

Arthur Miller (*New Yorker*)

In 1996, Miller wrote this essay for the *New Yorker*, in which he reflects on the changing politics surrounding his play "The Crucible," which he wrote in 1952, and which is now in revival on Broadway, at the Virginia Theatre.

As I watched "The Crucible" taking shape as a movie over much of the past year, the sheer depth of time that it represents for me kept returning to mind. As those powerful actors blossomed on the screen, and the children and the horses, the crowds and the wagons, I thought again about how I came to cook all this up nearly fifty years ago, in an America almost nobody I know seems to remember clearly. In a way, there is a biting irony in this film's having been made by a Hollywood studio, something unimaginable in the fifties. But there they are – Daniel Day-Lewis (John Proctor) scything his sea-bordered field, Joan Allen (Elizabeth) lying pregnant in the frigid jail, Winona Ryder (Abigail) stealing her minister-uncle's money, majestic Paul Scofield (Judge Danforth) and his righteous empathy with the Devil-possessed children, and all of them looking as inevitable as rain.

I remember those years – they formed "The Crucible" 's skeleton – but I have lost the dead weight of the fear I had then. Fear doesn't travel well; just as it can warp judgment, its absence can diminish memory's truth. What terrifies one generation is likely to bring only a puzzled smile to the next. I remember how in 1964, only twenty years after the war, Harold Clurman, the director of "Incident at Vichy," showed the cast a film of a Hitler speech, hoping to give them a sense of the Nazi period in which my play took place. They watched as Hitler, facing a vast stadium full of adoring people, went up on his toes in ecstasy, hands clasped under his chin, a sublimely self-gratified grin on his face, his body swiveling rather cutely, and they giggled at his overacting.

Likewise, films of Senator Joseph McCarthy are rather unsettling – if you remember the fear he once spread. Buzzing his truculent sidewalk brawler's snarl through the hairs in his nose, squinting through his cat's eyes and

sneering like a villain, he comes across now as nearly comical, a self-aware performer keeping a straight face as he does his juicy threat-shtick.

McCarthy's power to stir fears of creeping Communism was not entirely based on illusion, of course; the paranoid, real or pretended, always secretes its pearl around a grain of fact. From being our wartime ally, the Soviet Union rapidly became an expanding empire. In 1949, Mao Zedong took power in China. Western Europe also seemed ready to become Red – especially Italy, where the Communist Party was the largest outside Russia, and was growing. Capitalism, in the opinion of many, myself included, had nothing more to say, its final poisoned bloom having been Italian and German Fascism. McCarthy – brash and ill-mannered but to many authentic and true – boiled it all down to what anyone could understand: we had "lost China" and would soon lose Europe as well, because the State Department – staffed, of course, under Democratic Presidents – was full of treasonous pro-Soviet intellectuals. It was as simple as that.

If our losing China seemed the equivalent of a flea's losing an elephant, it was still a phrase – and a conviction – that one did not dare to question; to do so was to risk drawing suspicion on oneself. Indeed, the State Department proceeded to hound and fire the officers who knew China, its language, and its opaque culture – a move that suggested the practitioners of sympathetic magic who wring the neck of a doll in order to make a distant enemy's head drop off. There was magic all around; the politics of alien conspiracy soon dominated political discourse and bid fair to wipe out any other issue. How could one deal with such enormities in a play?

"The Crucible" was an act of desperation. Much of my desperation branched out, I suppose, from a typical Depression-era trauma – the blow struck on the mind by the rise of European Fascism and the brutal anti-Semitism it had brought to power. But by 1950, when I began to think of writing about the hunt for Reds in America, I was motivated in some great part by the paralysis that had set in among many liberals who, despite their discomfort with the inquisitors' violations of civil rights, were fearful, and with good reason, of being identified as covert Communists if they should protest too strongly.

In any play, however trivial, there has to be a still point of moral reference

against which to gauge the action. In our lives, in the late nineteen-forties and early nineteen-fifties, no such point existed anymore. The left could not look straight at the Soviet Union's abrogations of human rights. The anti-Communist liberals could not acknowledge the violations of those rights by congressional committees. The far right, meanwhile, was licking up all the cream. The days of "J'accuse" were gone, for anyone needs to feel right to declare someone else wrong. Gradually, all the old political and moral reality had melted like a Dali watch. Nobody but a fanatic, it seemed, could really say all that he believed.

President Truman was among the first to have to deal with the dilemma, and his way of resolving it – of having to trim his sails before the howling gale on the right – turned out to be momentous. At first, he was outraged at the allegation of widespread Communist infiltration of the government and called the charge of "coddling Communists" a red herring dragged in by the Republicans to bring down the Democrats. But such was the gathering power of raw belief in the great Soviet plot that Truman soon felt it necessary to institute loyalty boards of his own.

The Red hunt, led by the House Committee on Un-American Activities and by McCarthy, was becoming the dominating fixation of the American psyche. It reached Hollywood when the studios, after first resisting, agreed to submit artists' names to the House Committee for "clearing" before employing them. This unleashed a veritable holy terror among actors, directors, and others, from Party members to those who had had the merest brush with a front organization.

The Soviet plot was the hub of a great wheel of causation; the plot justified the crushing of all nuance, all the shadings that a realistic judgment of reality requires. Even worse was the feeling that our sensitivity to this onslaught on our liberties was passing from us – indeed, from me. In "Timebends," my autobiography, I recalled the time I'd written a screenplay ("The Hook") about union corruption on the Brooklyn waterfront. Harry Cohn, the head of Columbia Pictures, did something that would once have been considered unthinkable: he showed my script to the F.B.I. Cohn then asked me to take the gangsters in my script, who were threatening and murdering their opponents, and simply change them to Communists. When I declined to commit

this idiocy (Joe Ryan, the head of the longshoremen's union, was soon to go to Sing Sing for racketeering), I got a wire from Cohn saying, "The minute we try to make the script pro-American you pull out." By then – it was 1951 – I had come to accept this terribly serious insanity as routine, but there was an element of the marvelous in it which I longed to put on the stage.

In those years, our thought processes were becoming so magical, so paranoid, that to imagine writing a play about this environment was like trying to pick one's teeth with a ball of wool: I lacked the tools to illuminate miasma. Yet I kept being drawn back to it.

I had read about the witchcraft trials in college, but it was not until I read a book published in 1867 – a two-volume, thousand-page study by Charles W. Upham, who was then the mayor of Salem – that I knew I had to write about the period. Upham had not only written a broad and thorough investigation of what was even then an almost lost chapter of Salem's past but opened up to me the details of personal relationships among many participants in the tragedy.

I visited Salem for the first time on a dismal spring day in 1952; it was a sidetracked town then, with abandoned factories and vacant stores. In the gloomy courthouse there I read the transcripts of the witchcraft trials of 1692, as taken down in a primitive shorthand by ministers who were spelling each other. But there was one entry in Upham in which the thousands of pieces I had come across were jogged into place. It was from a report written by the Reverend Samuel Parris, who was one of the chief instigators of the witch-hunt. "During the examination of Elizabeth Procter, Abigail Williams and Ann Putnam" – the two were "afflicted" teen-age accusers, and Abigail was Parris's niece – "both made offer to strike at said Procter; but when Abigail's hand came near, it opened, whereas it was made up into a fist before, and came down exceeding lightly as it drew near to said Procter, and at length, with open and extended fingers, touched Procter's hood very lightly. Immediately Abigail cried out her fingers, her fingers, her fingers burned ... "

In this remarkably observed gesture of a troubled young girl, I believed, a play became possible. Elizabeth Proctor had been the orphaned Abigail's mistress, and they had lived together in the same small house until Elizabeth

fired the girl. By this time, I was sure, John Proctor had bedded Abigail, who had to be dismissed most likely to appease Elizabeth. There was bad blood between the two women now. That Abigail started, in effect, to condemn Elizabeth to death with her touch, then stopped her hand, then went through with it, was quite suddenly the human center of all this turmoil.

All this I understood. I had not approached the witchcraft out of nowhere, or from purely social and political considerations. My own marriage of twelve years was teetering and I knew more than I wished to know about where the blame lay. That John Proctor the sinner might overturn his paralyzing personal guilt and become the most forthright voice against the madness around him was a reassurance to me, and, I suppose, an inspiration: it demonstrated that a clear moral outcry could still spring even from an ambiguously unblemished soul. Moving crab-wise across the profusion of evidence, I sensed that I had at last found something of myself in it, and a play began to accumulate around this man.

But as the dramatic form became visible, one problem remained unyielding: so many practices of the Salem trials were similar to those employed by the congressional committees that I could easily be accused of skewing history for a mere partisan purpose. Inevitably, it was no sooner known that my new play was about Salem than I had to confront the charge that such an analogy was specious – that there never were any witches but there certainly are Communists. In the seventeenth century, however, the existence of witches was never questioned by the loftiest minds in Europe and America; and even lawyers of the highest eminence, like Sir Edward Coke, a veritable hero of liberty for defending the common law against the king's arbitrary power, believed that witches had to be prosecuted mercilessly. Of course, there were no Communists in 1692, but it was literally worth your life to deny witches or their powers, given the exhortation in the Bible, "Thou shalt not suffer a witch to live." There had to be witches in the world or the Bible lied. Indeed, the very structure of evil depended on Lucifer's plotting against God. (And the irony is that klatsches of Luciferians exist all over the country today, there may even be more of them now than there are Communists.)

As with most humans, panic sleeps in one unlighted corner of my soul. When I walked at night along the empty, wet streets of Salem in the week

that I spent there, I could easily work myself into imagining my terror before a gaggle of young girls flying down the road screaming that somebody's "familiar spirit" was chasing them. This anxiety-laden leap backward over nearly three centuries may have been helped along by a particular Upham footnote. At a certain point, the high court of the province made the fatal decision to admit, for the first time, the use of "spectral evidence" as proof of guilt. Spectral evidence, so aptly named, meant that if I swore that you had sent out your "familiar spirit" to choke, tickle, or poison me or my cattle, or to control my thoughts and actions, I could get you hanged unless you confessed to having had contact with the Devil. After all, only the Devil could lend such powers of invisible transport to confederates, in his everlasting plot to bring down Christianity.

Naturally, the best proof of the sincerity of your confession was your naming others whom you had seen in the Devil's company – an invitation to private vengeance, but made official by the seal of the theocratic state. It was as though the court had grown tired of thinking and had invited in the instincts: spectral evidence – that poisoned cloud of paranoid fantasy – made a kind of lunatic sense to them, as it did in plot-ridden 1952, when so often the question was not the acts of an accused but the thoughts and intentions in his alienated mind.

The breathtaking circularity of the process had a kind of poetic tightness. Not everybody was accused, after all, so there must be some reason why you were. By denying that there is any reason whatsoever for you to be accused, you are implying, by virtue of a surprisingly small logical leap, that mere chance picked you out, which in turn implies that the Devil might not really be at work in the village or, God forbid, even exist. Therefore, the investigation itself is either mistaken or a fraud. You would have to be a crypto-Luciferian to say that – not a great idea if you wanted to go back to your farm.

The more I read into the Salem panic, the more it touched off corresponding images of common experiences in the fifties: the old friend of a blacklisted person crossing the street to avoid being seen talking to him; the overnight conversions of former leftists into born-again patriots; and so on. Apparently, certain processes are universal. When Gentiles in Hitler's Germany, for ex-

ample, saw their Jewish neighbors being trucked off, or farmers in Soviet Ukraine saw the Kulaks vanishing before their eyes, the common reaction, even among those unsympathetic to Nazism or Communism, was quite naturally to turn away in fear of being identified with the condemned. As I learned from non-Jewish refugees, however, there was often a despairing pity mixed with "Well, they must have done something." Few of us can easily surrender our belief that society must somehow make sense. The thought that the state has lost its mind and is punishing so many innocent people is intolerable. And so the evidence has to be internally denied.

I was also drawn into writing "The Crucible" by the chance it gave me to use a new language – that of seventeenth-century New England. That plain, craggy English was liberating in a strangely sensuous way, with its swings from an almost legalistic precision to a wonderful metaphoric richness. "The Lord doth terrible things amongst us, by lengthening the chain of the roaring lion in an extraordinary manner, so that the Devil is come down in great wrath," Deodat Lawson, one of the great witchhunting preachers, said in a sermon. Lawson rallied his congregation for what was to be nothing less than a religious war against the Evil One – "Arm, arm, arm!" – and his concealed anti-Christian accomplices.

But it was not yet my language, and among other strategies to make it mine I enlisted the help of a former University of Michigan classmate, the Greek-American scholar and poet Kimon Friar. (He later translated Kazantzakis.) The problem was not to imitate the archaic speech but to try to create a new echo of it which would flow freely off American actors' tongues. As in the film, nearly fifty years later, the actors in the first production grabbed the language and ran with it as happily as if it were their customary speech.

"The Crucible" took me about a year to write. With its five sets and a cast of twenty-one, it never occurred to me that it would take a brave man to produce it on Broadway, especially given the prevailing climate, but Kermit Bloomgarden never faltered. Well before the play opened, a strange tension had begun to build. Only two years earlier, the "Death of a Salesman" touring company had played to a thin crowd in Peoria, Illinois, having been boycotted nearly to death by the American Legion and the Jaycees. Before

that, the Catholic War Veterans had prevailed upon the Army not to allow its theatrical groups to perform, first, "All My Sons," and then any play of mine, in occupied Europe. The Dramatists Guild refused to protest attacks on a new play by Sean O'Casey, a self-declared Communist, which forced its producer to cancel his option. I knew of two suicides by actors depressed by upcoming investigation, and every day seemed to bring news of people exiling themselves to Europe: Charlie Chaplin, the director Joseph Losey, Jules Dassin, the harmonica virtuoso Larry Adler, Donald Ogden Stewart, one of the most sought-after screenwriters in Hollywood, and Sam Wanamaker, who would lead the successful campaign to rebuild the Old Globe Theatre on the Thames.

On opening night, January 22, 1953, I knew that the atmosphere would be pretty hostile. The coldness of the crowd was not a surprise; Broadway audiences were not famous for loving history lessons, which is what they made of the play. It seems to me entirely appropriate that on the day the play opened, a newspaper headline read "ALL 13 REDS GUILTY" – a story about American Communists who faced prison for "conspiring to teach and advocate the duty and necessity of forcible overthrow of government." Meanwhile, the remoteness of the production was guaranteed by the director, Jed Harris, who insisted that this was a classic requiring the actors to face front, never each other. The critics were not swept away. "Arthur Miller is a problem playwright in both senses of the word," wrote Walter Kerr of the *Herald Tribune*, who called the play "a step backward into mechanical parable." *The Times* was not much kinder, saying, "There is too much excitement and not enough emotion in 'The Crucible.'" But the play's future would turn out quite differently.

About a year later, a new production, one with younger, less accomplished actors, working in the Martinique Hotel ballroom, played with the fervor that the script and the times required, and "The Crucible" became a hit. The play stumbled into history, and today, I am told, it is one of the most heavily demanded trade-fiction paperbacks in this country; the Bantam and Penguin editions have sold more than six million copies. I don't think there has been a week in the past forty-odd years when it hasn't been on a stage somewhere in the world. Nor is the new screen version the first. Jean-Paul Sartre, in

his Marxist phase, wrote a French film adaptation that blamed the tragedy on the rich landowners conspiring to persecute the poor. (In truth, most of those who were hanged in Salem were people of substance, and two or three were very large landowners.)

It is only a slight exaggeration to say that, especially in Latin America, "The Crucible" starts getting produced wherever a political coup appears imminent, or a dictatorial regime has just been overthrown. From Argentina to Chile to Greece, Czechoslovakia, China, and a dozen other places, the play seems to present the same primeval structure of human sacrifice to the furies of fanaticism and paranoia that goes on repeating itself forever as though imbedded in the brain of social man.

I am not sure what "The Crucible" is telling people now, but I know that its paranoid center is still pumping out the same darkly attractive warning that it did in the fifties. For some, the play seems to be about the dilemma of relying on the testimony of small children accusing adults of sexual abuse, something I'd not have dreamed of forty years ago. For others, it may simply be a fascination with the outbreak of paranoia that suffuses the play – the blind panic that, in our age, often seems to sit at the dim edges of consciousness. Certainly its political implications are the central issue for many people; the Salem interrogations turn out to be eerily exact models of those yet to come in Stalin's Russia, Pinochet's Chile, Mao's China, and other regimes. (Nien Cheng, the author of "Life and Death in Shanghai," has told me that she could hardly believe that a non-Chinese – someone who had not experienced the Cultural Revolution – had written the play.) But below its concerns with justice the play evokes a lethal brew of illicit sexuality, fear of the supernatural, and political manipulation, a combination not unfamiliar these days. The film, by reaching the broad American audience as no play ever can, may well unearth still other connections to those buried public terrors that Salem first announced on this continent.

One thing more – something wonderful in the old sense of that word. I recall the weeks I spent reading testimony by the tome, commentaries, broadsides, confessions, and accusations. And always the crucial damning event was the signing of one's name in "the Devil's book." This Faustian agreement to hand over one's soul to the dreaded Lord of Darkness was the

ultimate insult to God. But what were these new inductees supposed to have *done* once they'd signed on? Nobody seems even to have thought to ask. But, of course, actions are as irrelevant during cultural and religious wars as they are in nightmares. The thing at issue is buried intentions – the secret allegiances of the alienated heart, always the main threat to the theocratic mind, as well as its immemorial quarry

### 0.0.2 The Trauma Trap, Frederick C. Crews (*New York Review of Books*), March 11, 2004

March 11, 2004

Frederick C. Crews (*New York Review of Books*)

*Remembering Trauma*
by Richard J. McNally
Belknap/Harvard University Press, 420 pp., $35

*Memory, Trauma Treatment, and the Law*
by Daniel Brown, Alan W. Scheflin, and D. Corydon Hammond
W. W. Norton, 768 pp., $100.00

Every now and then a book appears that can be instantly recognized as essential for its field – a work that must become standard reading if that field is to be purged of needless confusion and fortified against future errors of the same general kind. Such a book is *Remembering Trauma*, by the Harvard psychology professor Richard J. McNally. To be sure, the author's intention is not revolutionary but only consolidating; he wants to show what has already been learned, through well-designed experiments and analyses of records, about the effects that psychological trauma typically exerts on our memory. But what has been learned is not what is widely believed, and McNally is obliged to clear away a heap of junk theory. In doing so, he provides a brilliant object lesson in the exercise of rational standards that are common to every science deserving of the name.

McNally's title *Remembering Trauma* neatly encapsulates the opposing views that, for a whole generation now, have made the study of trauma into psychology's most fiercely contested ground. Are scarring experiences well remembered in the usual sense of the term, or can some of them be remembered only much later, after the grip of a self-protective psychological mechanism has been relaxed? This is the pivotal issue that McNally decisively resolves. In the process, he also sheds light on a number of related questions. Does memory of trauma stand apart neurologically from normal memory? Does a certain kind of traumatic experience leave recognizable long-term effects that can vouch for its historical reality? What memory problems

typify post-traumatic stress disorder, and does the disorder itself "occur in nature" or is it a cultural construct? And is memory retrieval a well-tested and effective means of helping adults to shed depression, anxiety, and other psychological afflictions?

One extended trauma, a public one, that won't be soon forgotten by the involved parties is central to McNally's argument. I refer to the great sex panic that gripped this continent from about 1985 to 1994. It wasn't just an epidemic of runaway fear, rumor, and persecution but a grimly practical test of the theories whose currency made it possible. And the theories at issue were precisely those that are exhaustively reviewed in *Remembering Trauma*. McNally uses that chapter of our history to show just how much damage can be done when mistaken ideas about the mind get infused with ideological zeal.

In the 1980s, as McNally relates, day care workers risked prosecution and imprisonment on the coerced testimony of bewildered and intimidated three-year-olds who were prodded to "remember" nonexistent molestations. Meanwhile, poorly trained social workers, reasoning that signs of sexual curiosity in children must be "behavioral memories" of rape, were charging parents with incest and consigning their stunned offspring to foster homes. And most remarkably, whole communities were frantically attempting to expose envisioned covens of Satan worshipers who were said, largely on the basis of hypnotically unlocked "memories," to be raising babies for sexual torture, ritual murder, and cannibal feasts around the patio grill.

In the same period many psychotherapists, employing hypnosis, dream analysis, "guided imagery," "age regression," and other suggestion-amplifying devices, persuaded their mostly female patients to "remember" having been molested by their fathers or stepfathers through much of their childhood, in some cases with the active participation of their mothers. The "perpetrators" thus fingered were devastated, embittered, and often publicly shamed, and only a minority of their accusers eventually recanted. Many, in fact, fell in with their therapists' belief that young victims of sexual trauma, instead of consciously recalling what was done to them, are likely to develop multiple personalities. Disintegrating further, those unfortunates were then sent off to costly "dissociative identity" wards, where their fantasies of containing five,

a dozen, or even hundreds of inner selves were humored until their insurance coverage expired and they were abandoned in a crazed condition. At the height of the scare, influential traumatologists were opining that "between twenty and fifty percent of psychiatric patients suffer from dissociative disorders" [1] – disorders whose reported incidence plummeted toward zero as soon as some of the quacks who had promoted them began to be sued for malpractice [2].

What we experienced, McNally shows, was a perfect storm, with forces for mischief converging from every side. The fraudulent 1973 bestseller *Sybil* had already helped to relaunch the long-dormant fad of multiple personality and to link it to childhood sexual abuse [3]. Beginning in the early 1980s, the maverick Swiss psychoanalyst Alice Miller taught many American readers what Sigmund Freud had once believed, that memories of early abuse are typically repressed and must be therapeutically unlocked if the resultant neuroses are to be cured. Jeffrey Masson's melodramatic book *The Assault on Truth* (1984), misrepresenting Freud's "seduction" patients as self-aware incest victims rather than as the doubters that they remained, fanned the feminist anger that Miller had aroused, encouraging women to believe that molestation by fathers must be pervasive [4]. Self-help manuals such as *The Courage to Heal* (1988) then equipped scientifically ignorant psychotherapists with open-ended "symptom checklists," ensuring that their patients would be diagnosed as suffering from buried memories of violation. And all the while, Geraldo Rivera and less cynical alarmists were whipping up fear of murderous devil cults.

If the origins of our mass delusion were complex, its dissipation in the mid-1990s is easily explained. Like the Salem witch hunt three centuries earlier, the sex panic had no internal brake that could prevent its accusations from racing beyond all bounds of credibility. The stirring motto "Believe the children" began to sound hollow when preschoolers who finally agreed that they must have been inappropriately touched went on to describe having been dropped into a pool of sharks or turned into a mouse. The medical records of some alleged rape victims showed that they had still been virgins at a later period. In one notorious case, influential at first in promoting recovered memory but later in discrediting it, a woman who got her father sentenced to

life in prison for a murder/rape she had remembered in hypnotic trances went on to recall his killing of another person who proved to be wholly imaginary. And many patients, when urged to dig deeper after producing a vague scene or two, reduced the process to self-travesty by conjuring surreal orgies with Daddy's bridge partners, visiting uncles, and the family pets.

One recovered memory case in particular, less absurd than most but nevertheless lacking in *prima facie* plausibility, set in motion what the movement's loyalists now bitterly characterize as "the backlash." In 1991 the future "betrayal trauma" psychologist Jennifer J. Freyd, after her therapist had pointedly asked her in their second encounter whether she had ever been abused, suddenly "remembered" that her father had continually molested her between the ages of three and sixteen. It was Freyd's mother, Pamela, convinced that she would surely have noticed some effects of countless domestic sex crimes against her daughter, who then made contact with other recently accused parents and established the False Memory Syndrome Foundation. Under Pamela Freyd's leadership, the foundation (on whose advisory board I serve) gathered and disseminated the most authoritative scientific judgments about trauma, memory, and suggestive influence – judgments that swayed enough jurists, legislators, and journalists to bring a healthy skepticism into play.

What put Jennifer Freyd's "memories" in question wasn't just their dissonance with her mother's close observation. By alleging fourteen years' worth of molestations that had been unknown to her conscious mind prior to a therapist's prompting, Freyd was invoking an outlandish new defense mechanism. Granted, some psychologists still believed in repression, or the sequestering of a disagreeable thought or memory inside "the unconscious"; and others subscribed to dissociation, the more radical knack of "splitting the self" so quickly that no narrative memory of the trauma gets formed at all. But Freyd's story, like many others that surfaced during the sex panic, stretched those principles to cover any number of serial traumatic incidents, as if a person could be subjected to the same outrage hundreds of times without taking cognitive note of it.

This cumulative forgetting of harmful experience is what the social psychologist Richard Ofshe disdainfully named robust repression – a startlingly

maladaptive behavior that, if actual, ought to have aroused wonder and consternation from the earliest times until now, if indeed it didn't lead to the extinction of our species. Before the American 1980s, however, it had apparently never once been remarked. Could robust repression itself have been robustly repressed throughout the millennia?

Most recovered memory advocates have ducked the conundrum of robust repression, and some have dismissed it as an alien notion devised by their adversaries. But the alleged phenomenon, McNally shows, is nothing other than the "massive repression" posited by such prominent traumatologists as Judith Lewis Herman, Judith L. Alpert, Lenore C. Terr, and Jennifer J. Freyd herself, each of whom understood that claims of sudden access to a long string of previously unsuspected horrors require a basis in theory. What could that basis be? McNally makes short work of the only systematic attempts, Terr's and Freyd's, to maintain that serial traumas are easier to forget than single ones. Moreover, all such efforts are doomed to be question begging, because the only evidence favoring robust repression consists of the very memories whose authenticity hangs in doubt.

The same stricture applies, however, to repression and dissociation per se. Those notions became current in the 1880s and 1890s when Freud and Pierre Janet independently attempted to trace the then fashionable complaint of hysteria to pathogenic hidden memories and to expunge the ailment through hypnotically induced recall. Freud, by far the more influential figure, clung to repression – though rendering it progressively more elastic and ambiguous – even while repeatedly distancing himself from the diagnostic and curative claims he had inferred from its supposed workings.

Before he was finished, Freud had conceived of repression as both a conscious and an unconscious process acting upon feelings, thoughts, ideas, and fantasies as well as memories. Such profligacy left repression without any operational meaning; "the repressed" was simply any material that Freud, who was given to ascribing his own punning associations to his patients' minds, chose to identify as having been dismissed from awareness. Yet the long vogue of psychoanalysis kept the concept alive, enabling it to be virulently readapted, a century after its formal introduction, to the same task of recruiting patients to victimhood that had preoccupied its champion in

1895–96.

As McNally explains through deftly analyzed examples, it isn't just therapists and their patients who fail to ask prudent questions about the repression or dissociation of trauma. The body of research purporting to validate those mechanisms is riddled with procedural errors, most of which stem from naive trust in the retrospection of subjects who have already been led to believe that they must have undergone a trauma that was then sequestered from memory. Along with such other inquirers as David Holmes and Harrison G. Pope, Jr., McNally understands that a good test of repression or dissociation has to be prospective. That is, it must track down people who are known with certainty to have lived through ordeals that would be expected to have triggered a self-protective loss of memory, and it must then ascertain how many of those people are unable to recall the event.

Holocaust survivors make up the most famous class of such subjects, but whatever group or trauma is chosen, the upshot of well-conducted research is always the same. Like Holmes and Pope, McNally finds that no unanswerable evidence has been adduced to prove that anyone, anywhere, has ever repressed or dissociated the memory of any occurrence. Traumatic experiences may not always remain in the forefront of memory, but, unlike "repressed" ones, they can be readily called to mind again. Unless a victim received a physical shock to the brain or was so starved or sleep deprived as to be thoroughly disoriented at the time, those experiences are typically better remembered than ordinary ones. Thus Judith Herman's much-quoted maxim, "The ordinary response to atrocities is to banish them from consciousness," [5] would appear to be exactly opposite to the truth. And once that fact is understood, the improvised and precarious edifice of recovered memory theory collapses into rubble.

It would be a serious mistake, however, to assume that reckless traumatology has now been permanently laid to rest. The conviction that fathers are naturally prone to incestuous rape is still current. In some academic departments, a dogged literalism about the repression/dissociation of trauma has become oddly wedded to postmodernist suspicion of science [6]. Furthermore, most of the "trauma centers" that sprang up in the 1990s to study and treat psychogenic amnesia are still operating under the same premises as before.

As for the theoreticians of recovered memory, they continue to use their positions of authority in universities, hospitals, and professional organizations to advance the views whose hollowness McNally has exposed, and they can still count on a surprising level of support from their colleagues.

Consider, in this regard, the following example of deafness to the lessons of the sex panic. Each year the American Psychiatric Association, the body that sets the most basic guidelines for sound practice in our mental health professions, bestows its Manfred S. Guttmacher Award on what it deems to be the best recent publication on legal psychiatry. The prize for 1999 went to a 768-page tome by Daniel Brown, Alan W. Scheflin, and D. Corydon Hammond, *Memory, Trauma Treatment, and the Law*. The authors characterize themselves as "voices of moderation in the middle" opposing "zealots on both sides" (p. 1). Their book, however, consists largely of sophistical pleading for already lost causes: the forensic value of therapeutically retrieved memories, the genuineness of multiple personality disorder, the likelihood that some reports of ritual abuse cults are accurate, and the desirability of allowing evidence obtained through hypnosis to be admissible in court.

*Memory, Trauma Treatment, and the Law* isn't just a disingenuous book, hiding its partisanship behind a screen of sanctimony; it is also a noxious one. Lightly granting the possibility that therapy may occasionally lead to pseudomemories, it trivializes the problem, deeming it serious only "when the patient takes legal action or publically [sic] discloses abuse" (p. 37) – as if the suffering of privately shattered families counted for nothing. And the book's strategy of superficially "reviewing the literature," citing both skeptical and (always more numerous) credulous studies and then tilting the scales toward the latter, merely simulates scientific neutrality.

These authors' activism in the cause of recovered memory was well known long before they collaborated on their prize-winning volume. Daniel Brown and Alan Scheflin had often served as expert witnesses minimizing the hazards of memory retrieval, claiming to have found overwhelming experimental support for the concept of repression, and denying that a therapist could ever deceive a patient into thinking that she suffered from multiple personality; and their collaborative papers were similarly one-sided [7]. In 1995, moreover, Scheflin had delivered a warmly received address to a Texas conference held

by the Society for the Investigation, Treatment and Prevention of Ritual and Cult Abuse, whose other speakers asserted, inter alia, that there were 500 Satanic cults in New York City alone, committing 4000 human sacrifices per year, that Bill Clinton was serving as the Antichrist in the worldwide Satanic fraternity of the Illuminati and that the False Memory Syndrome Foundation is "a Central Intelligence Agency action." Expressing solidarity with the assembled psychotherapists whose diagnoses of ritual abuse were exposing them to malpractice suits, Scheflin counseled them on the best means of foiling the legal machinations of "the false memory people," whom he characterizes as "the enemy." [8]

But it is hypnotherapist D. Corydon Hammond, well known for his low regard for experimental research on memory [9], whose name on the title page of *Memory, Trauma Treatment, and the Law* ought to have prompted especial wariness among the Guttmacher judges. Like Scheflin, Hammond has affirmed the reality of both Satanic abuse cults and multiple personality disorder. But whereas Scheflin stops short of asserting a proven link between those two phenomena, Hammond is on record as a flamboyant true believer.

In a notorious 1992 lecture at a conference on sexual abuse and MPD, Hammond revealed his conviction that many MPD sufferers have acquired their split personalities through subjection, from early childhood onward, to ritual sexual abuse, sadistic torture, and mind control programming. The aim of the programmers, he disclosed, has been to produce remotely guided "alters" who, unbeknownst to their core selves, will be slaves to a worldwide intergenerational cult that is organized into "Illuminatic councils." The cult, said Hammond, is headed by a shadowy "Dr. Greenbaum," a Hasidic Jewish collaborator with the Nazis who once assisted in death camp experiments and later used the CIA to further his nefarious ends. "My best guess," Hammond confided, ... "is that they want an army of Manchurian Candidates, tens of thousands of mental robots who will do prostitution, do child pornography, smuggle drugs, engage in international arms smuggling, do snuff films," ... and eventually the megalomaniacs at the top believe they'll create a Satanic order that will rule the world [10].

These colorful fantasies are significant, but not because they point to a failure of reality testing on Hammond's part. Closely related ideas were

voiced in the heyday of the recovered memory movement by other prominent MPD specialists such as Bennett Braun and Colin Ross. What matters is that Hammond and the others all claim to have learned about the grand cabal from their hypnotized patients, who, until they were placed in trances, hadn't even known they were molestation victims, much less robotic smugglers, whores, and assassins [11]. As Brown, Scheflin, and Hammond now put it in arguing in favor of hypnotically obtained evidence in the courtroom, "for some victims, hypnosis may provide the only avenue to the repressed memories" (p. 647). Exactly. Without that means of exchanging and embroidering false beliefs, Hammond himself could never have learned from his patients about the evil Dr. Greenbaum and his thirst for absolute power over us all.

The illogicalities and distortions in *Memory, Trauma Treatment, and the Law* do not go unremarked in McNally's *Remembering Trauma*. Thus, when Brown et al. cite one study as evidence that "amnesia for Nazi Holocaust camp experiences has also been reported," McNally quotes that study's rather different conclusion: "There is no doubt that almost all witnesses remember Camp Erika in great detail, even after 40 years" (p. 192). And when Brown et al., again straining to make psychologically motivated amnesia look commonplace, cite another study to the effect that "two of the 38 children studied after watching lightning strike and kill a playmate had no memory of the event," McNally informs us that those two children "had themselves been struck by side flashes from the main lightning bolt, knocked unconscious, and nearly killed" (p. 192).

Such corrections, however damning, are peripheral to McNally's fundamental critique of Brown and his colleagues. The heart of the matter is that Brown et al. have miscast the entire debate over recovered memory by marshaling evidence against a straw-man "extreme false memory position." Supposedly, the extremists hold that all refreshed memories of abuse are necessarily wrong. Then one could put the extremists in their place just by citing a few cases of authenticated recall. But as McNally shows, critics of recovered memory fully allow that a period of forgetfulness can precede a genuine recollection. Indeed, that pattern is just what we would expect if the young subject at the time of the act, never having been warned against sexual predators, was unsure how to regard that act. What the critics deny

is that "memories" of trauma, surfacing for the first time many years later, are so intrinsically reliable that they can serve as useful evidence that the experience was real. Brown, Scheflin, and Hammond want that extremism to be embraced once again by the legal system that has finally learned to distrust it.

It would be reassuring to think that the the American Psychiatric Association's Guttmacher jury merely skimmed *Memory, Trauma Treatment, and the Law* and misconstrued it as a bland eclectic survey. Already in 1991, however, another Guttmacher Award had been bestowed on co-author Scheflin for a work that made several of the same legal arguments [12]. A more likely explanation for the subsequent prize is that Brown et al., having mounted a brief for the deep knowledge and expert testimony of theory-minded clinicians, were gratefully perceived as siding with mental health providers against their adversaries. If so, a larger question comes into view. What role did our major societies representing psychotherapists – the American Psychoanalytic Association, the American Psychological Association, and the American Psychiatric Association itself – play in condoning or actually facilitating the recovered memory movement, and how much enlightened guidance can we expect from them in the future?

As I have noted on several occasions [13], and as McNally confirms, in the 1990s recovered memory therapy made significant inroads into the practice of North American psychoanalysis. Even today, feminist clinicians bearing diplomas from analytic institutes are probing for missing memories of abuse and vigorously defending that practice in psychoanalytic books and journals. But the American Psychoanalytic Association, representing over 3,000 members, has turned a blind eye to this trend – and one can understand why. The psychoanalytic movement is already embattled, and too much about the historical ties between Freudianism and recovered memory would prove embarrassing if attention were called to it. The elected custodians of Freud's legacy have no desire to confront his early phase as a self-deceived abuse detector; or to admit the precedent he set, during that phase and thereafter, in treating dreams, tics, obsessional acts, and agitation in the consulting room as "behavioral memories" of inferrable traumas; or to revisit the grave doubts that have been raised about repression; or to be reminded of the way psycho-

analysts, until quite recently, insulted real victims of molestation by telling them that their "screen memories" covered a repressed desire to have sex with their fathers [14]. No longer given to excommunicating dissidents, the tottering Freudian patriarchy has made its peace with "recovered memory psychoanalysis" by pretending that it doesn't exist.

The largest of the three societies riven by the issue of recovered memory, the 95,000-member American Psychological Association (hereafter APA), is nominally responsible for quality control in the administration of therapy by the nation's clinical psychologists. Hence one APA division's commendable effort in the 1990s to identify the most effective treatment methods for specific complaints such as phobias and obsessive-compulsive disorder. That initiative, however, met with disapproval from APA members whose favorite regimens had not been found to give superior results. Some practitioners worried that insurers would use the list of approved treatments as an excuse to cut off reimbursement for all but the preferred therapies, and others complained that the association seemed on the verge of putting soulless experimentation ahead of clinical know-how. For now at least, the organization as a whole is not recommending treatments, to say nothing of disavowing dangerous ones [15]. Recovered memory thus gets the same free pass from the APA as "attachment therapy," "therapeutic touch," "eye movement desensitization and reprocessing," "facilitated communication," and the hypnotic debriefing of reincarnated princesses and UFO abductees [16].

This reluctance to challenge the judgment of its therapist members is deeply rooted in the APA's philosophy. Ever since 1971, when the association gave its blessing to Ph.D. and Psy.D. programs that omitted any scientific training, the APA has guided its course by reference to studies indicating that the intuitive competence of clinicians, not their adherence to one psychological doctrine or another, is what chiefly determines their effectiveness [17]. Those studies, however, were conducted before recovered memory practitioners, using a mixture of peremptory guesswork and unsubstantiated theory, began wrenching patients away from their families and their remembered past.

In 1995 the APA did publish a brochure, "Questions and Answers about Memories of Childhood Abuse," which can still be found on the "APA On-

line" Web site. The document combined some prudent advice to patients with soothing reassurance that "the issue of repressed or suggested memories has been overreported and sensationalized." Further inquiry into the phenomenon, it said, "will profit from collaborative efforts among psychologists who specialize in memory research and those clinicians who specialize in working with trauma and abuse victims."

But the APA directors already knew that such collaboration was impossible. In 1993 they had established a "task force," the Working Group on the Investigation of Memories of Childhood Abuse, self-defeatingly composed of three research psychologists and three clinicians favorably disposed to retrieval, and the task force had immediately degenerated into caucusing and wrangling. After years of stalemate, the group predictably submitted two reports that clashed on every major point; and the abashed APA, presented with this vivid evidence that "clinical experience" can lead to scientific heterodoxy, declined to circulate photocopies of the two documents even to its own members except by individual demand.

Meanwhile, the organization repeatedly compromised its formal neutrality. In 1994, for example, the APA's publishing house lent its prestigious imprint to a book that not only recommended recovered memory therapy but recycled the most heedless advice found in pop-psychological manuals. The book, Lenore E. A. Walker's *Abused Women and Survivor Therapy: A Practical Guide for the Psychotherapist*, touted hypnotism as a legitimate means of gaining access to "buried memories of incest" and "different personalities" within the victim (pp. 425–426). Walker provided a list of telltale symptoms, any one of which might indicate a history of forgotten molestation. These included "ambivalent or conflict ridden relationships," "poor body image," "quiet-voiced," "inability to trust or indiscriminate trust," "high risk taking or inability to take risks," "fear of losing control and need for intense control," "great appreciation of small favors by others," "no sense of humor or constant wisecracking," and "blocking out early childhood years" (p. 113) – years which in fact are not remembered by anyone.

Then in 1996 the APA published and conspicuously endorsed another book, *Recovered Memories of Abuse*, aimed at equipping memory therapists and their expert witnesses with every argument and precaution that could

thwart malpractice suits [18]. The book's co-authors were well-known advocates of recovered memory treatment, and one of them, Laura S. Brown, was actually serving at the time on the deadlocked task force. She had also supplied a foreword to Lenore Walker's bumbling *Abused Women and Survivor Therapy*, calling it "invaluable and long overdue" (p. vii). Unsurprisingly, then, *Recovered Memories of Abuse* characterized false memory as an overrated problem and drew uncritically on much of the research whose weaknesses Richard McNally has now exposed. The APA's unabated promotion of that book, even today, suggests that the organization remains more concerned with shielding its most wayward members than with warning the public against therapeutic snake oil.

There remains, once again, the American Psychiatric Association – "the voice and conscience of modern psychiatry," as its Web site proclaims. Putting aside the fiasco of the 1999 Guttmacher Award, we might expect that a society representing 37,000 physicians, all of whom have been schooled in the standard of care that requires treatments to be tested for safety and effectiveness, would be especially vigilant against the dangers of retrieval therapy. Thus far, however, that expectation has not been fulfilled.

To be sure, the Psychiatric Association's 1993 "Statement on Memories of Sexual Abuse" did warn clinicians not to "exert pressure on patients to believe in events that may not have occurred ... " Yet the statement inadvertently encouraged just such tampering by avowing that the "coping mechanisms" of molested youngsters can "result in a lack of conscious awareness of the abuse" and by characterizing "dissociative disorders" as a typical outcome of that abuse. Those remarks constituted a discreet but unmistakable vote of confidence in multiple personality disorder and its imagined sexual etiology. And indeed, a year later the fourth edition of the Psychiatric Association's *Diagnostic and Statistical Manual of Mental Disorders (DSM-IV)* reaffirmed the validity of MPD under the more dignified and marketable name of dissociative identity disorder.

The Psychiatric Association's 1993 declaration on abuse memories performed still another service, a subtle one, for the repression/dissociation lobby. In explaining "implicit" memory – the kind that is exercised in the routine execution of skills or in the coloring of emotions by past impressions

that aren't being explicitly called to mind – the statement proffered a curiously strained example. "In the absence of explicit recall," it said, implicit memory can torment "a combat veteran who panics when he hears the sound of a helicopter, but cannot remember that he was in a helicopter crash which killed his best friend." Here was an elision of the crucial gap between merely not thinking about a past event, as in the normal operation of implicit memory, and having total, psychologically motivated amnesia for that event.

Knowledgeable readers would have seen that in taking this unusual step, the statement's drafters were lending their authority to one controversial interpretation of post-traumatic stress disorder (PTSD), which the Psychiatric Association had first stamped as genuine in DSM-III of 1980. But why should a primarily martial ailment have figured even indirectly in a position paper on childhood sexual abuse? The mystery vanishes, however, if we know that the recovered memory movement's favorite means of courting respectability has been to fold the symptoms of repressed/dissociated abuse into PTSD.

In 2000 the Psychiatric Association's trustees, eschewing risky flights into theory, approved a lower-profile "Position Statement on Therapies Focused on Memories of Childhood Physical and Sexual Abuse." This declaration, however, was more pussyfooting than its predecessor. The validity of recovered memory treatment, it whispered, "has been challenged" in some quarters. While pointing out that memories can be altered as a result of suggestions from "a trusted person or authority figure," the drafters tactfully refrained from mentioning that the suggesting party is usually a therapist. And clinicians were advised to avoid "prejudging the veracity of the patient's reports" of abuse, as if false reports were typically delivered to therapists out of the blue, without influence from confabulation-enhancing devices employed within the treatment. The absence of any mention of those devices, such as hypnosis and sodium amytal, marked a step backward from the association's 1993 statement.

These equivocations neither helped nor impeded the already withering recovered memory movement. As we will now see, however, the movement's hopes of a comeback have been pinned on the Psychiatric Association's fateful decision to treat post-traumatic stress disorder as an integral and historically invariable malady. And that decision was a medically unwarranted one. As

McNally indicates with reference to several recent studies, PTSD, like Victorian hysteria and like recovered memory itself, can now be understood as an artifact of its era – a sociopolitical invention of the post-Vietnam years, meant to replace "shell shock" and "combat fatigue" with an enduring affliction that would tacitly indict war itself as a psychological pathogen [19]. However crippling the symptoms associated with it may be for many individuals, the PTSD diagnosis itself has proved to be a modern contagion.

Once certified by the American Psychiatric Association as natural and beyond the sufferer's control, post-traumatic stress disorder began attracting claimants, both civilian and military, who schooled themselves in its listed symptoms and forged a new identity around remaining uncured. By now, as McNally relates, PTSD compensation is demanded for such complaints as "being fired from a job, one-mile-per-hour fender benders, age discrimination, living within a few miles of an explosion (although unaware that it had happened), and being kissed in public" (p. 281). According to Paula Jones among others, PTSD can even be the outcome of a consensual love affair. In view of such examples, the attempt to subsume forgotten abuse under post-traumatic stress makes more cultural than scientific sense; the same atmosphere of hypersensitivity and victimhood brought both diagnoses to life [20].

As McNally shows in his concise and undemonstrative style, the national sex panic left its mark on each successive version of the Psychiatric Association's bible, which in turn congealed folklore into dogma. The 1980 DSM-III entry on post-traumatic stress disorder, mindful only of wars and other shocking disasters, had defined a PTSD-triggering event as one that falls "generally outside the range of usual human experience" and that "would evoke significant symptoms of distress in almost everyone." In 1994, however, the fourth edition generously expanded the category of precipitating causes to include "developmentally inappropriate sexual experiences without threatened or actual violence or injury." Thus a single-minded therapeutic sleuth could now place a questionably retrieved incident of infantile genital fondling on the same etiological plane as the Bataan death march or an ambush in the Mekong Delta.

It was the diagnostic manual, once again, that removed the largest ob-

stacle of all to the merger of post-traumatic stress and recovered memory. The key sign of PTSD, as first conceived, was that accurate recollections of the trauma keep intruding on the patient's conscious mind; this was just the opposite of repressed or dissociated memory. But between DSM-III and its revised edition of 1987, PTSD patients were discovered to have been harboring a convenient new symptom. In 1980 they had shown only some incidental "memory impairment or trouble concentrating" on daily affairs, but the updated edition replaced routine forgetfulness with "inability to recall an important aspect of the trauma."

This retroactive infusion of amnesia into the clinical picture of PTSD explains why the Psychiatric Association's illustrative helicopter pilot could have been troubled by a memory that had left no conscious imprint on his mind. Here, too, was the opening needed to give dissociation an appearance of hard-scientific concreteness. Post-traumatic stress, it was now claimed, short-circuits narrative memory and finds another, precognitive, channel through which it can flood the subject with anxiety. Accordingly, diehard recovered memory theorists took up a last refuge in neurobiology, now maintaining that dissociated sexual abuse generates signature alterations of brain tissue.

With the arrival of McNally's *Remembering Trauma*, there is no longer any excuse for such obfuscation. It makes no sense, McNally shows, to count forgetfulness for some "aspect of the trauma" within the definition of PTSD, because normal people as well as PTSD sufferers get disoriented by shocking incidents and fail to memorize everything about the event, even while knowing for the rest of their lives that it occurred. Likewise, it has never been established, and it seems quite unbelievable, that people can be haunted by memories that were never cognitively registered as such. Nor can specific brain markers vouch for the reality of a long-past sexual trauma, because, among other reasons, those features could have been present from birth. "It is ironic," McNally reflects, "that so much has been written about the biological mechanisms of traumatic psychological amnesia when the very existence of the phenomenon is in doubt. What we have here is a set of theories in search of a phenomenon" (p. 182n.).

*Remembering Trauma* is neither a polemic nor a sermon, and McNally offers little counsel to psychotherapists beyond warning them against turn-

ing moral disapproval of pedophilia into overconfidence that they can infer its existence from behavioral clues observed twenty or thirty years after the fact. But another lesson is implied throughout this important book. Attention to the chimerical task of divining a patient's early traumas is attention subtracted from sensible help in the here and now. The reason why psychotherapists ought to familiarize themselves with actual knowledge about the workings of memory, and why their professional societies should stop waffling and promulgating misinformation about it, is not that good science guarantees good therapy; it is simply that pseudoscience inevitably leads to harm.

Notes:

[1] Bessel A. van der Kolk and Onno van der Hart, "The Intrusive Past: The Flexibility of Memory and the Engraving of Trauma," *American Imago*, vol. 48 (1991), pp. 425–454; the quotation is from p. 432.

[2] The fullest treatment of the recovered memory episode and its historical antecedents is Mark Pendergrast, *Victims of Memory: Sex Abuse Accusations and Shattered Lives*, 2nd ed. (Upper Access, 1996). For a concise and pointed account of the multiple personality fad, see Joan Acocella, *Creating Hysteria: Women and Multiple Personality Disorder* (Jossey-Bass, 1999). The best extended discussion is Nicholas P. Spanos, *Multiple Identities and False Memories: A Sociocognitive Perspective* (American Psychological Association, 1996). On Satanic abuse, see Jeffrey S. Victor, *Satanic Panic: The Creation of a Contemporary Legend* (Open Court, 1993), and Debbie Nathan and Michael Snedeker, *Satan's Silence: Ritual Abuse and the Making of a Modern American Witch Hunt* (Basic Books, 1995). The plight of daycare workers who remain imprisoned even today is treated by Dorothy Rabinowitz, *No Crueler Tyrannies: Accusation, False Witness, and Other Terrors of Our Times* (Wall Street Press Books/Free Press, 2003).

[3] For the current state of knowledge about "Sybil," see Mikkel Borch-Jacobsen, *Folie à plusieurs: De l'hystérie à la dépression* (Les Empêcheurs de penser en rond/Le Seuil, 2002), pp. 111–168.

[4] For Masson's errors about Freud's "seduction" phase, see Allen Esterson, "Jeffrey Masson and Freud's Seduction Theory: A New Fable Based on Old Myths," *History of the Human Sciences*, vol. 11 (1998), pp. 1–21. In

his preface to a recently reprinted edition of *The Assault on Truth* (Random House, 2003), Masson at last concedes that Freud's patients in 1895–6 resisted the incest stories that he tried to force upon them. Bizarrely, however, Masson still counts those patients among the likely victims of sexual abuse in Freud's day.

[5] Judith Lewis Herman, *Trauma and Recovery* (Basic Books, 1992), p. 1.

[6] See, in this connection, the final chapter of Ruth Leys's *Trauma: A Genealogy* (Univ. of Chicago Press, 2000).

[7] In one paper, for example, Scheflin and Brown addressed the problem of patients' suggestibility, but the danger they envisioned from that quarter was only "false litigant syndrome," or surrender to "pro-false-memory suggestive influences" emanating from "plaintiffs' attorneys and expert witnesses" brought into malpractice suits against their former therapists. See Alan W. Scheflin and Daniel Brown, "The False Litigant Syndrome: 'Nobody Would Say That Unless It Was the Truth,'" *Journal of Psychiatry and Law*, vol. 27 (1999), pp. 649–705. This same argument surfaces in *Memory, Trauma Treatment, and the Law*, which states that pressures exerted in therapy "pale in comparison" (p. 398) with those that can turn a patient into a litigious ingrate.

[8] Transcripts of the Texas conference proceedings have been available from Toronto radio station CKLN. See also Evan Harrington, "Conspiracy Theories and Paranoia: Notes from a Mind-Control Conference," *Skeptical Inquirer*, vol. 20 (September/October 1996), pp. 35–42.

[9] "I think it's time somebody called for an open season on academicians and researchers," Hammond said in 1997; " ... it's time for clinicians to begin bringing ethics charges for scientific malpractice against researchers and journal editors" who disparage recovered memory theory. "Investigating False Memory for the Unmemorable: A Critique of Experimental Hypnosis and Memory Research," 14th International Congress of Hypnosis and Psychosomatic Medicine, San Diego, June 1997. Tapes of Hammond's talk have been offered by The Sound of Knowledge, Inc.

[10] D. Corydon Hammond, "Hypnosis in MPD: Ritual Abuse," a paper delivered at the Fourth Annual Eastern Regional Conference on Abuse and

Multiple Personality, Alexandria, VA, June 25, 1992. Understandably, tapes of this talk have been withdrawn from sale; but a transcript, which repays reading from start to finish, can be found at www.heart7.net/mcf/greenbaum.htm.

[11] Patients of hypnosis-wielding MPD enthusiasts really have acquired crippling beliefs about their cult participation. That is why Bennett Braun, in 1997, had his license to practice suspended and why his insurers paid one of his tormented ex-patients a sobering malpractice settlement of $10.6 million.

[12] Alan W. Scheflin and Jerrold Lee Shapiro, *Trance on Trial* (Guilford Press, 1989).

[13] See, e.g., "The Memory Wars: Freud's Legacy in Dispute" (*New York Review of Books*, 1995), pp. 15–29; *Unauthorized Freud: Doubters Confront a Legend* (Viking, 1998), pp. x–xi; and "Forward to 1896? Commentary on Papers by Harris and Davies," *Psychoanalytic Dialogues*, vol. 6 (1996), pp. 231–250. That special number of *Psychoanalytic Dialogues* became a book edited by Richard B. Gartner, *Memories of Sexual Betrayal: Truth, Fantasy, Repression, and Dissociation* (Jason Aronson, 1997). My own contribution, however, was excised and replaced by an attack on my earlier criticisms of psychoanalysis.

[14] On this last point, see Bennett Simon, "'Incest – See Under Oedipus Complex': The History of an Error in Psychoanalysis," *Journal of the American Psychoanalytic Association*, vol. 40 (1992), pp. 955–988.

[15] See David Glenn, "Nightmare Scenarios," *Chronicle of Higher Education*, Oct. 24, 2003, pp. 14–17.

[16] A welcome new critique of fad therapies is *Science and Pseudoscience in Clinical Psychology*, ed. Scott O. Lilienfeld, Steven Jay Lynn, and Jeffrey M. Lohr (Guilford Press, 2003).

[17] See Robyn M. Dawes, *House of Cards: Psychology and Psychotherapy Built on Myth* (Free Press, 1994), especially pp. 10–22.

[18] Kenneth S. Pope and Laura S. Brown, *Recovered Memories of Abuse: Assessment, Therapy, Forensics* (American Psychological Association, 1996).

[19] See especially Allan Young, *The Harmony of Illusions: Inventing Post-Traumatic Stress Disorder* (Princeton Univ. Press, 1995), and Herb Kutchins and Stuart A. Kirk, *Making Us Crazy: DSM: The Psychiatric Bible and the Creation of Mental Disorders* (Free Press, 1997).

[20] As the Pied Pipers of recovered memory, Ellen Bass and Laura Davis, told prospective survivors in 1988, "When you first remember your abuse or acknowledge its effects, you may feel tremendous relief. Finally there is a reason for your problems. There is someone, and something, to blame." *The Courage to Heal: A Guide for Women Survivors of Child Sexual Abuse* (Harper & Row, 1988), p. 173.

### 0.0.3 Health Care: Who Knows 'Best'?, Jerome Groopman (*New York Review of Books*), February 11, 2010

February 11, 2010

Jerome Groopman (*New York Review of Books*)

One of the principal aims of the current health care legislation is to improve the quality of care. According to the President and his advisers, this should be done through science. The administration's stimulus package already devoted more than a billion dollars to "comparative effectiveness research," meaning, in the President's words, evaluating "what works and what doesn't" in the diagnosis and treatment of patients.

But comparative research on effectiveness is only part of the strategy to improve care. A second science has captured the imagination of policymakers in the White House: behavioral economics. This field attempts to explain pitfalls in reasoning and judgment that cause people to make apparently wrong decisions; its adherents believe in policies that protect against unsound clinical choices. But there is a schism between presidential advisers in their thinking over whether legislation should be coercive, aggressively pushing doctors and patients to do what the government defines as best, or whether it should be respectful of their own autonomy in making decisions. The President and Congress appear to be of two minds. How this difference is resolved will profoundly shape the culture of health care in America.

The field of behavioral economics is rooted in the seminal work of Amos Tversky and Daniel Kahneman begun some three decades ago. Drawing on data from their experiments on how people process information, particularly numerical data, these psychologists challenged the prevailing notion that the economic decisions we make are rational. We are, they wrote, prone to incorrectly weigh initial numbers, draw conclusions from single cases rather than a wide range of data, and integrate irrelevant information into our analysis. Such biases can lead us astray.

The infusion of behavioral economics into public policy is championed by Cass Sunstein, a respected professor of law and longtime friend of President Obama; he is now in the White House, overseeing regulatory affairs, and will have an important voice in codifying the details of any bill that is passed. In

*Nudge: Improving Decisions About Health, Wealth, and Happiness*, Sunstein and Richard Thaler, a professor of behavioral science and economics at the University of Chicago, propose that people called "choice architects" should redesign our social structures to protect against the incompetencies of the human mind. Those who understand thinking better can make life better for us all.

Thaler and Sunstein build on behavioral economic research that reveals inertia to be a powerful element in how we act. Most people, they argue, will choose the "default option" – i.e., they will follow a particular course of action that is presented to them instead of making an effort to find an alternative or opt out. Further, they write,

These behavioral tendencies toward doing nothing will be reinforced if the default option comes with some implicit or explicit suggestion that it represents the normal or even the recommended course of action.

Sunstein and Thaler propose to use default options as "nudges" in the service of "libertarian paternalism." For example, to promote a healthy diet among teenagers, broccoli and carrots would be presented at eye level in the cafeteria and would be easily available, while it would take considerable effort for students to locate junk food, thereby nudging them into accepting a healthier diet. But all choices should be "libertarian" – people should be free to opt out of "undesirable arrangements if they want to do so." The soft paternalistic nudge Sunstein and Thaler envisage should try "to influence choices in a way that will make choosers better off, as judged by themselves." They are very clear that nudges are not mandates, and that behavior should not be forcefully directed by changing economic incentives. Your doctor should not be paid less if she follows a course of treatment that she can defend as reasonable, even if she deviates from officially issued guidelines. To prevent policy planners from going down the slippery slope of coercion, there should, in Sunstein's view, be safety rails. Whatever the proposal put forward, he has written, people must retain "freedom of choice" and be able to oppose the more objectionable kinds of government intervention.

Such freedom of choice, however, is not supported by a second key Obama adviser, Peter Orszag, director of the Office of Management and Budget. In June 2008, testifying before Max Baucus's Senate Finance Committee,

Orszag – at the time director of the Congressional Budget Office – expressed his belief that behavioral economics should seriously guide the delivery of health care. In subsequent testimony, he made it clear that he does not trust doctors and health administrators to do what is "best" if they do no more than consider treatment guidelines as the "default setting," the procedure that would generally be followed, but with freedom to opt out. Rather, he said,

To alter providers' behavior, it is probably necessary to combine comparative effectiveness research with aggressive promulgation of standards and changes in financial and other incentives.

The word "probably" is gone in the Senate health care bill. Doctors and hospitals that follow "best practices," as defined by government-approved standards, are to receive more money and favorable public assessments. Those who deviate from federal standards would suffer financial loss and would be designated as providers of poor care. In contrast, the House bill has explicit language repudiating such coercive measures and protecting the autonomy of the decisions of doctors and patients.

On June 24, 2009, when President Obama convened a meeting on health care at the White House, Diane Sawyer of ABC News asked him whether federally designated "best practices" would be mandated or simply suggested. That is, would he recommend Orszag's shove or Sunstein's nudge?

Obama: Let's study and figure out what works and what doesn't. And let's encourage doctors and patients to get what works. Let's discourage what doesn't. Let's make sure that our payment incentives allow doctors to do the right thing. Because sometimes our payment incentives don't allow them to do the right things. And if we do that, then I'm confident that we can drive down costs significantly.

Sawyer: Will it just be encouragement? Or will there be a board making Solomonic decisions about best practices?

Obama: What I've suggested is that we have a commission made up of doctors, made up of experts, that helps set best practices.

Sawyer: By law?

Obama: If we know what those best practices are, then I'm confident that doctors are going to want to engage in best practices. But I'm also confident

patients are going to insist on it. In some cases, people just don't know what the best practices are. And certain cultures build up. And we can change those cultures, but it's going to require some work.

Sawyer: But a lot of people – say – "I'm very concerned that there's going be a reduction in treatment someplace in all of this." And the question is if there is a board that is recommending, that's one thing. If there is a board that is dictating through cost or through some other instruction, that's another thing. Will it have the weight of law? Will it have the weight of regulations?

Obama: I don't think that there's anybody who would argue for us continuing to pay for things that don't make us feel better. That doesn't make any sense. [Yet] that's the reason why, in America, we typically pay 50 percent more for our health care than other advanced countries that actually have better health care outcomes.

Still, the President appears not to be entirely in Orszag's camp. He has repeatedly deflected accusations of a "government takeover of health care" by asserting that no federal bureaucrat will come between the doctor and patient in clinical decision-making. The President has also repeatedly told physicians that reform would sustain them as healers, not make them into bean counters and paper pushers. In an interview on NPR two days before passage of the Senate bill, the President said that changes in how doctors and patients think about health care should come from giving them the "best information possible" and did not invoke the coercive measures favored by Orszag.

How do we reconcile this apparent difference between Sunstein and Orszag? The President contends that sound policies are built on data, but which data? Here the evidence is strongly in favor of Sunstein and his insistence on the need for freedom of choice and retaining the ability to oppose objectionable forms of government intervention. Over the past decade, federal "choice architects" – i.e., doctors and other experts acting for the government and making use of research on comparative effectiveness – have repeatedly identified "best practices," only to have them shown to be ineffective or even deleterious.

For example, Medicare specified that it was a "best practice" to tightly

control blood sugar levels in critically ill patients in intensive care. That measure of quality was not only shown to be wrong but resulted in a higher likelihood of death when compared to measures allowing a more flexible treatment and higher blood sugar. Similarly, government officials directed that normal blood sugar levels should be maintained in ambulatory diabetics with cardiovascular disease. Studies in Canada and the United States showed that this "best practice" was misconceived. There were more deaths when doctors obeyed this rule than when patients received what the government had designated as subpar treatment (in which sugar levels were allowed to vary).

There are many other such failures of allegedly "best" practices. An analysis of Medicare's recommendations for hip and knee replacement by orthopedic surgeons revealed that conforming to, or deviating from, the "quality metrics" – i.e., the supposedly superior procedure – had no effect on the rate of complications from the operation or on the clinical outcomes of cases treated. A study of patients with congestive heart failure concluded that most of the measures prescribed by federal authorities for "quality" treatment had no major impact on the disorder. In another example, government standards required that patients with renal failure who were on dialysis had to receive statin drugs to prevent stroke and heart attack; a major study published last year disproved the value of this treatment.

Other "quality measures" recommended by the government were carried out in community health centers to improve the condition of patients with asthma, diabetes, and hypertension. The conclusion of subsequent research was that there was, as a result, no change in outcome for any of these three disorders. Finally, Medicare, following the recommendations of an expert panel, specified that all patients with pneumonia must receive antibiotics within four hours of arrival at the emergency room. Many doctors strongly disagreed with such a rigid rule, pointing out that an accurate diagnosis cannot be made so quickly, and the requirement to treat within four hours was not based on convincing evidence. But the government went ahead, and the behavior of physicians was altered by the new default setting – for the worse. Many cases of heart failure or asthma, where the chest X-ray can resemble a pulmonary infection, were wrongly diagnosed as pneumonia; the misdiagnosed patients were given high doses of antibiotics, resulting in some

cases of antibiotic-induced colitis. The "quality measure" was ultimately rescinded.

What may account for the repeated failures of expert panels to identify and validate "best practices"? In large part, the panels made a conceptual error. They did not distinguish between medical practices that can be standardized and not significantly altered by the condition of the individual patient, and those that must be adapted to a particular person. For instance, inserting an intravenous catheter into a blood vessel involves essentially the same set of procedures for everyone in order to assure that the catheter does not cause infection. Here is an example of how studies of comparative effectiveness can readily prove the value of an approach by which "one size fits all." Moreover, there is no violation of autonomy in adopting "aggressive" measures of this kind to assure patient safety.

But once we depart from such mechanical procedures and impose a single "best practice" on a complex malady, our treatment is too often inadequate. Ironically, the failure of experts to recognize when they overreach can be explained by insights from behavioral economics. I know, because I contributed to a misconceived "best practice."

My early research involved so-called growth factors: proteins that stimulate the bone marrow to produce blood cells. I participated in the development of erythropoietin, the red cell growth factor, as a treatment for anemic cancer patients. Erythropoietin appeared to reduce the anemia, lessening the frequency of transfusion. With other experts, I performed a "meta-analysis," i.e., a study bringing together data from multiple clinical trials. We concluded that erythropoietin significantly improved the health of cancer patients and we recommended it to them as their default option. But our analysis and guidelines were wrong. The benefits ultimately were shown to be minor and the risks of treatment sometimes severe, including stroke and heart attack.

After this failure, I came to realize that I had suffered from a "Pygmalion complex." I had fallen in love with my own work and analytical skills. In behavioral economics, this is called "overconfidence bias," by which we overestimate our ability to analyze information, make accurate estimates, and project outcomes. Experts become intoxicated with their past success and fail to be sufficiently self-critical.

A second flaw in formulating "best practices" is also explained by behavioral economics – "confirmation bias." This is the tendency to discount contradictory data, staying wed to assumptions despite conflicting evidence. Inconsistent findings are rationalized as being "outliers." There were, indeed, other experts who questioned our anemia analysis, arguing that we had hastily come to a conclusion, neglecting findings that conflicted with our position. Those skeptics were right.

Yet a third powerful bias identified in behavioral economics can plague expert panels: this is the "focusing illusion," which occurs when, basing our predictions on a single change in the status quo, we mistakenly forecast dramatic effects on an overall condition. "If only I moved from the Midwest to sunny California, I would be so much happier" is a classical statement of a focusing illusion, proven to be such by studies of people who have actually moved across the country. Another such illusion was the prescription of estrogen as the single remedy to restore feminine youth and prevent heart disease, dementia, and other complications of the complex biology of aging. Such claims turned out to be seriously flawed.

There is a growing awareness among researchers, including advocates of quality measures, that past efforts to standardize and broadly mandate "best practices" were scientifically misconceived. Dr. Carolyn Clancy of the Agency for Healthcare Research and Quality, the federal body that establishes quality measures, acknowledged that clinical trials yield averages that often do not reflect the "real world" of individual patients, particularly those with multiple medical conditions. Nor do current findings on best practices take into account changes in an illness as it evolves over time. Tight control of blood sugar may help some diabetics, but not others. Such control may be prudent at one stage of the malady and not at a later stage. For years, the standards for treatment of the disease were blind to this clinical reality.

Orszag's mandates not only ignore such conceptual concerns but also raise ethical dilemmas. Should physicians and hospitals receive refunds after they have suffered financial penalties for deviating from mistaken quality measures? Should public apologies be made for incorrect reports from government sources informing the public that certain doctors or hospitals were not providing "quality care" when they actually were? Should a physician who is

skeptical about a mandated "best practice" inform the patient of his opinion? To aggressively implement a presumed but still unproven "best practice" is essentially a clinical experiment. Should the patient sign an informed consent document before he receives the treatment? Should every patient who is treated by a questionable "best practice" be told that there are credible experts who disagree with the guideline?

But even when there are no coercive measures, revising or reversing the default option requires a more complicated procedure than the one described by the President at the White House meeting. In November, the United States Preventive Services Task Force, reversing a long-standing guideline, recommended that women between the ages of forty and forty-nine do not need to have routine mammograms. To arrive at this conclusion, researchers made both a meta-analysis and computer models of data from seven clinical trials. The task force found that routine mammograms result in a 15 percent reduction in the relative risk of death from breast cancer for women in the forty to forty-nine age group, a similar level of benefit as in earlier analyses. For women in their forties, this means one life is saved for every 1,904 women screened. For older women in their fifties, one life is saved for every 1,359 women screened.

If these estimates are correct, then how many lives might be saved in the United States for each age group if every woman received a mammogram? The 2008 US Census estimates the number of women between forty and forty-nine at 22.3 million. So if mammography were available to all these women, nearly 12,000 deaths could be potentially averted during these ten years in their lives. As for the 20.5 million women in their fifties, some 15,000 deaths could potentially be averted.

What are the risks of mammography for women in their forties? The task force estimated a higher rate of false positive findings in mammograms in women in their forties compared to older women. This translates into increased anxiety when women are told that there may be a cancer and there is not. A false positive reading may also result in a woman having a biopsy. For every case of invasive breast cancer in a young woman diagnosed by mammography, five women with benign findings will have biopsies. In addition, there are potential risks of radiation from the mammogram itself, although no one

really knows how significant these are. Then there is an unanswered question in the biology of breast cancer: Which tumors are indolent and which are aggressive? We lack the molecular tools to distinguish between slow- and fast-growing cancers. Some slow-growing ones detected in young women might be treated later in life without any disadvantage in the rate of survival. But aggressive breast cancers in young women are notoriously difficult to treat and frequently result in death. And as with essentially all screening tests in a population, the majority of women receiving mammograms do not have any disorder.

These, roughly, are the statistics and state of the science with regard to breast cancer. How do we weigh the evidence and apply it to individuals and to society at large? Setting the default option that doctors will present to patients requires us to make value judgments. Dr. Otis Brawley of the American Cancer Society, an oncologist who worked for decades at the National Cancer Institute, is well versed in preventive care; he disagrees with the new default setting, based on findings that mammograms save lives. (Brawley also happens to be an African-American and has long been concerned about the meager access among minority and poor groups to potentially lifesaving screenings.)

Dr. Diana Petitti, a professor of bioinformatics at Arizona State University and vice-chair of the task force, appeared with Brawley on November 17, 2009, on the PBS NewsHour. She had no disagreement with him about what the studies show, and emphasized that the task force did not say that women in their forties should not get mammograms, only that they were no longer routinely recommended since the benefit to patients did not clearly outweigh the risks. Cost considerations were not part of the task force's deliberations.

Other supporters of the new recommendations took a less temperate view. A statistician who developed computer models for the task force told *New York Times* that "this decision is a no-brainer." It did not appear to be so clear to Melissa Block of NPR when she interviewed an internist who agreed with the task force. The doctor said that stopping routine mammography for young women would spare them anxiety, distress, and unnecessary biopsies. Block replied, "I've heard this before. When people say, you know, there's unnecessary anxiety and false positives and fear and worry." That, she said,

is "a very patronizing approach to take toward women's health. Women may very well be willing to assume those harms if it means that they may be diagnosed earlier." The internist replied that each woman should talk with her doctor and figure out what is best. Sunstein's Nudge coauthor, the behavioral economist Richard Thaler, wrote a thoughtful analysis of the pros and cons of mammography in *New York Times* and concluded that "one can make a good case that we don't want the government making these choices" for us.

Two days after the task force recommendations were released, Health and Human Services Secretary Kathleen Sebelius put some distance between the Obama administration and the task force's conclusions, saying:

My message to women is simple. Mammograms have always been an important life-saving tool in the fight against breast cancer and they still are today. Keep doing what you have been doing for years.

Dr. Petitti later appeared before Congress to apologize for any "confusion" caused by the task force report. Petitti was not recanting a scientific truth. She correctly described the new recommendations as "qualitative." That is, they were offered as value judgments that could be modified or revised; and the political process offers one way of doing so. As Sunstein has written, if default options embody standards that many people judge as not better for themselves, those standards can be changed.

Shortly after the new mammography guidelines were announced, an expert panel of obstetricians and gynecologists recommended that teenage girls no longer have routine pap smears for cervical cancer. The incidence of deadly cervical cancer among teens is at most one in a million and screening does not appear to save that one life. When false positive results from screenings are followed by cervical surgery, the risk may be injury that can predispose a young woman to later premature labor. There was no public uproar following this changed default setting for many women. It was consistent with how most people value the benefit of lives saved versus risks incurred. This is the reality of "comparative effectiveness" research. It is not simply a matter of "what works and what doesn't." Nor will patients always "insist" on being treated according to what experts define as "best practice." They should be aware that there are numerous companies, some of them "not for profit,"

issuing standards for treatment that are congenial to the insurance industry but are often open to the kinds of counterevidence I have described here.

What of the President's statement that doctors will want to engage in federally approved "best practices"? The American College of Physicians, composed of internists, agreed with the task force conclusions about mammography. The American Society of Clinical Oncology, representing oncologists, did not. I am a member of both professional organizations. What do I do? As a physician who has cared for numerous young women with breast cancer, many dying an untimely death, my bias was that the dangers of mammograms do not outweigh the reduction in mortality. Notably, the oncologists who head the breast cancer programs at Minnesota's Mayo Clinic and Utah's Intermountain Health – described by President Obama as pinnacles of quality care using guidelines – also disagreed with the task force.

Such challenges to "best practice" do not imply that doctors should stand alone against received opinion. Most physicians seek data and views on treatments from peers and, as needed, specialists, and then present information and opinion to patients who ultimately decide.

While costs were not part of the task force calculations, they prominently entered the national debate on them. Dr. Robert Truog of Boston Children's Hospital allowed that mammography saves lives, but asked if it is "cost effective." That is, should policy planners set a price on saving those young women?

Cost-effectiveness is going to be a hard sell to the American public, not only because of the great value placed on each life in the Judeo-Christian tradition, but because the federal government has devoted many hundreds of billions of dollars to bail out Wall Street. To perform mammograms for all American women in their forties costs some $3 billion a year, a pittance compared to the money put into the bank rescue. The Wall Street debacle also made many Americans suspicious of "quants," the math whizzes who developed computer models that in theory accurately assessed value in complex monetary instruments but in fact nearly brought down the worldwide financial system. When a medical statistician says that imposing a limit on mammography is a "no-brainer," people may recall George Tenet's claim that the case for invading Iraq was a "slam-dunk."

At the White House gathering, the President portrayed comparative effectiveness as equivalent to cost-effectiveness, noting that other countries spend half of what we do by only paying for "what works." This contention is not supported by evidence. Theodore Marmor, a professor of health care policy at Yale, writes in *Fads, Fallacies and Foolishness in Medical Care Management and Policy* that movements for "quality improvement" in Britain have failed to reduce expenditures. Marmor, with Jonathan Oberlander, a professor at the University of North Carolina, has written in these pages that the President has offered up rosy scenarios to avoid the harsh truth that there is no "painless cost control." Lower spending in countries like France and Germany is accounted for not by comparative effectiveness studies but by lower costs of treatment attained through their systems of medical care and by reduced medical budgets. In Europe, prescription drugs cost between 50 and 60 percent of what they do in the US, and doctor's salaries are lower. (Insurance premiums also are tightly constrained.) France and Germany have good records in health care, but in Great Britain, where costs are strictly controlled by the National Health Service, with rationing of expensive treatments, outcomes for many cancers are among the worst in Europe.

The care of patients is complex, and choices about treatments involve difficult tradeoffs. That the uncertainties can be erased by mandates from experts is a misconceived panacea, a "focusing illusion." If a bill passes, Cass Sunstein will be central in drawing up the regulations that carry out its principles. Let's hope his thinking prevails.

– January 14, 2010

On August 7, 2008, addressing the Retirement Research Consortium in Washington, D.C., Orszag presented "Behavioral Economics: Lessons from Retirement Research for Health Care and Beyond." Here, he states the likely need for aggressive measures. The Senate Finance Committee, under Max Baucus, was widely reported to have worked closely with the White House, and many of Orszag's proposals are prominent in the bill that Majority Leader Harry Reid brought to the floor. See Senate Bill HR 3590, Title III – Improving the Quality and Efficiency of Health Care. The House rejected many of the ideas from the President's advisers in favor of safeguards on patient-physician autonomy, causing Rahm Emanuel, the White House chief of staff,

to quip that politics trumps "ideal" plans made in the shade of the "Aspen Institute." See Sheryl Gay Stolberg, "Democrats Raise Alarms over Health Bill Costs," *New York Times*, November 9, 2009. Explicit language in the House bill is intended to safeguard patient-physician autonomy. See House Bill HR 3962, Title IV – Quality; Subtitle A – Comparative Effectiveness Research.

---

To the Editors:

In his cautionary essay, Jerome Groopman writes about the dangers of governments and regulators taking a prescriptive approach to medical practice based on "best practice" guidelines. In support of his skepticism about their value, he gives examples of guidelines that have been overturned after accumulating evidence indicated practices were being recommended that were at best useless. Among others, these included recommendations that ambulatory diabetics should have their blood glucose very tightly controlled for the sake of their cardiovascular health, chronic renal failure patients on dialysis should take statin drugs to reduce their vascular event rate, patients with pneumonia must be treated with antibiotics within four hours, and anemic cancer patients should be treated with erythropoietin.

But critically, what Dr. Groopman fails to mention is that none of these recommendations was supported by high-quality evidence even when it was written. None was supported by a large randomized trial showing improvements in real clinical outcomes. Instead the studies on which the guidelines were based measured "surrogate" outcomes, which were supposed to be as good as clinical outcomes, simple examples being the measurement of cholesterol levels in the case of statins and the measurement of red blood cell counts in the case of erythropoetin.

There are probably many reasons guideline writers are way out in front of the available evidence, not limited to their financial ties to industry previously documented in *New York Review* [Marcia Angell, "Drug Companies & Doctors" *NYR*, January 15, 2009]. The biggest problem is not that there is a real likelihood of future regulators being dangerously overzealous in their application of guidelines, but that many guidelines are simply not justified by evidence.

Current examples are easy to find: the American College of Cardiology/American Heart Association guidelines on the management of (non-ST elevation) acute coronary syndromes recommends the routine use of ACE inhibitors despite no support from trials in this clinical scenario. The same guidelines also recommend treatment with cholesterol-lowering agents to achieve certain low cholesterol targets, the consequence of which is a huge industry involving repeat visits to clinicians for cholesterol measurements and dose adjustment or addition of drugs when these targets are not met. This is despite there being no convincing evidence that this is any better than simply prescribing a high-potency statin drug and sending the patient on his way.

Dr. Groopman seems to miss all this, believing that the major "conceptual error" guideline writers make is failing to recognize that their guidelines may not be applicable to "the individual patient." I would have thought a bigger issue is whether they are applicable to any patient at all.

Matthew Pincus
Senior Lecturer, Department of Medicine
University of Queensland
Brisbane, Australia

To the Editors:

Jerome Groopman glosses over the real danger in mammography: overdiagnosis. He did concede that one life would be saved for screening 1,904 women in their forties, but he left out that ten healthy women would be treated unnecessarily (i.e., surgery, radiation, chemo). This is based on an overdiagnosis rate of 30 – 70 percent (!) by Norwegian and Danish epidemiologists, which is naturally disputed by the American Cancer Society and others. If this is correct, it means that it is not just a matter of harmlessly delaying treatment for an indolent cancer; it is a matter of many of the invasive cancers identified by mammography and biopsy disappearing on their own, and better left undetected and "untreated."

So it boils down to a philosophic choice. It would be nice if these discussions with patients could be encouraged, so they can make up their own minds. For those patients who just want the doctor to tell them what to do, they are asking the doctor to pretend he knows the right answer.

I appreciate Groopman drawing attention to "the focusing illusion," but it may be more prevalent than he realizes. All oncologists have suffered through the premature deaths of wonderful women from breast cancer; how could they possibly know that something they treated brilliantly and humanely was never going to cause problems? Now there's a potential focusing illusion.

Finally, Groopman brilliantly highlights the irony that the White House is pushing best practices as a cost-cutting measure, while this effort has not been shown to cut costs elsewhere, nor improve care.

Richard Ganz, M.D. Healdsburg, California

Jerome Groopman replies:

Dr. Pincus omits the first of the mandated "best practices" enumerated in my article: tight regulation of blood glucose in critically ill patients in the intensive care unit. This was among the most aggressively promulgated guidelines by the government and insurers. Contrary to his contention that prior recommendations relied on surrogates rather than meaningful clinical outcomes, tight regulation of blood glucose in ICU patients was based on randomized prospective clinical trials that measured death as the outcome.

These studies, as well as subsequent research that contradicted their findings, were published in *New England Journal of Medicine* and are cited in the footnotes of my essay. The recommendation that I prematurely endorsed on erythropoietin treatment for cancer patients was based not only on increasing the "surrogate" of red blood cell counts but also data on improving quality of life, sparing patients the risks of transfusion, and preserving the precious resource of donated blood for other patients like those who hemorrhage.

These facts belie Dr. Pincus's critique. Furthermore, Dr. Pincus oversimplifies the difficulties in crafting "prescriptive guidelines" that standardize therapies for the kinds of patients seen in daily clinical practice. Statistics from randomized trials, as he surely knows, represent averages of selected groups of patients. Knowing how to "best" treat an individual patient, particularly one who has concurrent medical problems that would have barred him from the clinical trial, requires referring to the guidelines but not necessarily adhering to them. But there is an even more important flaw in Dr. Pincus's analysis. How experts judge the "quality" of evidence is hardly a uniform or objective process. Randomized controlled clinical trials, which are

taken as usually yielding more reliable data, nonetheless are hotly debated among experts with respect to their design: which patients were included and excluded, which outcomes were "primary," meaning the overriding aims of the study, and which were "secondary," meaning providing information on other possible benefits of the treatment.

Different experts bring their own mind-sets and biases to bear in judging not only the quality of evidence from these clinical trials but the tradeoffs between risks and benefits of the therapy. For example, in the randomized prospective studies of tight control of blood glucose in ambulatory diabetics with cardiovascular disease, there are indications of possible benefit with regard to protecting small blood vessels from the deleterious effects of diabetes, thereby sustaining kidney function and averting blindness; but offsetting these potential gains, tight control may promote heart attack and stroke and increase the risk of death.

Indeed, every physician has attended clinical conferences where credible specialists debate the sagacity of trial design and the trade-offs between risk and benefit of the treatment. But one need not enter a medical center to witness such a debate. The medical journals routinely publish editorials in conjunction with large randomized clinical trials in which independent researchers in the field point out the strengths and weaknesses of the studies. And within weeks of publication of the data, the same medical journals are filled with letters from credible critics who point out pitfalls in design and in the execution of the clinical trial and weigh in with their own interpretation of the risk versus benefit trade-off. The better press coverage of these trials includes the expert voices of both advocates and dissenters when presenting results of clinical research to the public. It is very rare that we have situations in clinical medicine in which a black or white answer is apparent.

Does this mean we should do away with guidelines? Not at all. Rather, the major point of my essay was the probity of mandates versus suggestions. If an expert committee is convened with the imperative to come to a consensus and write a mandated guideline, then it will do just that. But if patients and their physicians are provided with the full range of expert opinions, zealous and conservative, and specialists articulate such views, explaining how they weighed the "quality" of the evidence, then a great service is done to foster

truly informed choices. Dr. Pincus ignores the reality that clinical data from randomized trials are imperfect and credible experts bring their biases to bear when they judge evidence to be of high or low quality, even when untainted by financial conflicts of interest.

Dr. Ganz cites an inference from a single epidemiological study that is far from proven, as the authors forthrightly state in their publication. There are no direct, prospective, and compelling data on breast cancer spontaneously remitting. A more important issue in breast cancer diagnosis and treatment is what is termed *ductal carcinoma in situ*, or DCIS. This is a very early stage of the malignancy, often detected by mammogram. There is considerable controversy about how often and how quickly DCIS grows into an invasive cancer, and whether it should be treated by surgery or radiation or hormonal blockers like Tamoxifen. There are a number of well-designed ongoing clinical studies to obtain better knowledge about DCIS, and with it, hopefully, provide women and their physicians with a sounder basis to make decisions.

---

To the Editors:

Jerome Groopman is right that treatment guidelines often do not improve quality of care because they fail to take into account variability in co-morbidity and course of illness. They fail for other reasons as well, particularly the objectivity of those making the guidelines. So many "experts" are influenced by emoluments from pharmaceutical and medical device companies that finding any with unencumbered objectivity is difficult. Interestingly, every example Groopman gives of "best practices" that went awry, and that involved drugs, erred on the side of what proved to be excessive use.

Dr. Groopman points out "that there are numerous companies ... issuing standards for treatment that are congenial to the insurance industry" but he fails to mention how "standards" issued by pharmaceutical and medical device manufacturers influence doctors. Pharmaceutical companies dispatch thousands of representatives to persuade doctors to use their companies' drugs and spend billions of dollars on advertising to consumers and physicians.

Finally, he fails to consider how the development of guidelines can be

improved. One step is to prohibit physicians and others with conflicts of interest from serving on expert panels. Taken together with strict limitations on what any physician can receive from drug and device companies, it could improve quality of care.

Neil A. Holtzman, M.D., M.P.H.

Emeritus Professor, School of Medicine The Johns Hopkins University Baltimore, Maryland

Jerome Groopman replies:

Dr. Holtzman raises an important and urgent issue. Not only has there been considerable marketing by pharmaceutical companies to individual physicians, but the government has outsourced the task of creating treatment guidelines to expert panels where some members may have significant conflicts of interest.

National guidelines that are adopted by Medicare should not be outsourced. Rather, there should be direct federal oversight of "in-house" panels, similar to the process used by the Food and Drug Administration to evaluate data and approve or deny drugs for specific clinical indications.

Expert panels assembled by the Center for Medicare and Medicaid Services should be rigorously vetted for potential conflicts, and an expert should not be writing guidelines that would financially benefit a pharmaceutical company that supports him as a consultant or researcher. The Institute of Medicine of the National Academy of Science is currently pondering recommendations to the government on how to regulate such conflicts of interest.

The concerns raised by Dr. Holtzman and others are valid and should inform new federal policies that result in better guidelines.

### 0.0.4 Trawling the Brain, Laura Sanders (*Science News*), December 19, 2009

December 19, 2009

Laura Sanders (*Science News*)

New findings raise questions about reliability of fMRI as gauge of neural activity.

The 18-inch-long Atlantic salmon lay perfectly still for its brain scan. Emotional pictures – a triumphant young girl just out of a somersault, a distressed waiter who had just dropped a plate – - flashed in front of the fish as a scientist read the standard instruction script aloud. The hulking machine clunked and whirred, capturing minute changes in the salmon's brain as it assessed the images. Millions of data points capturing the fluctuations in brain activity streamed into a powerful computer, which performed herculean number crunching, sorting out which data to pay attention to and which to ignore.

By the end of the experiment, neuroscientist Craig Bennett and his colleagues at Dartmouth College could clearly discern in the scan of the salmon's brain a beautiful, red-hot area of activity that lit up during emotional scenes.

An Atlantic salmon that responded to human emotions would have been an astounding discovery, guaranteeing publication in a top-tier journal and a life of scientific glory for the researchers. Except for one thing. The fish was dead.

The scanning technique used on the salmon – called functional magnetic resonance imaging – allows scientists to view the innards of a working brain, presumably reading the ebbs and flows of activity that underlie almost everything the brain does. Over the last two decades, fMRI has transformed neuroscience, enabling experiments that researchers once could only dream of. With fMRI, scientists claim to have found the brain regions responsible for musical ability, schadenfreude, Coca-Cola or Pepsi preference, fairness and even tennis skill, among many other highly publicized conclusions.

But many scientists say that serious issues have been neglected during fMRI's meteoric rise in popularity. Drawing conclusions from an fMRI experiment requires complex analyses relying on chains of assumptions. When subjected to critical scrutiny, inferences from such analyses and many of the

assumptions don't always hold true. Consequently, some experts allege, many results claimed from fMRI studies are simply dead wrong.

"It's a dirty little secret in our field that many of the published findings are unlikely to replicate," says neuroscientist Nancy Kanwisher of MIT.

A reanalysis of the salmon's postmortem brain, using a statistical check to prevent random results from accidentally seeming significant, showed no red-hot regions at all, Bennett, now at the University of California, Santa Barbara, and colleagues report in a paper submitted to *Human Brain Mapping*. In other words, the whole brain was as cold as a dead fish.

Less dramatic studies have also called attention to flawed statistical methods in fMRI studies. Some such methods, in fact, practically guarantee that researchers will seem to find exactly what they're looking for in the tangle of fMRI data. Other new research raises questions about one of the most basic assumptions of fMRI – that blood flow is a sign of increased neural activity. At least in some situations, the link between blood flow and nerve action appears to be absent. Still other papers point out insufficient attention to insidious pitfalls in interpreting the complex enigmatic relationship between an active brain region and an emotion or task.

Make no mistake: fMRI is a powerful tool allowing neuroscientists to elucidate some of the brain's deepest secrets. It "provides you a different window into how mental processes work in the brain that we wouldn't have had without it," says Russell Poldrack of the University of Texas at Austin.

But like any powerful tool, fMRI must be used with caution. "All methods have shortcomings – conclusions they support and conclusions they don't support," Kanwisher says. "Neuroimaging is no exception."

BOLD assumptions:

fMRI machines use powerful magnets, radio transmitters and detectors to peer into the brain. First, strong magnets align protons in the body with a magnetic field. Next, a radio pulse knocks protons out of that alignment. A detector then measures how long it takes for the protons to recover and emit telltale amounts of energy. Such energy signatures act as beacons, revealing the locations of protons ensconced in specific molecules.

fMRI is designed to tell researchers which brain regions are active – the areas where nerve cells are abuzz with electrical signals. Scientists have known

for a long time how to record these electrical communiques with electrodes, which can sit on the scalp or be implanted in brain tissue. Yet electrodes outside the skull can't precisely pinpoint active regions deep within the brain, and implanting electrodes in the brain comes with risks. fMRI, on the other hand, offers a nonintrusive way to measure neuron activity, requiring nothing more of the subject than an ability to lie in a big tube for a while.

But fMRI doesn't actually measure electrical signals. Instead, the most common fMRI method, BOLD (for blood oxygen level-dependent), relies on tiny changes in oxygenated blood as a proxy for brain activity. The assumption is that when neurons are working hard, they need more energy, brought to them by fresh, oxygen-rich blood. Protons in oxygen-laden hemoglobin molecules, whisked along in blood, respond to magnetic fields differently than protons in oxygen-depleted blood. Detecting these different signatures allows researchers to follow the oxygenated blood to track brain activity – presumably.

(Most brain regions and mental tasks don't match up one-to-one, confounding the interpretation of fMRI results. Pain activates many regions throughout the brain. One such region, the anterior cingulate cortex, is also activated by many other functions.)

"There's still some mystery," Bennett says. "There are still some things we don't understand about the coupling between neural activity and the BOLD signal that we're measuring in fMRI."

Researchers use BOLD because it's the best approximation to neural activity that fMRI offers. And for the most part, it works. But a study published in January in *Nature* reported that the link between blood flow and neural activity is not always so clear. In their experiments, Aniruddha Das and Yevgeniy Sirotin, both of Columbia University, found that in monkeys some blood changes in the brain had nothing to do with localized neuron firing.

Das and Sirotin used electrodes to measure neuronal activity at the same time and place as blood flow in monkeys who were looking at an appearing and disappearing dot. As expected, when vision neurons detected the dot and fired, blood rushed into the scrutinized brain region. But surprisingly, at times when the dot never appeared and the neurons remained silent, the researchers also saw a dramatic change in blood flow. This unprompted change

in blood flow occurred when the monkeys were anticipating the dot, the researchers found. The imperfect correlations between blood flow and neural firing can confound BOLD signals and muddle the resulting conclusions about brain activity.

Mass action:

Another fMRI difficulty arises from its view-from-the-top scale. Predicting a single neuron's activity from fMRI is like trying to tell which way an ant on the ground is crawling from the top of the Washington Monument, without binoculars. The smallest single unit measured by BOLD fMRI, called a voxel, is often a few millimeters on each side, dwarfing the size of individual neurons. Each voxel – a mashup of volume and pixel – holds around 5.5 million neurons, calculates Nikos Logothetis of the Max Planck Institute for Biological Cybernetics in Tbingen, Germany. Assuming that the millions of neurons in a voxel perform identically is like assuming every single ant on the National Mall crawls north at noon.

"fMRI is a measure of mass action," Logothetis says. "You almost have to be a professional moron to think you're saying something profound about the neural mechanisms. You're nowhere close to explaining what's happening, but you have a nice framework, an excellent starting point." BOLD signals could reflect many different events, he says. For instance, some neurons send signals that stop other neurons from firing, so increased activity of these dampening neurons could actually lead to an overall decrease in neuron activity.

Kanwisher points out that words such as "activity" and "response," mainstays of fMRI paper titles, are intentionally vague. Pinning down the details from such a zoomed-out view, she says, is impossible. "What exactly are the neurons doing in there? Is one inhibiting the other? Are there action potentials? Is there synaptic activity? Well, we have no idea," she says. "It would be nice to know what the neurons are doing, but we don't with this method. And that's life."

Inadvertent mischief:

After BOLD signals have been measured and the patient has been released from the machine, researchers must sort the red-hot voxels from the dead fish. Statistics for dealing with these gigantic data sets are so complex that some

researchers outsource the analyses to professional number crunchers. Choosing criteria to catch real and informative brain changes, and guarding against spurious results, is one of the most important parts of an fMRI experiment, and also one of the most opaque.

"It's hellishly complicated, this data analysis," says Hal Pashler, a psychologist at the University of California, San Diego. "And that creates great opportunity for inadvertent mischief."

Making millions, often billions, of comparisons can skew the numbers enough to make random fluctuations seem interesting, as with the dead salmon. The point of the salmon study, Bennett says, was to point out how easy it is to get bogus results without the appropriate checks.

Bennett and colleagues have written an editorial to appear in *Social Cognitive and Affective Neuroscience* that argues for strong measures to protect against false alarms. Another group takes the counterpoint position, arguing that these protections shouldn't be so strong that the real results are tossed too, like a significant baby with the statistical bathwater.

One of the messiest aspects of fMRI analysis is choosing which part of the brain to scrutinize. Some studies have dealt with this problem by selecting defined anatomical regions in advance. Often, though, researchers don't know where to focus, instead relying on statistics to tell them which voxels in the entire brain are worth a closer look.

In a paper originally titled "Voodoo correlations in social neuroscience" in the May issue of *Perspectives on Psychological Science*, Edward Vul of MIT, Pashler and colleagues called out 28 fMRI papers (of 53 analyzed) for committing the statistical sin of "nonindependence." In nonindependent analyses, the hypothesis in question is not an innocent bystander, but in fact distorts the experiment's outcome. In other words, the answer is influenced by how the question is asked.

One version of this error occurs when researchers define interesting voxels with one set of criteria – say, those that show a large change when a person is scared – and then use those same voxels to test the strength of the link between voxel and fear. Not surprisingly, the correlation will be big. "If you have many voxels to choose from, and you choose the largest ones, they'll be large," Vul says.

In a paper in the May *Nature Neuroscience*, Nikolaus Kriegeskorte of the Medical Research Council in Cambridge, England, and colleagues call the non-independence issue the error that "beautifies" results. "It tends to clean things up at the expense of a veritable representation of the data," Kriegeskorte says.

Digging through the methods sections of fMRI papers published in 2008 in *Nature*, *Science*, *Nature Neuroscience*, *Neuron* and the *Journal of Neuroscience* turned up some sort of nonindependence error in 42 percent, Kriegeskorte and colleagues report in their paper. Authors "do very complicated analyses, and they don't realize that they're actually walking in a very big circle, logically," Kriegeskorte says.

Kanwisher, who just cowrote a book chapter with Vul about the non-independence error, says that researchers can lean too heavily on "fancy" math. "Statistics should support common sense," she says. "If the math is so complicated that you don't understand it, do something else."

The problem with blobology:

An issue that particularly irks some researchers has little to do with statistical confounders in fMRI, but rather with what the red-hot blobs in the brain images actually mean. Just because a brain region important for a particular feeling is active does not mean a person must be feeling that feeling. It's like concluding that a crying baby must be hungry. True, a hungry baby does cry, but a crying baby might be tired, feverish, frightened or wet while still well-fed.

Likewise, studies have found that a brain structure called the insula is active when a person is judging fairness. But if a scan shows the insula to be active, the person is not necessarily contemplating fairness; studies have found that the insula also responds to pain, tastes, interoceptive awareness, speech and memory.

In most cases, the brain does not rely on straightforward relationships, with a specific part of the brain responsible for one and only one task, making these reverse inferences risky, Poldrack points out.

"Researchers often assume that there are one-to-one relations between brain areas and mental functions," he says. "But we don't actually know if that is true, and there are many reasons to think that it's not." Inferring

complex human emotions from the activity of a single brain region is not something that should be done casually, as it often is, he says.

Sometimes, reverse inference is warranted, though, as long as it is done with care. "There's nothing wrong with saying there's a brain region for x," Kanwisher says. "It just takes many years to establish that. And like all other results, you establish it, and it can still crash if somebody presents a new piece of data that argues against it."

Marco Iacoboni of the University of California, Los Angeles and colleagues drew heat from fellow neuroscientists for *New York Times* op-ed in November 2007 in which the team claimed to have ascertained the emotional states of undecided voters as they were presented with pictures of candidates. For instance, the researchers concluded that activity in the anterior cingulate cortex meant that subjects were "battling unacknowledged impulses to like Mrs. Clinton." Poldrack and 16 other neuroscientists quickly wrote their own editorial, saying that the original article's claims had gone too far.

Iacoboni counters that reverse inference has a valuable place in research, as long as readers realize that it is a probabilistic measure. "A little bit of reverse inference, to me, is almost necessary," he says.

Careful language and restrained conclusions may solve some of the issues swirling around fMRI interpretations, but a more serious challenge comes from fMRI's noise. Random fluctuations masquerading as bona fide results are insidious, but the best way to flush them out is simple: Do the experiment again and see if the results hold up. This built-in reality check is time-consuming and expensive, Kanwisher says, but it's the best line of defense against spurious results.

A paper published April 15 in *NeuroImage* clearly illustrates the perils of one-off experiments. In an fMRI experiment, Bradley Schlaggar of Washington University in St. Louis and colleagues found differences in 13 brain regions between men and women during a language task. To see how robust these results were, the researchers scrambled the groups to create random mixes of men and women. Any differences found between these mixed-up groups could be chalked up to noise or unknown factors, the researchers reasoned. The team found 14 "significant" different regions between the scrambled groups, undermining the original finding and rendering the experiment

uninterpretable.

"The upshot of the paper is really a cautionary one," Schlaggar says. "It's easy and common to find some group differences at some statistical threshold. So go ahead and do the study again."

In many ways, fMRI has earned its reputation as a powerful neuroscience tool. In the laboratories of capable, thoughtful researchers, the challenges, exceptions and assumptions that plague fMRI can be overcome. Its promise to decode the human brain is real. fMRI "is a great success story of modern science, and I think historically it will definitely be viewed as that," Kriegeskorte says. "Overwhelmingly it is a very, very positive thing."

But the singing of fMRI's praises ought to be accompanied by a chorus of caveats. fMRI cannot read minds nor is it bogus neophrenology, as Logothetis pointed out in *Nature* in 2008. Rather, fMRI's true capabilities fall somewhere between those extremes. Ultimately, understanding the limitations of neuroimaging, instead of ignoring them, may propel scientists toward a deeper understanding of the brain.

Citations & References:

Bennett, C. In press. Neural correlates of interspecies perspective taking in the post-mortem Atlantic salmon: an argument for proper multiple comparisons correction. *Human Brain Research.*

Sirotin, Y.B., & A. Das (2009). Anticipatory haemodynamic signals in sensory cortex not predicted by local neuronal activity. *Nature 457* (Jan. 22):475–479.

Poldrack, R.A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences 10* (February):59–63.

Logothetis, N.K. (2008). What we can do and what we cannot do with fMRI. *Nature 453* (June 12):869–878.

Bennett, C., G. Wolford, & M. Miller. In press. The principled control of false positives in neuroimaging. *Social Cognitive and Affective Neuroscience.*

Kriegeskorte, N., et al. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature Neuroscience 12* (May):535–540.

Vul, E., & N. Kanwisher. In press. Begging the Question: The Non-Independence Error in fMRI Data Analysis. In Hanson, S. & Bunzl, M. (Eds.). *Foundations and Philosophy for Neuroimaging.*

New York Times OpEd by Iacoboni, et al. *This is your brain on politics.*

New York Times OpEd response by Poldrack, et al. *Politics and the brain.*

Ihnen, S.K.Z., et al. (2009). Lack of generalizability of sex differences in the fMRI BOLD activity associated with language processing in adults. *NeuroImage 45*:1020 – -1032.

### 0.0.5 The Cancer-Cluster Myth, Atul Gawande (*New Yorker*), February 8, 1999

February 8, 1999

Atul Gawande (*New Yorker*)

When a dozen people in a neighborhood develop tumors, it can't be coincidence. Or can it?

Is it something in the water? During the past two decades, reports of cancer clusters – communities in which there seems to be an unusual number of cancers – have soared. The place names and the suspects vary, but the basic story is nearly always the same. The Central Valley farming town of McFarland, California, came to national attention in the eighties after a woman whose child was found to have cancer learned of four other children with cancer in just a few blocks around her home. Soon doctors identified six more cases in the town, which had a population of sixty-four hundred. The childhood-cancer rate proved to be four times as high as expected. Suspicion fell on groundwater wells that had been contaminated by pesticides, and lawsuits were filed against six chemical companies.

In 1990, in Los Alamos, New Mexico, a local artist learned of seven cases of brain cancer among residents of a small section of the town's Western Area. How could seven cases of brain cancer in one neighborhood be merely a coincidence? "I think there is something seriously wrong with the Western Area," the artist, Tyler Mercier, told the *Times*. "The neighborhood may be contaminated." In fact, the Los Alamos National Laboratory, which was the birthplace of the atomic bomb, had once dumped millions of gallons of radioactive and toxic wastes in the surrounding desert, without providing any solid documentation about precisely what was dumped or where. In San Ramon, California, a cluster of brain cancers was discovered at a high-school class reunion. On Long Island, federal, state, and local officials are currently spending twenty-one million dollars to to try to find out why towns like West Islip and Levittown have elevated rates of breast cancer.

I myself live in a cancer cluster. A resident in my town – Newton, Massachusetts – became suspicious of a decades-old dump next to an elementary

school after her son developed cancer. She went from door to door and turned up forty-two cases of cancer within a few blocks of her home. The cluster is being investigated by the state health department.

No doubt, one reason for the veritable cluster of cancer clusters in recent years is the widespread attention that cases like those in McFarland and Los Alamos received, and the ensuing increase in public awareness and concern. Another reason, though, is the way in which states have responded to that concern: they've made available to the public data on potential toxic sites, along with information from "cancer registries" about local cancer rates. The result has been to make it easier for people to find worrisome patterns, and more and more, they've done so. In the late eighties, public-health departments were receiving between thirteen hundred and sixteen hundred reports of feared cancer clusters, or "cluster alarms," each year. Last year, in Massachusetts alone, the state health department responded to between three thousand and four thousand cluster alarms. Under public pressure, state, and federal agencies throughout the country are engaging in "cancer mapping" to find clusters that nobody has yet reported.

A community that is afflicted with an unusual number of cancers quite naturally looks for a cause in the environment – in the ground, the water, the air. And correlations are sometimes found: the cluster may arise after, say, contamination of the water supply by a possible carcinogen. The problem is that when scientists have tried to confirm such causes, they haven't been able to. Raymond Richard Neutra, California's chief environmental health investigator and an expert on cancer clusters, points out that among hundreds of exhaustive, published investigations of residential clusters in the United States, not one has convincingly identified an underlying environmental cause. Abroad, in only a handful of cases has a neighborhood cancer cluster been shown to arise from an environmental cause. And only one of these cases ended with the discovery of an unrecognized carcinogen. It was in a Turkish village called Karain, where twenty-five cases of mesothelioma, a rare form of lung cancer, cropped up among fewer than eight hundred villagers. (Scientists traced the cancer to a mineral called erionite, which is abundant in the soil there.) Given the exceedingly poor success rate of such investigations, epidemiologists tend to be skeptical about their worth.

When public-health investigators fail to turn up any explanation for the appearance of a cancer cluster, communities can find it frustrating, even suspicious. After all, these investigators are highly efficient in tracking down the causes of other kinds of disease clusters. "Outbreak" stories usually start the same way: someone has an intuition that there are just too many people coming down with some illness and asks the health department to investigate. With outbreaks, though, such intuitions are vindicated in case after case. Consider the cluster of American Legionnaires who came down with an unusual lung disease in Philadelphia in 1976; the startling number of limb deformities among children born to Japanese women in the sixties; and the appearance of rare *Pneumocystis carinii* pneumonia in five young homosexual men in Los Angeles in 1981. All these clusters prompted what are called "hot-pursuit investigations" by public-health authorities, and all resulted in the definite identification of a cause: namely, *Legionella* pneumonitis, or Legionnaires' disease; mercury poisoning from contaminated fish; and H.I.V. infection. In fact, successful hot-pursuit investigations of disease clusters take place almost every day. A typical recent issue of the Centers for Disease Control's *Morbidity and Motality Weekly Report* described a cluster of six patients who developed muscle pain after eating fried fish. Investigation by health authorities identified the condition as Haff disease, which is caused by a toxin sometimes present in buffalo fish. Four of the cases were traced to a single Louisiana wholesaler, whose suppliers fished the same tributaries of the Mississippi River.

What's more, for centuries scientists have succeeded in tracking down the causes of clusters of cancer that aren't residential. In 1775, the surgeon Percivall Pott discovered a cluster of scrotal-cancer cases among London chimney sweeps. It was common practice then for young boys to do their job naked, the better to slither down chimneys, and so high concentrations of carcinogenic coal dust would accumulate in the ridges of their scrota. Pott's chimney sweeps proved to be a classic example of an "occupational" cluster. Scientists have also been successful in investigating so-called "medical" clusters. In the late nineteen-sixties, for example, the pathologist Arthur Herbst was surprised to come across eight women between the ages of fifteen and twenty-two who had clear-cell adenocarcinoma, a type of cervical cancer that had never

been seen in women so young. In 1971, he published a study linking the cases to an anti-miscarriage drug call diethylstilbestrol, or DES, which was taken by some five million pregnant women between 1938 and 1971. The investigation of medical and occupational cancer clusters has lead to the discovery of dozens of carcinogens, including asbestos, vinyl chloride, and certain artificial dyes.

So why don't hot-pursuit investigations of neighborhood cancer clusters yield such successes? For one thing, many clusters fall apart simply because they violate basic rules of cancer behavior. Cancer develops when a cell starts multiplying out of control, and the process by which this happens isn't straightforward. A carcinogen doesn't just flip some cancer switch to "one." Cells have a variety of genes that keep them functioning normally, and it takes an almost chance combination of successive mutations in these genes – multiple "hits," as cancer biologists put it – to make a cell cancerous rather than simply killing it. A carcinogen provides one hit. Other hits may come from a genetic defect, a further environmental exposure, a spontaneous mutation. Even when people have been subjected to a heavy does of a carcinogen and many cells have been damaged, they will not all get cancer. (For example, DES causes clear-cell adenocarcinoma in only one out of a thousand women exposed to it in utero.) As a rule, it takes a long time before a cell receives enough hits to produce the cancer, and so, unlike infections or acute toxic reactions, the effect of a carcinogen in a community won't be seen for years. Besides, in a mobile society like ours, cancer victims who seem to be clustered may not all have lived in an area long enough for their cancers to have a common cause.

To produce a cancer cluster, a carcinogen has to hit a great many cells in a great many people. A brief, low-level exposure to a carcinogen is unlikely to do the job. Raymond Richard Neutra has calculated that for a carcinogen to produce a sevenfold increase in the occurrence of a cancer (a rate of increase not considered particularly high by epidemiologists) a population would have to be exposed to seventy per cent of the maximum tolerated dose in the course of a full year, or the equivalent. "This kind of exposure is credible as part of chemotherapy or in some work settings," he wrote in a 1990 paper," but it must be very rare for most neighborhood and school settings." For

that reason, investigations of occupational cancer clusters have been vastly more successful than investigations of residential cancer clusters.

Matters are further complicated by the fact that cancer isn't one disease. What turns a breast cell into breast cancer isn't what turns a white blood cell into leukemia: the precise combination of hits varies. Yet some clusters lump together people with tumors that have entirely different biologies and are unlikely to have the same cause. The cluster in McFarland, for example, involved eleven children with nine kinds of cancer. Some of the brain-cancer cases in the Los Alamos cluster were really cancers of other organs which had metastasized to the brain.

If true neighborhood clusters – that is, local clusters arising from a common environmental cause – are so rare, why do we see so many? In a sense, we're programmed to: nearly all of them are the result of almost irresistible errors in perception. In a pioneering article published in 1971, the cognitive psychologists Daniel Kahneman and Amos Tversky identified a systematic error in human judgment which they called the Belief in the Law of Small Numbers. People assume that the pattern of a large population will be replicated in all its subsets. But clusters will occur simply through chance. After seeing a long sequence of red on the roulette wheel, people find it hard to resist the idea that black is "due" – or else they start to wonder whether the wheel is rigged. We assume that a sequence of R-R-R-R-R-R is somehow less random than, say, R-R-B-R-B-B. But the two sequences are equally likely. (Casinos make a lot of money from the Belief in the Law of Small Numbers.) Truly random patterns often doesn't appear random to us. The statistician William Feller studied one classic example. During the Germans' intensive bombing of South London in the second World War, a few areas were hit several times and others were not hit at all. The places that were not hit seemed to have been deliberately spared, and, Kahneman says, people became convinced that those places were where the Germans had their spies. When Feller analyzed the statistics of the bomb hits, however, he found that the distribution matched a random pattern.

Daniel Kahneman himself was involved in a similar case. "During the Yom Kippur War, in 1973, I was approached by people in the Israeli Air Force,"

he told me. "They had two squads that had left base, and when the squads came back one had lost four planes and the other had lost none. They wanted to investigate for all kinds of differences between the squadrons, like whether pilots in one squadron had seen their wives more than in the other. I told them to stop wasting their time." A difference of four lost planes could easily have occurred by chance. Yet Kahneman knew that if Air Force officials investigated they would inevitably find some measurable differences between the squadrons and feel compelled to act on them.

Human beings evidently have a deep-seated tendency to see meaning in the ordinary variations that are bound to appear in small samples. For example, most basketball players and fans believe that players have hot and cold streaks in shooting. In a paper entitled "The Hot Hand in Basketball," Tversky and two colleagues painstakingly analyzed the shooting of individual players in more than eighty games played by the Philadelphia 76ers, the New Jersey Nets, and the New York Knicks during the 1980–81 season. It turned out that basketball players – even notorious "streak shooters" – have no more runs of hits or misses than would be expected by chance. Because of the human tendency to perceive clusters in random sequences, however, Tversky and his colleagues found that "no amount of exposure to such sequences will convince the player, the coach, or the fan that the sequences are in fact random. The more basketball one watches and plays, the more opportunities one has to observe what appears to be streak shooting."

In epidemiology, the tendency to isolate clusters from their context is known as the Texas-sharpshooter fallacy. Like a Texas sharpshooter who shoots at the side of a barn and then draws a bull's eye around the bullet holes, we tend to notice cases first – four cancer patients on one street – and then define the population base around them. With rare conditions, such as Haff disease or mercury poisoning, even a small clutch of cases really would represent a dramatic excess, no matter how much Texas sharpshooting we did. But most cancers are common enough that noticeable residential clusters are bound to occur. Raymond Richard Neutra points out that, given a typical registry of eight different cancers, you could expect twenty-seven hundred and fifty of California's five thousand census tracts to have statistically significant but perfectly random elevations of cancer. So if you check to see whether your

neighborhood has an elevated rate of a specific cancer, chances are better than even that it does – and it almost certainly won't mean a thing. Even when you've established a correlation between a specific cancer and a potential carcinogen, scientists have hardly any way to distinguish the "true" cancer cluster that's worth investigating from the crowd of cluster impostors.

One helpful tip-off is an extraordinarily high cancer rate. In Karain, Turkey, the incidence of mesothelioma was more than *seven thousand times* as high as expected. In even the most serious cluster alarms that public-health departments have received, however, the cancer rate has been nowhere near that high. (The lawyer Jan Schlichtmann, of "Civil Action" fame, is now representing victims of a cancer cluster in Dover Township, New Jersey, where the childhood-cancer rate is thirty per cent higher than expected.)

This isn't to say that carcinogens in the local environment can't raise cancer rates; it's just that such increases disappear in all the background variation that occurs in small populations. In larger populations, it's a different story. The 1986 Chernobyl disaster exposed hundreds of thousands of people to radiation; scientists were able to establish that it caused a more than one-hundred-fold increase in thyroid cancer among children years later. By contrast, investigating an isolated neighborhood cancer cluster is almost always a futile exercise. Investigators knock on doors, track down former residents, and check medical records. They sample air, soil, and water. Thousands, sometimes millions, of dollars are spent. And, with all those tests, correlations inevitably turn up. Yet, years later, in case after case, nothing definite is confirmed.

"The reality is that they're an absolute, total, and complete waste of taxpayer dollars," says Alan Bender, an epidemiologist with the Minnesota Department of Health, which investigated more than a thousand cancer clusters in the state between 1984 and 1995. The problem of perception and politics, however, remains. If you're a public health official, try explaining why a dozen children with cancer in one neighborhood doesn't warrant investigation. According to a national study, health departments have been able to reassure people by education in more than seventy per cent of cluster alarms. Somewhere between one and three per cent of alarms, however, re-

sult in expensive on-site investigations. And the cases that are investigated aren't even the best-grounded ones: they are the cases pushed by the media, enraged citizens, or politicians. "Look, you can't just kiss people off," Bender says. In fact, Minnesota has built such an effective public-response apparatus that is has not needed to conduct a formal cluster investigation in three years.

Public-health departments aren't lavishly funded, and scientists are reluctant to see money spent on something that has proved to be as unproductive as neighborhood cluster alarms or cancer mapping. Still, public confidence is poorly served by officials who respond to inquiries with a scientific brushoff and a layer of bureaucracy. To be part of a cancer cluster is a frightening thing, and it magnifies our ordinary response when cancer strikes: we want to hold something or someone responsible, even allocate blame. Health officials who understand the fear and anger can have impressive success, as the ones in Minnesota have shown. But there are times when you cannot maintain public trust without acting on public concerns. Science alone won't put to rest questions like the one a McFarland mother posed to the Los Angeles *Times*: "How many more of our children must die before something is done?"

### 0.0.6 Duped, Margaret Talbot (*New Yorker*), July 2, 2007

July 2, 2007

Margaret Talbot (*New Yorker*)

Can brain scans uncover lies?

In theory, a markedly improved method of lie detection could have as profound an impact as DNA evidence.

The most egregious liar I ever knew was someone I never suspected until the day that, suddenly and irrevocably, I did. Twelve years ago, a young man named Stephen Glass began writing for *New Republic*, where I was an editor. He quickly established himself as someone who was always onto an amusingly outlandish story – like the time he met some Young Republican types at a convention, gathered them around a hotel-room minibar, then, with guileless ferocity, captured their boorishness in print. I liked Steve; most of us who worked with him did. A baby-faced guy from suburban Chicago, he padded around the office in his socks. Before going on an errand, Steve would ask if I wanted a muffin or a sandwich; he always noticed a new scarf or a clever turn of phrase, and asked after a colleague's baby or spouse. When he met with editors to talk about his latest reporting triumph, he was self-effacing and sincere. He'd look us in the eye, wait for us to press him for details, and then, without fidgeting or mumbling, supply them.

One day, the magazine published an article by Steve about a teen-ager so diabolically gifted at hacking into corporate computer networks that C.E.O.s paid him huge sums just to stop messing with them. A reporter for the online edition of *Forbes* was assigned to chase down the story. You can see how Steve's journalism career unraveled if you watch the movie "Shattered Glass": *Forbes* challenged the story's veracity, and Steve – after denying the charges, concocting a fake Web site, and enlisting his brother to pose as a victimized C.E.O. – finally confessed that he'd made up the whole thing. Editors and reporters at the magazine investigated, and found that Steve had been inventing stories for at least a year. The magazine disavowed twenty-seven articles.

After Steve's unmasking, my colleagues and I felt ashamed of our gullibility. But maybe we shouldn't have. Human beings are terrible lie detectors.

In academic studies, subjects asked to distinguish truth from lies answer correctly, on average, fifty-four per cent of the time. They are better at guessing when they are being told the truth than when they are being lied to, accurately classifying only forty-seven per cent of lies, according to a recent meta-analysis of some two hundred deception studies, published by Bella DePaulo, of the University of California at Santa Barbara, and Charles Bond, Jr., of Texas Christian University. Subjects are often led astray by an erroneous sense of how a liar behaves. "People hold a stereotype of the liar – as tormented, anxious, and conscience-stricken," DePaulo and Bond write. (The idea that a liar's anxiety will inevitably become manifest can be found as far back as the ancient Greeks, Demosthenes in particular.) In fact, many liars experience what deception researchers call "duping delight."

Aldert Vrij, a psychologist at the University of Portsmouth, in England, argues that there is no such thing as "typical" deceptive behavior – "nothing as obvious as Pinocchio's growing nose." When people tell complicated lies, they frequently pause longer and more often, and speak more slowly; but if the lie is simple, or highly polished, they tend to do the opposite. Clumsy deceivers are sometimes visibly agitated, but, over all, liars are less likely to blink, to move their hands and feet, or to make elaborate gestures – perhaps they deliberately inhibit their movements. As DePaulo says, "To be a good liar, you don't need to know what behaviors really separate liars from truthtellers, but what behaviors people think separate them."

A liar's testimony is often more persuasive than a truthteller's. Liars are more likely to tell a story in chronological order, whereas honest people often present accounts in an improvised jumble. Similarly, according to DePaulo and Bond, subjects who spontaneously corrected themselves, or said that there were details that they couldn't recall, were more likely to be truthful than those who did not – though, in the real world, memory lapses arouse suspicion.

People who are afraid of being disbelieved, even when they are telling the truth, may well look more nervous than people who are lying. This is bad news for the falsely accused, especially given that influential manuals of interrogation reinforce the myth of the twitchy liar. "Criminal Interrogation and Confessions" (1986), by Fred Inbau, John Reid, and Joseph Buckley, claims

that shifts in posture and nervous "grooming gestures," such as "straightening hair" and "picking lint from clothing," often signal lying. David Zulawski and Douglas Wicklander's "Practical Aspects of Interview and Interrogation" (1992) asserts that a liar's movements tend to be "jerky and abrupt" and his hands "cold and clammy." Bunching Kleenex in a sweaty hand is another damning sign – one more reason for a sweaty-palmed, Kleenex-bunching person like me to hope that she's never interrogated.

Maureen O'Sullivan, a deception researcher at the University of San Francisco, studies why humans are so bad at recognizing lies. Many people, she says, base assessments of truthfulness on irrelevant factors, such as personality or appearance. "Baby-faced, non-weird, and extroverted people are more likely to be judged truthful," she says. (Maybe this explains my trust in Steve Glass.) People are also blinkered by the "truthfulness bias": the vast majority of questions we ask of other people – the time, the price of the breakfast special – are answered honestly, and truth is therefore our default expectation. Then, there's the "learning-curve problem." We don't have a refined idea of what a successful lie looks and sounds like, since we almost never receive feedback on the fibs that we've been told; the co-worker who, at the corporate retreat, assured you that she loved your presentation doesn't usually reveal later that she hated it. As O'Sullivan puts it, "By definition, the most convincing lies go undetected."

Maybe it's because we're such poor lie detectors that we have kept alive the dream of a foolproof lie-detecting machine. This February, at a conference on deception research, in Cambridge, Massachusetts, Steven Hyman, a psychiatrist and the provost of Harvard, spoke of "the incredible hunger to have some test that separates truth from deception – in some sense, the science be damned."

This hunger has kept the polygraph, for example, in widespread use. The federal government still performs tens of thousands of polygraph tests a year – even though an exhaustive 2003 National Academy of Sciences report concluded that research on the polygraph's efficacy was inadequate, and that when it was used to investigate a specific incident after the fact it performed "well above chance, though well below perfection." Polygraph advocates cite accuracy estimates of ninety per cent – which sounds impressive until

you think of the people whose lives might be ruined by a machine that fails one out of ten times. The polygraph was judged thoroughly unreliable as a screening tool; its accuracy in "distinguishing actual or potential security violators from innocent test takers" was deemed "insufficient to justify reliance on its use." And its success in criminal investigations can be credited, in no small part, to the intimidation factor. People who believe that they are in the presence of an infallible machine sometimes confess, and this is counted as an achievement of the polygraph. (According to law-enforcement lore, the police have used copy machines in much the same way: They tell a suspect to place his hand on a "truth machine" – a copier in which the paper has "LIE " printed on it. When the photocopy emerges, it shows the suspect's hand with "LIE " stamped on it.)

Over the past two decades, inventors have attempted to supplant the polygraph with new technologies: voice-stress analysis; thermal imaging of the face; and, most recently and spectacularly, brain imaging. Though these methods remain in an embryonic stage of development, they have already been greeted with considerable enthusiasm, especially in America. Private companies are eager to replace traditional modes of ascertaining the truth – such as the jury system – with a machine that can be patented and sold. And law-enforcement agencies yearn to overcome the problem of suspects who often remain maddeningly opaque, even in the face of sustained interrogation. Although one immediate result of the September 11th attacks was the revival of an older, and even more controversial, form of interrogation – torture – the war on terror has also inflamed the desire for a mind-reading machine.

Not long ago, I met with an entrepreneur named Joel Huizenga, who has started a company, based in San Diego, called No Lie MRI. Most methods of lie detection look at the activity of the sympathetic nervous system. The polygraph, for instance, is essentially an instrument for measuring stress. Heart and respiration rates, blood volume, and galvanic skin response – a proxy for palm sweat – are represented as tracings on graph paper or on a screen, which fluctuate with every heartbeat or breath. The method that Huizenga is marketing, which employs a form of body scanning known as functional magnetic resonance imaging, or fMRI, promises to look inside the brain. "Once you jump behind the skull, there's no hiding," Huizenga told

me.

Functional MRI technology, invented in the early nineties, has been used primarily as a diagnostic tool for identifying neurological disorders and for mapping the brain. Unlike MRIs, which capture a static image, an fMRI makes a series of scans that show changes in the flow of oxygenated blood preceding neural events. The brain needs oxygen to perform mental tasks, so a rise in the level of oxygenated blood in one part of the brain can indicate cognitive activity there. (Blood has different magnetic properties when it is oxygenated, which is why it is helpful to have a machine that is essentially a big magnet.) Brain-scan lie detection is predicated on the idea that lying requires more cognitive effort, and therefore more oxygenated blood, than truth-telling.

Brain scanning promises to show us directly what the polygraph showed us obliquely. Huizenga expects his company to be a force for justice, exonerating customers who are, as he put it, "good people trying to push back the cruel world that is indicting them unfairly." Brain scans already have clout in the courtroom; during death-penalty hearings, judges often allow images suggesting neurological impairment to be introduced as mitigating evidence. In theory, an improved method of lie detection could have as profound an impact as DNA evidence, which has freed more than a hundred wrongly accused people since its introduction, in the late eighties. If Huizenga has perfected such a technology, he's onto something big.

At Huizenga's suggestion, we met at a restaurant called the Rusty Pelican, on the Pacific Coast Highway, in Newport Beach. A television screen on one wall showed a surfing contest; Huizenga, who is fifty-three, with dirty-blond hair in a boyish cut, is a surfer himself. He has a bachelor's degree from the University of Colorado, a master's degree in biology from Stony Brook, and an M.B.A. from the University of Rochester. No Lie is Huizenga's second startup. The first, ISCHEM Corporation, uses body scanning to look for plaque in people's arteries. Before that, he worked for Pantox, a company that offers blood tests to gauge a person's antioxidant levels.

After we sat down, Huizenga recounted the origins of No Lie. A few years ago, he came across an item in the *Times* about some tantalizing research conducted by Daniel Langleben, a psychiatrist and neuroscientist at the Uni-

versity of Pennsylvania. Subjects were placed inside an fMRI machine and told to make some true statements and some false ones. Brain scans taken while the subjects were lying frequently showed a significantly increased level of activity in three discrete areas of the cerebral cortex. Langleben suggested that "intentional deception" could be "anatomically localized" by fMRI scanning. Huizenga immediately saw a business opportunity. "I jumped on it," he told me. "If I wasn't here sitting in front of you, somebody else would be."

The Web site for No Lie claims that its technology, which is based on the Penn protocol, "represents the first and only direct measure of truth verification and lie detection in human history!" No Lie just started offering tests commercially, and has charged about a dozen clients approximately ten thousand dollars apiece for an examination. (No Lie sends customers to an independent imaging center in Tarzana, a suburb of Los Angeles, to insure that "quality testing occurs according to standardized test protocols.") Some of these initial clients are involved in civil and criminal cases; the first person to use the service, Harvey Nathan, was accused in 2003 of deliberately setting fire to a deli that he owns in South Carolina. A judge dismissed the charges, but Nathan wanted to bring suit against his insurance company, and he thought that documented evidence of his innocence would further his cause. So in December he flew to California and took No Lie's test. He passed. Nathan said, "If I hadn't, I would have jumped from the seventeenth floor of the hotel where I was staying. How could I have gone back to South Carolina and said, 'Oh that machine must not have worked right'? I believed in it then and I believe in it now." Nathan's exam was filmed for the Discovery Channel, which may soon launch a reality series centering on brain-scanning lie detection.

Several companies have expressed interest in No Lie's services, Huizenga told me. (He would not name them.) He said that he will be able to accommodate corporate clients once he has signed deals with other scanning facilities; he is in talks with imaging centers in a dozen cities, including New York and Chicago. No Lie also plans to open a branch in Switzerland later this year.

Huizenga has been criticized for his company's name, but he said, "It's not

about being dignified – it's about being remembered." He believes that the market for fMRI-based lie detection will one day exceed that of the polygraph industry, which brings in hundreds of millions of dollars annually. Investment analysts say that it is too soon to judge if Huizenga's optimism is warranted, but No Lie has attracted some prominent backing. One of its prime investors is Alex Hart, the former C.E.O. of MasterCard International, who is also serving as a management consultant. And it has a "scientific board" consisting of four paid advisers, among them Terrence Sejnowski, the director of the Crick-Jacobs Center for theoretical and computational biology at the Salk Institute. In an e-mail, Sejnowski explained that he offers counsel on "advanced signal processing and machine-learning techniques that can help improve the analysis of the data and the accuracy of the performance." He said of No Lie, "The demand is there, and to succeed as a company the new technology only needs to be better than existing approaches."

Huizenga speaks of his company's goals in blunt terms. "What do people lie about?" he asked me. "Sex, power, and money – probably in that order." (The company's Web site recommends No Lie's services for "risk reduction in dating," "trust issues in interpersonal relationships," and "issues concerning the underlying topics of sex, power, and money.") "Parents say, 'Yes, this is perfect for adolescents,'" he went on. "People who are dating say, 'Yes, this is great for dating, because people never tell you the truth.'"

He said that his company receives dozens of inquiries a week: from divorcing men accused of child abuse; from women wanting to prove their fidelity to jealous spouses or boyfriends; from people representing governments in Africa and the former Soviet republics; from "the Chinese police department." He said that he understood why governments were interested in lie-detection technology. "Look at Joe Stalin," he said. "Joe wanted power, he wanted to be on top. Well, it's hard to murder massive numbers of opponents. People in our government, and in others', need more effective ways of weeding out those who aren't their puppets." Some potential foreign clients had explained to him, he said, that in societies that lacked "civilization, there is not trust, and lie detection could help build that trust." (He wasn't sure about that – he was "mulling it over.") Huizenga said that the United States government was "interested" in the kind of technology offered by No Lie; the company

has hired Joel S. Lisker, a former F.B.I. agent, to be its "sales liaison for the federal government." (Lisker declined to be interviewed, saying that his government contacts were "confidential.")

The Pentagon has supported research into high-tech lie detection, including the use of fMRI. The major scientific papers in the field were funded, in part, by the Defense Advanced Research Projects Agency, which develops new technologies for military use, and by the Department of Defense Polygraph Institute, which trains lie-detection experts at Fort Jackson, South Carolina. (The Polygraph Institute underwent a name change in January – it's now the Defense Academy for Credibility Assessment – apparently in deference to new technologies such as fMRI.) Last June, the A.C.L.U. filed several Freedom of Information Act requests in an attempt to learn more about the government's involvement with the technology. Chris Calabrese, an A.C.L.U. lawyer, said that the C.I.A. would neither "confirm nor deny" that it is investigating fMRI applications; the Pentagon produced PowerPoint presentations identifying brain scans as a promising new technology for lie detection. Calabrese went on, "We were motivated by the fact that there are companies trying to sell this technology to the government. This Administration has a history of using questionable techniques of truth verification."

Many scholars also think that Huizenga's effort is premature. Steven Hyman, the Harvard professor, told me that No Lie was "foolish." But the history of lie-detection machines suggests that it would be equally foolish to assume that a few scholarly critics can forestall the adoption of such a seductive new technology. "People are drawn to it," Huizenga said, smiling. "It's a magnetic concept."

In comic books of the nineteen-forties, Wonder Woman, the sexy Amazon superhero, wields a golden "lasso of truth." Anybody she captures is rendered temporarily incapable of lying. Like the golden lasso, the polygraph, its inventors believed, compelled the body to reveal the mind's secrets. But the connection between the lasso and the lie detector is even more direct than that: Wonder Woman's creator, William Moulton Marston, was also a key figure in the development of the polygraph. Marston, like other pioneers of lie detection, believed that the conscious mind could be circumvented, and the truth uncovered, through the measurement of bodily signals.

This was not a new idea. In 1730, Daniel Defoe published "An Effectual Scheme for the Immediate Preventing of Street Robberies and Suppressing All Other Disorders of the Night," in which he proposed an alternative to physical coercion: "Guilt carries fear always about with it, there is a tremor in the blood of a thief, that, if attended to, would effectually discover him; and if charged as a suspicious fellow, on that suspicion only I would feel his pulse."

In the late nineteenth century, the Italian criminologist Cesare Lombroso invented his own version of a lie detector, based on the physiology of emotion. A suspect was told to plunge his hand into a tank filled with water, and the subject's pulse would cause the level of liquid to rise and fall slightly; the greater the fluctuation, the more dishonest the subject was judged to be.

Lombroso's student Angelo Mosso, a physiologist, noticed that shifts in emotion were often detectable in fair-skinned people in the flushing or blanching of their faces. Based on this observation, he designed a bed that rested on a fulcrum. If a suspect reclining on it told a lie, Mosso hypothesized, resulting changes in blood flow would alter the distribution of weight on the bed, unbalancing it. The device, known as Mosso's cradle, apparently never made it past the prototype.

William Moulton Marston was born in 1893, in Boston. He attended Harvard, where he worked in the lab of Hugo Münsterberg, a German émigré psychologist, who had been tinkering with an apparatus that registered responses to emotions, such as horror and tenderness, through graphical tracing of pulse rates. One student volunteer was Gertrude Stein. (She later wrote of the experience in the third person: "Strange fancies begin to crowd upon her, she feels that the silent pen is writing on and on forever.")

In 1917, Marston published a paper arguing that systolic blood pressure could be monitored to detect deception. As Ken Alder, a history professor at Northwestern, notes in his recent book, "The Lie Detectors: The History of an American Obsession," Münsterberg and Marston's line of inquiry caught the imagination of police detectives, reporters, and law-enforcement reformers across the country, who saw a lie-detecting machine as an alternative not only to the brutal interrogation known as the third degree but also to the jury system. In 1911, an article in the *Times* predicted a future in which "there

will be no jury, no horde of detectives and witnesses, no charges and countercharges, and no attorney for the defense. These impediments of our courts will be unnecessary. The State will merely submit all suspects in a case to the tests of scientific instruments." John Larson, a police officer in Berkeley, California, who also had a doctorate in physiology, expanded on Marston's work. He built an unwieldy device, the "cardio-pneumo-psychograph," which used a standard cuff to measure blood pressure, and a rubber hose wrapped around the subject's chest to measure his breathing. Subjects were told to answer yes-or-no questions; their physiological responses were recorded by styluses that scratched black recording paper on revolving drums.

In 1921, as Alder writes, Larson had his first big chance to test his device. He was seeking to identify a thief at a residence hall for female students at Berkeley. Larson gave several suspects a six-minute exam, in which he asked various questions: "How much is thirty times forty?" "Will you graduate this year?" "Do you dance?" "Did you steal the money?" The result foretold the way in which a polygraph would often "work": as a goad to confession. A student nurse confessed to the crime – a few days after she'd stormed out during the exam.

In the early twenties, another member of the Berkeley police force, Leonarde Keeler, increased the number of physical signs that the lie detector monitored. His portable machine recorded pulse rate, blood pressure, respiration, and "electrodermal response" – again, palm sweat. Today's lie detector looks much like Keeler's eighty-year-old invention. And it bears the same name: the polygraph.

Polygraphs never caught on in Europe. But here their advent coincided with the Prohibition-era crime wave; with a new fascination with the unconscious (this was also the era of experimentation with so-called truth serums); and with the wave of technological innovation that had brought Americans electricity, radios, telephones, and cars. The lie detector quickly insinuated itself into American law enforcement: at the end of the thirties, a survey of thirteen city police departments showed that they had given polygraphs to nearly nine thousand suspects.

In 1923, Marston tried without success to get a polygraph test introduced as evidence in the Washington, D.C., murder trial of James Alphonso Frye.

In its ruling, the Court of Appeals for the D.C. Circuit declared that a new scientific method had to have won "general acceptance" from experts before judges could give it credence. Since this decision, the polygraph has been kept out of most courtrooms, but there is an important exception: about half the states allow a defendant to take the test, generally on the understanding that the charges will be dropped if he passes and the results may be entered as evidence if he fails.

The polygraph became widely used in government and in business, often with dubious results. In the fifties, the State Department deployed the lie detector to help purge suspected homosexuals. As late as the seventies, a quarter of American corporations used the polygraph on their employees. Although Congress banned most such tests when it passed the Polygraph Protection Act, in 1988, the federal government still uses the polygraph for security screenings – despite high-profile mistakes. The polygraph failed to cast suspicion on Aldrich Ames, the C.I.A. agent who spied for the Soviets, and wrongly implicated Wen Ho Lee, the Department of Energy scientist, as an agent of the Chinese government.

One excellent way to gauge the polygraph's effectiveness would be to compare it with an equally intimidating fake machine, just as a drug is compared with a placebo. But, strangely, no such experiment has ever been performed. In 1917, the year that Marston published his first paper on lie detection, his research encountered strong skepticism. John F. Shepard, a psychologist at the University of Michigan, wrote a review of Marston's research. Though the physical changes that the machine measured were "an index of activity," Shepard wrote, the same results "would be caused by so many different circumstances, anything demanding equal activity (intelligence or emotional)." The same criticism holds true today. All the physiological responses measured by the polygraph have causes other than lying, vary greatly among individuals, and can be affected by conscious effort. Breathing is particularly easy to regulate. Advice on how to beat the lie detector is a cottage industry. "Deception Detection: Winning the Polygraph Game" (1991) warns potential subjects, "Don't complain about a dry mouth. An examiner will interpret this as fear of being found out and will press you even harder." (Many people do get dry-mouthed when they're nervous – which is appar-

ently why, during the Inquisition, a suspect was sometimes made to swallow a piece of bread and cheese: if it stuck in his throat, he was deemed guilty.) Other well-known "countermeasures" include taking a mild sedative; using mental imagery to calm yourself; and biting your tongue to make yourself seem anxious in response to random questions.

Why, then, is the polygraph still used? Perhaps the most vexing thing about the device is that, for all its flaws, it's not pure hokum: a racing pulse and an increased heart rate can indicate guilt. Every liar has felt an involuntary flutter, at least once. Yet there are enough exceptions to insure that the polygraph will identify some innocent people as guilty and some guilty people as innocent.

At the Cambridge conference, Jed S. Rakoff, a United States district judge in New York, told a story about a polygraph and a false confession. Days after September 11th, an Egyptian graduate student named Abdallah Higazy came to the attention of the F.B.I. Higazy had been staying at the Millennium Hotel near Ground Zero on the day of the attacks. A hotel security guard claimed that he had found a pilot's radio in Higazy's room. Higazy said that it wasn't his, and when he appeared before Rakoff he asked to be given a polygraph. As Rakoff recalled, "Higazy very much believed in them and thought it would exonerate him." During a four-hour interrogation by an F.B.I. polygrapher, Higazy first repeated that he knew nothing about the radio, and then said that maybe it was his. He was charged with lying to the F.B.I. and went to prison. Within a month, a pilot stopped by the hotel to ask about a radio that he had accidentally left there. The security guard who found the radio admitted that it hadn't been in Higazy's room; he was prosecuted and pled guilty. Higazy was exonerated, and a subsequent investigation revealed that he had felt dizzy and ill during the examination, probably out of nervousness. But when Higazy asked the polygrapher if anyone had ever become ill during a polygraph test he was told that "it had not happened to anyone who told the truth."

To date, there have been only a dozen or so peer-reviewed studies that attempt to catch lies with fMRI technology, and most of them involved fewer than twenty people. Nevertheless, the idea has inspired a torrent of media attention, because scientific studies involving brain scans dazzle people, and

because mind reading by machine is a beloved science-fiction trope, revived most recently in movies like "Minority Report" and "Eternal Sunshine of the Spotless Mind." Many journalistic accounts of the new technology – accompanied by colorful bitmapped images of the brain in action – resemble science fiction themselves. In January, the *Financial Times* proclaimed, "For the first time in history, it is becoming possible to read someone else's mind with the power of science." A CNBC report, accompanied by the Eurythmics song "Would I Lie to You?," showed its reporter entering an fMRI machine, described as a "sure-fire way to identify a liar." In March, a cover story in the *Times Magazine* predicted transformations of the legal system in response to brain imaging; its author, Jeffrey Rosen, suggested that there was a widespread "fear" among legal scholars that "the use of brain-scanning technology as a kind of super mind-reading device will threaten our privacy and mental freedom." *Philadelphia* has declared "the end of the lie," and a *Wired* article, titled "Don't Even Think About Lying," proclaimed that fMRI is "poised to transform the security industry, the judicial system, and our fundamental notions of privacy." Such talk has made brain-scan lie detection sound as solid as DNA evidence – which it most definitely is not.

Paul Bloom, a cognitive psychologist at Yale, believes that brain imaging has a beguiling appeal beyond its actual power to explain mental and emotional states. "Psychologists can be heard grousing that the only way to publish in *Science* or *Nature* is with pretty color pictures of the brain," he wrote in an essay for the magazine *Seed*. "Critical funding decisions, precious column inches, tenure posts, science credibility, and the popular imagination have all been influenced by fMRI's seductive but deceptive grasp on our attentions." Indeed, in the past decade, Nature alone has published nearly a hundred articles involving fMRI scans. The technology is a remarkable tool for exploring the brain, and may one day help scientists understand much more about cognition and emotion. But enthusiasm for brain scans leads people to overestimate the accuracy with which they can pinpoint the sources of complex things like love or altruism, let alone explain them.

Brain scans enthrall us, in part, because they seem more like "real" science than those elaborate deductive experiments that so many psychologists perform. In the same way that an X-ray confirms a bone fissure, a brain scan

seems to offer an objective measure of mental activity. And, as Bloom writes, fMRI research "has all the trappings of work with great lab-cred: big, expensive, and potentially dangerous machines, hospitals and medical centers, and a lot of people in white coats."

Deena Skolnick Weisberg, a graduate student at Yale, has conducted a clever study, to be published in the *Journal of Cognitive Neuroscience*, which points to the outsized glamour of brain-scan research. She and her colleagues provided three groups – neuroscientists, neuroscience students, and ordinary adults – with explanations for common psychological phenomena (such as the tendency to assume that other people know the same things we do). Some of these explanations were crafted to be bad. Weisberg found that all three groups were adept at identifying the bad explanations, except when she inserted the words "Brain scans indicate." Then the students and the regular adults became notably less discerning. Weisberg and her colleagues conclude, "People seem all too ready to accept explanations that allude to neuroscience."

Some bioethicists have been particularly credulous, assuming that MRI mind reading is virtually a done deal, and arguing that there is a need for a whole new field: "neuroethics." Judy Illes and Eric Racine, bioethicists at Stanford, write that fMRI, by laying bare the brain's secrets, may "fundamentally alter the dynamics between personal identity, responsibility, and free will." A recent article in *The American Journal of Bioethics* asserts that brain-scan lie detection may "force a rexamination of the very idea of privacy, which up until now could not reliably penetrate the individual's cranium."

Legal scholars, for their part, have started debating the constitutionality of using brain-imaging evidence in court. At a recent meeting of a National Academy of Sciences committee on lie detection, in Washington, D.C., Hank Greely, a Stanford law professor, said, "When we make speculative leaps like these ... it increases, sometimes in detrimental ways, the belief that the technology works." In the rush of companies like No Lie to market brain scanning, and in the rush of scholars to judge the propriety of using the technology, relatively few people have asked whether fMRIs can actually do what they either hope or fear they can do.

Functional MRI is not the first digital-age breakthrough that was supposed

to supersede the polygraph. First, there was "brain fingerprinting," which is based on the idea that the brain releases a recognizable electric signal when processing a memory. The technique used EEG sensors to try to determine whether a suspect retained memories related to a crime – an image of, say, a murder weapon. In 2001, *Time* named Lawrence Farwell, the developer of brain fingerprinting, one of a hundred innovators who "may be the Picassos or the Einsteins of the 21st century." But researchers have since noted a big drawback: it's impossible to distinguish between brain signals produced by actual memories and those produced by imagined memories – as in a made-up alibi.

After September 11th, another technology was widely touted: thermal imaging, an approach based on the finding that the area around the eyes can heat up when people lie. The developers of this method – Ioannis Pavlidis, James Levine, and Norman Eberhardt – published journal articles that had titles like "Seeing Through the Face of Deception" and were accompanied by dramatic thermal images. But the increased blood flow that raises the temperature around the eyes is just another mark of stress. Any law-enforcement agency that used the technique to spot potential terrorists would also pick up a lot of jangly, harmless travellers.

Daniel Langleben, the Penn psychiatrist whose research underpins No Lie, began exploring this potential new use for MRIs in the late nineties. Langleben, who is forty-five, has spent most of his career studying the brains of heroin addicts and hyperactive boys. He developed a side interest in lying partly because his research agenda made him think about impulse control, and partly because his patients often lied to him. Five years ago, Langleben and a group of Penn colleagues published the study on brain scanning and lie detection that attracted Huizenga's attention. In the experiment, which was written up in *Neuroimage*, each of twenty-three subjects was offered an envelope containing a twenty-dollar bill and a playing card – the five of clubs. They were told that they could keep the money if they could conceal the card's identity when they were asked about it inside an MRI machine. The subjects pushed a button to indicate yes or no as images of playing cards flashed on a screen in front of them. After Langleben assembled the data, he concluded that lying seemed to involve more cognitive effort than

truth-telling, and that three areas of the brain generally became more active during acts of deception: the anterior cingulate cortex, which is associated with heightened attention and error monitoring; the dorsal lateral prefontal cortex, which is involved in behavioral control; and the parietal cortex, which helps process sensory input. Three years later, Langleben and his colleagues published another study, again involving concealed playing cards, which suggested that lying could be differentiated from truthtelling in individuals as well as in groups. The fMRI's accuracy rate for distinguishing truth from lies was seventy-seven per cent.

Andrew Kozel and Mark George, then at the Medical University of South Carolina, were doing similar work at the time; in 2005, they published a study of fMRI lie detection in which thirty people were instructed to enter a room and take either a watch or a ring that had been placed there. Then, inside a scanner, they were asked to lie about which object they had taken but to answer truthfully to neutral questions, such as "Do you like chocolate?" The researchers distinguished truthful from deceptive responses in ninety per cent of the cases. (Curiously, Kozel's team found that liars had heightened activity in different areas of the brain than Langleben did.)

Langleben and Kozel weren't capturing a single, crisp image of the brain processing a lie; an fMRI's record of a split-second event is considered unreliable. Instead, they asked a subject to repeat his answer dozens of times while the researchers took brain scans every couple of seconds. A computer then counted the number of "voxels" (the 3-D version of pixels) in the brain image that reflected a relatively high level of oxygenated blood, and used algorithms to determine whether this elevated activity mapped onto specific regions of the brain.

One problem of fMRI lie detection is that the machines, which cost about three million dollars each, are notoriously finicky. Technicians say that the scanners often have "bad days," in which they can produce garbage data. And a subject who squirms too much in the scanner can invalidate the results. (Even moving your tongue in your mouth can cause a problem.) The results for four of the twenty-three subjects in Langleben's first study had to be thrown out because the subjects had fidgeted.

The Langleben studies also had a major flaw in their design: the concealed

playing card came up only occasionally on the screen, so the increased brain activity that the scans showed could have been a result not of deception but of heightened attention to the salient card. Imagine that you're the research subject: You're lying on your back, trying to hold still, probably bored, maybe half asleep, looking at hundreds of cards that don't concern you. Then, at last, up pops the five of clubs – and your brain sparks with recognition.

Nearly all the volunteers for Langleben's studies were Penn students or members of the academic community. There were no sociopaths or psychopaths; no one on antidepressants or other psychiatric medication; no one addicted to alcohol or drugs; no one with a criminal record; no one mentally retarded. These allegedly seminal studies look exclusively at unproblematic, intelligent people who were instructed to lie about trivial matters in which they had little stake. An incentive of twenty dollars can hardly be compared with, say, your freedom, reputation, children, or marriage – any or all of which might be at risk in an actual lie-detection scenario.

The word "lie" is so broad that it's hard to imagine that any test, even one that probes the brain, could detect all forms of deceit: small, polite lies; big, brazen, self-aggrandizing lies; lies to protect or enchant our children; lies that we don't really acknowledge to ourselves as lies; complicated alibis that we spend days rehearsing. Certainly, it's hard to imagine that all these lies will bear the identical neural signature. In their degrees of sophistication and detail, their moral weight, their emotional valence, lies are as varied as the people who tell them. As Montaigne wrote, "The reverse side of the truth has a hundred thousand shapes and no defined limits."

Langleben acknowledges that his research is not quite the breakthrough that the media hype has suggested. "There are many questions that need to be looked into before we know whether this will work as lie detection," he told me. "Can you do this with somebody who has an I.Q. of ninety-five? Can you do it with somebody who's fifty or older? Somebody who's brain-injured? What kinds of real crimes could you ask about? What about countermeasures? What about people with delusions?"

Nevertheless, the University of Pennsylvania licensed the pending patents on his research to No Lie in 2003, in exchange for an equity position in the

company. Langleben didn't protest. As he explained to me, "It's good for your résumé. We're encouraged to have, as part of our portfolio, industry collaborations." He went on, "I was trying to be a good boy. I had an idea. I went to the Center of Technology Transfer and asked them, 'Do you like this?' They said, 'Yeah, we like that.' "

Steven Laken is the C.E.O. of Cephos, a Boston-based company that is developing a lie-detection product based on Kozel's watch-and-ring study. (It has an exclusive licensing agreement for pending patents that the Medical University of South Carolina applied for in 2002.) Cephos is proceeding more cautiously than No Lie. Laken's company is still conducting studies with Kozel, the latest of which involve more than a hundred people. (The sample pool is again young, healthy, and free of criminal records and psychological problems.) Cephos won't be offering fMRIs commercially until the results of those studies are in; Laken predicts that this will happen within a year. At the National Academy of Sciences committee meeting, he said, "I can say we're not at ninety-per-cent accuracy. And I have said, if we were not going to get to ninety per cent, we're not going to sell this product." (Nobody involved in fMRI lie detection seems troubled by a ten-per-cent error rate.)

In March, I went to a suburb of Boston to meet Laken. He is thirty-five years old and has a Ph.D. in cellular and molecular medicine from Johns Hopkins. Nine years ago, he identified a genetic mutation that can lead to colorectal cancer. He has a more conservative temperament than Joel Huizenga does, and he told me he thinks that spousal-fidelity cases are "sleazy." But he sees a huge potential market for what he calls a "truth verifier" – a service for people looking to exonerate themselves. "There are some thirty-five million criminal and civil cases filed in the U.S. every year," Laken said. "About twenty million are criminal cases. So let's just say that you never even do a criminal case – well, that still leaves roughly fifteen million for us to go after. Some you exclude, but you end up with several million cases that are high stakes: two people arguing about things that are important." Laken also thinks that fMRI lie detection could help the government elicit information, and confessions, from terrorist suspects, without physical coercion.

He calmly dismissed the suggestion that the application of fMRI lie detection is premature. "I've heard it said, 'This technology can't work because

it hasn't been tested on psychopaths, and it hasn't been tested on children, and it certainly hasn't been tested on psychopathic children,' " he said. "If that were the standard, there'd never be any medicine."

Laken and I spoke while driving to Framingham, Massachusetts, to visit an MRI testing center run by Shields, a company that operates twenty-two such facilities in the state. Laken was working on a deal with Shields to use their scanners. For Shields, it would be a smart move, Laken said, because customers would pay up front for the scan – there would be no insurance companies to contend with. (Cephos and Shields have since made an official arrangement.) Laken believes that Cephos will prosper primarily through referrals: lawyers will function as middlemen, ordering an fMRI for a client, much as a doctor orders an MRI for a patient.

We pulled into the parking lot, where a sign identifies Shields as "the MRI provider for the 3-X World Champion New England Patriots." Inside, John Cannillo, an imaging specialist at Shields, led us into a room to observe a woman undergoing an MRI exam. She lay on a platform that slid into a white tubular scanner, which hummed like a giant tuning fork.

During a brain scan, the patient wears a copper head coil, in order to enhance the magnetic field around the skull. The magnet is so powerful that you have to remove any metal objects, or you will feel a tugging sensation. If a person has metal in his body – for instance, shrapnel, or the gold grillwork that some hip-hop fans have bonded to their teeth – it can pose a danger or invalidate the results. At the N.A.S. meeting in Washington, one scientist wryly commented, "It could become a whole new industry – criminals having implants put in to avoid scanning."

A Shields technician instructed the woman in the scanner from the other side of a glass divide. "Take a breath in, and hold it, hold it," he said. Such exercises help minimize a patient's movements.

As we watched, Laken admitted that "the kinks" haven't been worked out of fMRI lie detection. "We make mistakes," he said of his company. "We don't know why we make mistakes. We may never know why. We hope we can get better." Some bioethicists and journalists may worry about the far-off threat to "cognitive freedom," but the real threat is simpler and more immediate: the commercial introduction of high-tech "truth verifiers" that

may work no better than polygraphs but seem more impressive and scientific. Polygraphs, after all, are not administered by licensed medical professionals.

Nancy Kanwisher, a cognitive scientist at M.I.T., relies a great deal on MRI technology. In 1997, she identified an area near the bottom of the brain that is specifically involved in perceiving faces. She has become a pointed critic of the rush to commercialize brain imaging for lie detection, and believes that it's an exaggeration even to say that research into the subject is "preliminary." The tests that have been done, she argues, don't really look at lying. "Making a false response when instructed to do so is not a lie," she says. The ninety-per-cent "accuracy" ascribed to fMRI lie detection refers to a scenario so artificial that it is nearly meaningless. To know whether the technology works, she believes, "you'd have to test it on people whose guilt or innocence hasn't yet been determined, who believe the scan will reveal their guilt or innocence, and whose guilt or innocence can be established by other means afterward." In other words, you'd have to run a legal version of a clinical trial, using real suspects instead of volunteers.

Langleben believes that Kanwisher is too pessimistic. He suggested that researchers could recruit people who had been convicted of a crime in the past and get them to lie retrospectively about it. Or maybe test subjects could steal a "bagel or something" from a convenience store (the researchers could work out an agreement with the store in advance) and then lie about it. But even these studies don't approximate the real-world scenarios Kanwisher is talking about.

She points out that the various brain regions that appear to be significantly active during lying are "famous for being activated in a wide range of different conditions – for almost any cognitive task that is more difficult than an easier task." She therefore believes that fMRI lie detection would be vulnerable to countermeasures – performing arithmetic in your head, reciting poetry – that involve concerted cognitive effort. Moreover, the regions that allegedly make up the brain's "lying module" aren't that small. Even Laken admitted as much. As he put it, "Saying 'You have activation in the anterior cingulate' is like saying 'You have activation in Massachusetts.' "

Kanwisher's complaint suggests that fMRI technology, when used cavalierly, harks back to two pseudosciences of the eighteenth and nineteenth

centuries: physiognomy and phrenology. Physiognomy held that a person's character was manifest in his facial features; phrenology held that truth lay in the bumps on one's skull. In 1807, Hegel published a critique of physiognomy and phrenology in "The Phenomenology of Spirit." In that work, as the philosopher Alasdair MacIntyre writes, Hegel observes that "the rules that we use in everyday life in interpreting facial expression are highly fallible." (A friend who frowns throughout your piano recital might explain that he was actually fuming over an argument with his wife.) Much of what Hegel had to say about physiognomy applies to modern attempts at mind reading. Hegel quotes the scientist Georg Christoph Lichtenberg, who, in characterizing physiognomy, remarked, "If anyone said, 'You act, certainly, like an honest man, but I can see from your face you are forcing yourself to do so, and are a rogue at heart,' without a doubt every brave fellow to the end of time when accosted in that fashion will retort with a box on the ear." This response is correct, Hegel argues, because it "refutes the fundamental assumption of such a 'science' of conjecture – that the reality of a man is his face, etc. The true being of man is, on the contrary, his act; individuality is real in the deed." In a similar vein, one might question the core presumption of fMRI – that the reality of man is his brain.

Elizabeth Phelps, a prominent cognitive neuroscientist at N.Y.U., who studies emotion and the brain, questions another basic assumption behind all lie-detection schemes – that telling a falsehood creates conflict within the liar. With the polygraph, the assumption is that the conflict is emotional: the liar feels guilty or anxious, and these feelings produce a measurable physiological response. With brain imaging, the assumption is that the conflict is cognitive: the liar has to work a little harder to make up a story, or even to stop himself from telling the truth. Neither is necessarily right. "Sociopaths don't feel the same conflict when they lie," Phelps says. "The regions of the brain that might be involved if you have to inhibit a response may not be the same when you're a sociopath, or autistic, or maybe just strange. Whether it's an emotional or a cognitive conflict you're supposed to be exhibiting, there's no reason to assume that your response wouldn't vary depending on what your personal tendencies are – on who you are."

When I talked to Huizenga, the No Lie C.E.O., a few months after I had

met him in California, he was unperturbed about the skepticism that he was encountering from psychologists. "In science, when you go out a little further than other people, it can be hard," he said. "The top people understand, but the middle layer don't know what you're talking about."

Huizenga told me that he was trying to get fMRI evidence admitted into a California court for a capital case that he was working on. (He would not go into the case's details.) Given courts' skepticism toward the polygraph, Huizenga's success is far from certain. Then again we are in a technology-besotted age that rivals the twenties, when Marston popularized lie detection. And we live in a time when there is an understandable hunger for effective ways to expose evildoers, and when concerns about privacy have been nudged aside by our desire for security and certainty. "Brain scans indicate": what a powerful phrase. One can easily imagine judges being impressed by these pixellated images, which appear so often in scientific journals and in the newspaper. Indeed, if fMRI lie detection is successfully marketed as a service that lawyers steer their clients to, then a refusal even to take such a test could one day be cause for suspicion.

Steven Hyman, the Harvard psychiatrist, is surprised that companies like No Lie have eluded government oversight. "Think of a medical test," he said. "Before it would be approved for wide use, it would have to be shown to have acceptable accuracy among the populations in whom it would be deployed. The published data on the use of fMRI for lie detection uses highly artificial tests, which are not even convincing models of lying, in very structured laboratory settings. There are no convincing data that they could be used accurately to screen a person in the real world." But, in the end, that might not matter. "Pseudo-colored pictures of a person's brain lighting up are undoubtedly more persuasive than a pattern of squiggles produced by a polygraph," he said. "That could be a big problem if the goal is to get to the truth."

Laken, meanwhile, thinks that people who find themselves in a jam, and who are desperate to exonerate themselves, simply have to educate themselves as consumers. "People have said that fMRI tests are unethical and immoral," he said. "And the question is, Why is it unethical and immoral if somebody wants to spend their money on a test, as long as they understand what it is

they're getting into? We've never said the test was perfect. We've never said we can guarantee that this is admissible in court and that's it – you're scot-free." Later that day, I looked again at the Cephos Web site. It contained a bolder proclamation. "The objective measure of truth and deception that Cephos offers," it said, "will help protect the innocent and convict the guilty."

### 0.0.7 Better Decisions Through Science, John A. Swets, Robyn M. Dawes, and John Monahan (*Scientific American*, October 2000)

John A. Swets, Robyn M. Dawes, and John Monahan (*Scientific American*, October 2000).

A physician stares at a breast x-ray, agonizing over whether an ambiguous spot is a tumor. A parole board weighs the release of a potentially violent criminal. A technician at an airport worries over a set of ultrasound readings: do they suggest a deadly crack in an airplane's wing?

All these people are grappling with diagnostic decisions. In spite of incomplete or ambiguous evidence, they must determine whether or not a certain condition exists (or will occur). Such problems abound in health care, public safety, business, environment, justice, education, manufacturing, information processing, the military and government. And the stakes can be high. In many cases, a wrong verdict means that people will die.

Perhaps surprisingly, the diagnostic decision-making process turns out to be essentially the same across fields. Hence, methods that improve the process in one industry can usually serve in others. At least two such methods are already available. Sadly, though, they remain unknown or unused in many realms. One increases accuracy, enhancing the odds that any given decision will be the correct one. The other improves the "utility" of a decision-making approach, ensuring that the number of true cases found does not come at the cost of an unreasonable number of false positive diagnoses ("false alarms"). These methods are statistical, but math phobics have nothing to fear; the basic logic is easy to grasp.

No one is saying that diagnosticians must always be slaves to mathematical formulas. In certain arenas (such as clinical medicine and weather forecasting), objective tools may function best as "second opinions" that inform a reviewer's decisions but do not have the final word. In other fields, however, statistical analyses have frequently been found to be more accurate than subjective judgments, even those made by highly experienced professionals.

We focus in this article on diagnoses that hinge on a choice between just two alternatives – yes or no (Is a tumor present? Is an airplane wing de-

fective?). Certainly the world is full of problems involving a wider range of options, but serious yes/no decisions are prevalent.

TOOLS OF THE TRADE

If diagnostic tests always produced straightforward answers, no one would need statistical decision-making tools. In reality, though, the raw results of diagnostic tests usually have to be interpreted. In a simple example, the fluid pressure in the eye is measured to detect whether a person has glaucoma, which robs vision by damaging the optic nerve and other parts of the eye. A very low score clearly means the eye is healthy, and a high score signifies glaucoma. But scores in between are ambiguous, unable to indicate which patients have the condition and which do not.

Statistics can clear some of that fog. For argument's sake, assume that pressure is the only diagnostic measure available for glaucoma. Assume, too, that pressures below 10 on the standard measuring scale always signify good health, pressures over 40 always signify disease, and readings between 10 and 40 can occur in affected as well as healthy eyes.

To cope with this ambiguity, analysts would first identify a large population of individuals whose scores on the pressure test were known. Then they would determine which people went on to have vision problems characteristic of glaucoma within a set period and which did not. And they would calculate the odds that people having each possible score will have glaucoma. Finally, guided by those probabilities (and by other considerations we will discuss), they would set a rational cut point, or diagnostic threshold: scores at or above that level would yield a positive diagnosis ("the patient has glaucoma"); scores below would yield a negative diagnosis ("the patient does not have glaucoma").

Of course, single diagnostic tests may not be as informative as a combination. To enhance the accuracy of a diagnosis, analysts can combine data from many tests that each provide unique information, giving greater weight to measurements that are most predictive of the condition under study. The mathematical algorithms that specify the best tests to include in a diagnostic workup and that calculate the likelihood, based on the combined results, that a condition is present are known as statistical prediction rules (SPRs).

Totally objective data, such as pressure readings, are not the only features

that can be incorporated to enhance the accuracy of statistical prediction rules; subjective impressions can be quantified and included as well. They can be objectified, for instance, by making an explicit list of perceptual criteria (such as the size and irregularity of a possibly malignant mole) that can be rated according to a scale, perhaps from one to five.

If more than one statistical prediction rule is available, decision makers have to determine which ones are most accurate. This challenge, too, can be met objectively. The overall accuracy of prediction rules can be evaluated by reviewing what are called ROC (receiver operating characteristic) curves. Such curves were first applied to assess how well radar equipment in World War II distinguished random interference (noise) from signals truly indicative of enemy planes.

Programs that generate ROC curves consider what will happen if a particular raw score on a diagnostic test (or set of tests) is selected as the diagnostic threshold for a yes/no decision. What percent of individuals who truly have the condition in question will correctly be deemed to have it (true positive decisions, or "hits")? And what percent of individuals free of the condition will mistakenly be deemed to have it (false positive decisions, or false alarms)?

Then, for each threshold, the programs plot the percentage of true positives against the percentage of false positives. The result is a bowed curve, rising from the lower left corner, where both percentages are zero, to the upper right corner, where both are 100. The more sharply the curve bends, the greater the accuracy of the rule, because the number of hits relative to the number of false alarms is higher.

Obviously, true positives and false positives are not the only outcomes possible. A yes/no diagnosis based on any particular threshold will also generate true negatives (individuals are correctly deemed to be free of the condition being evaluated) and false negatives, or "misses" (individuals are incorrectly deemed to be free of the condition). But these results are the exact complements of the others and thus can be ignored when constructing ROC curves. A true positive rate of 80 percent, for instance, automatically means that the miss rate is 20 percent.

Given that few diagnostic methods are perfect at sorting individuals who have a condition from individuals who do not, institutions have to decide how

important it is to find all or most true positives – because more true positives come at the cost of more false alarms. That is, they need to set a threshold that makes good sense for their particular situation.

Returning to our glaucoma example, clinicians who looked only at pressure could find virtually every case of glaucoma if they chose a very "lenient" diagnostic cutoff – say, a score of 10. After all, the test sample revealed that virtually everyone with glaucoma has a score above that level. Yet that cutoff would result in many healthy people being told they were ill; those people would then be subjected unnecessarily to both worry and treatment. To minimize such errors, clinicians could instead set a rather strict diagnostic threshold – an eye pressure of 35, perhaps; very few healthy people in the sample had pressures that high. But this strict criterion would miss more than half of all affected individuals, denying them treatment. In setting a threshold, decision makers weigh such issues as the consequences of misses and false alarms and the prevalence of the problem under consideration in the population being tested. Fortunately, some rules of thumb and mathematical aids for finding the optimal cutoff point have been developed. For instance, a high prevalence of a problem in a population or a large benefit associated with finding true cases generally argues for a lenient threshold; conversely, a low prevalence or a high cost for false alarms generally calls for a strict threshold.

RULES COME TO LIFE

Although statistical prediction rules and ROC curves are often sorely underused by diagnosticians, real-life examples of their value abound. One of the most dramatic illustrations comes from psychiatry.

Increasingly, psychiatrists and clinical psychologists are asked to determine whether incarcerated or disturbed individuals are likely to become violent. People who seem most likely to endanger others need to be identified and treated for their own good and for others' safety. At the same time, interfering in the lives of people who do not need care is unacceptable.

Disconcertingly, in 1993 the most sophisticated study of clinicians' unaided assessments uncovered a startling lack of accuracy. Clinicians who diagnosed consecutive patients coming to the emergency department of a metropolitan psychiatric hospital proved no more accurate than chance at predicting which

female patients would commit violence in the community within the next six months. Their success rate with male patients was only modestly better.

In response to such findings, a number of statistical prediction rules were developed for assessing the probability of violence. One of the most studied is the Violence Risk Appraisal Guide (VRAG), which measures 12 variables, among them scores on a checklist of features indicative of psychopathy and assessments of maladjustment in elementary school.

In a test of the rule's ability to predict whether criminals being discharged from a maximum-security hospital would commit violent acts over the next several years, the VRAG divided the subjects into two categories of risk: "high" and "low." Fifty-five percent of the high-risk group but only 19 percent of the low committed a new violent offense – an accuracy level well above that of chance. And a newer statistical prediction rule proved to be even better at forecasting violence in noncriminals about to be discharged from psychiatric facilities. Nevertheless, interested parties continue to disagree over whether clinicians should treat such rules as advisory or make decisions based solely on the statistics.

BETTER CANCER DIAGNOSES

Statistical prediction rules have also had impressive success in studies aimed at helping radiologists diagnose breast cancer. In one such investigation, radiologists in community hospitals evaluated mammograms in their usual, subjective way. Months later they examined the same mammograms according to a checklist of perceptual features (such as how fuzzy the borders of a mass seem to be) developed by radiologists who specialize in reviewing mammograms. Then a statistical prediction rule converted the ratings into probability assessments indicating the likelihood for each patient that breast cancer was present. The radiologists reviewed these probabilities but ultimately made their own judgments. The extra data helped considerably. General radiologists who took the statistical data into account became more accurate, reaching the precision of specialists who had used the checklist.

Physicians who treat prostate cancer are already making extensive use of statistical prediction rules. One rule in particular is getting a serious workout. Once a man is "clinically" deemed to have cancer of the prostate gland (on the basis of a checkup, a simple needle biopsy and noninvasive tests), the

question of the best treatment arises [see "Combating Prostate Cancer," by Marc B. Garnick and William R. Fair; *Scientific American*, December 1998]. Neither surgery to remove the affected gland nor radiation focused tightly on it (to limit side effects) will eliminate the tumor if it has grown beyond the gland or has spread to other parts of the body. Hence, physicians strive to determine the status of the tumor before any treatment is attempted. Unfortunately, a great many tumors that initially seem to be confined to the prostate later turn out to have been more advanced.

For years, doctors had few good ways of predicting which patients truly had confined disease and which did not. More recently, however, doctors and patients have been able to gain a clearer picture by consulting probability tables published in the May 14, 1997, issue of the *Journal of the American Medical Association.* The researchers who created the tables knew that three assessments each had independent predictive value: the tumor's "clinical stage" (a determination, based on noninvasive tests, of tumor size and spread), the level in the blood of a specific protein (PSA, or prostate-specific antigen) and the Gleason score (an indicator of tumor aggressiveness, based on microscopic analyses of a biopsy sample). The investigators therefore developed a statistical prediction rule that looked at virtually every combination of results for these three variables and calculated the odds that the initial diagnosis of "no spread" would be correct. Then they listed the probabilities in a user-friendly, tabular form.

GOOD CHANCE OF RAIN

It would be a mistake to think that only medical practitioners use statistical prediction rules. In fact, meteorologists adopted the tools for weather forecasting more then 25 years ago.

The National Weather Service routinely feeds weather-related data into statistical programs designed to estimate the likelihood that tornadoes, hurricanes, heavy rains and other hazards will arise in different parts of the nation. The weather service then conveys these objective predictions to meteorologists in local areas, who modify the predictions in light of new information or of factors they think the computer programs did not address adequately.

Other groups have embraced the techniques as well – among them, graduate admissions committees at universities. In a typical example, a committee

will project first-year grades from two variables – undergraduate grades and graduate school aptitude exams, on the assumption that students scoring above some preselected high level should generally be admitted and those scoring below a specified lower level should generally be rejected. Then the committee will more subjectively evaluate the credentials of applicants who have not been admitted or rejected by the school's statistical prediction rule.

One law school objectively rates two variables that were formerly assessed subjectively: the quality of the student's undergraduate institution and the extent of grade inflation at that institution. Along with the student's grade point average and scores on the aptitude exam required for law school, it considers the mean exam score of all students from the applicant's college who took the test and the mean grade point average of students from that college who applied to law school. The revised formula predicts first-year law-school grades significantly better than the two-variable scheme.

THORNY THRESHOLDS

So far we have highlighted success stories. But the merit of statistical analyses may be best illustrated by examples of failure to apply them for setting rational diagnostic thresholds – such as for tests that detect the human immunodeficiency virus (HIV), the cause of AIDS.

HIV screening relies initially on a relatively simple test that detects the presence of anti-HIV antibodies, molecules produced when the immune system begins to react against HIV. Sometimes these antibodies arise for reasons other than the presence of HIV, however. Hence, if the outcome (based on some antibody threshold) is positive, laboratories will run a different, more sophisticated test. This two-test requirement is meant to help limit false positives. The antibody tests are particularly problematic in that, illogically, the several approved tests differ in their accuracies and thresholds. Varied thresholds would make sense if each test were aimed at a distinct population, but that is not the case.

The thresholds are disturbing in another way as well. They were originally set to distinguish clean from tainted donated blood; then they were left unchanged when they were enlisted to identify people infected with the virus. Throwing out a pint of uncontaminated blood because of a false positive is a cheap mistake; sending an alarmed, uninfected person for further HIV test-

ing is not. Worse still, the original thresholds have been applied mindlessly to low-risk blood donors, high-risk donors, military recruits and methadone-clinic visitors – groups whose infection rates vary over an enormous range. For the high-risk groups, the threshold should be set more leniently than for the low-risk populations (to maximize discovery), even if the price is a higher rate of false positives.

Recent years have seen the introduction of confirmatory tests that are more accurate and of HIV therapies that prolong life and health. Consequently, false positive diagnoses are rare these days, and people who are infected with HIV benefit much more from being diagnosed than was true in the past. These advances mean that the diagnostic problem has shifted from whom to call positive to whom to test. The time has come for doctors to lower their thresholds for deciding when to test; they should not be waiting until patients show up with obvious symptoms of infection. We would even argue that almost every adult should be screened and that federal agencies should take the lead in encouraging such testing.

Objective methods for establishing thresholds are also being dangerously underused in parts of the aerospace industry. This industry must constantly diagnose conditions that are serious but arise relatively infrequently, among them cracked wings and life-threatening hazards during flights. The costs of missing a cracked wing are large and obvious: many passengers may die if the plane crashes. On the other hand, a false-positive decision takes a plane out of service unnecessarily, potentially causing inconvenience and lost income. At first blush, the benefits and costs point toward a lenient threshold, favoring lives over dollars. Yet such cracks occur rarely; therefore a lenient threshold yields an unworkable number of false positives. Unfortunately, no one has yet tackled this issue with the available statistical techniques.

Purchasers of cockpit alarms (such as airlines and the military) have similarly failed to come to grips with how best to set decision thresholds. Alarms go off in flight under many circumstances – when sensing devices determine that another plane is too close, that the plane is getting too near to the ground, that an engine is dying or that wind shear is threatening the landing area. But they cry wolf too often, largely because the sensors are only moderately accurate and because the thresholds set for them are rather lenient.

Pilots are reluctant to act on the warnings unnecessarily, because doing so can be quite disruptive. This situation has raised fears that the high number of false alarms will cause pilots to ignore or respond slowly to a real emergency. To date, though, no one has forced manufacturers to consider the false positive rate when they establish alarm thresholds.

A PLEA

Clearly, statistical prediction rules can often raise the accuracy of repetitive diagnostic decisions, and formulas for setting decision thresholds can improve the utility of those decisions. But these tools provide other advantages as well. By standardizing the features that are assessed to make a diagnosis, the prediction rules can hasten the speed with which professionals recognize key diagnostic features. They also give decision makers a way to communicate more easily and precisely about impressionistic features. And they can help teach newcomers to a field.

Yet they are often met with resistance, especially if they are seen as replacing or degrading clinicians. Further, diagnosticians want to feel that they understand their own diagnoses and recommendations and that they can give a narrative of their thought processes. The results of a statistical prediction rule may be hard to include in such an account, particularly if the logic behind the analysis is not self-evident. We understand all these concerns. Nevertheless, the benefits that statistical tools provide surely justify consideration by decision makers who hold others' lives and futures in their hands.

### 0.0.8 Do Fingerprints Lie?, Michael Specter (*New Yorker*), May 27, 2002

May 27, 2002

Michael Specter (*New Yorker*)

The gold standard of forensic evidence is now being challenged.

Late one afternoon in the spring of 1998, a police detective named Shirley McKie stood by the sea on the southern coast of Scotland and thought about ending her life. A promising young officer, the thirty-five-year-old McKie had become an outcast among her colleagues in the tiny hamlet of Strathclyde. A year earlier, she had been assigned to a murder case in which an old woman was stabbed through the right eye with a pair of sewing scissors. Within hours of the killing, a team of forensic specialists had begun working their way through the victim's house. Along with blood, hair, and fibres, the detectives found some unexpected evidence: one of the prints lifted from the room where the murder took place apparently matched the left thumb of Detective McKie.

Crime scenes are often contaminated by fingerprints belonging to police officers, and investigators quickly learn to eliminate them from the pool of suspects. But McKie said that she had never entered the house. Four experts from the Scottish Criminal Record Office – the agency that stores and identifies fingerprints for Scotland's police – insisted, however, that the print was hers. Though McKie held to her story, even her father doubted her. "I love my daughter very much," Iain McKie, who served as a police officer in Scotland for more than thirty years, told me earlier this year. "But when they said the print was Shirley's I have to admit I assumed the worst. My entire career I had heard that fingerprints never lie."

Nobody actually suspected McKie of murder, and in fact the victim's handyman, David Asbury, was charged with the crime. The sole physical evidence against him consisted of two fingerprints – one of his, lifted from an unopened Christmas gift inside the house, and one of the victim's, found on a biscuit tin in Asbury's home. The last thing prosecutors needed was for their own witness to raise questions in court about the quality of the

evidence. Yet McKie did just that – repeating under oath that she had never entered the house. Asbury was convicted anyway, but Scottish prosecutors were enraged by McKie's testimony. As far as they were concerned, McKie had not only lied; she had challenged one of the evidentiary pillars of the entire legal system. Despite their victory in the murder trial, they charged McKie with perjury.

Desperate, she went to the public library and searched the Internet for somebody who might help her. Among the names she came upon was that of Allan Bayle, a senior forensic official at New Scotland Yard and perhaps the United Kingdom's foremost fingerprint expert. (It was Bayle's expertise and supporting evidence that helped convict one of the principal Libyan suspects in the 1988 bombing of Pan Am Flight 103, over Lockerbie, Scotland.) He agreed to review the prints, and what he saw astonished him. "It was obvious the fingerprint was not Shirley's," Bayle told me recently. "It wasn't even a close call. She was identified on the left thumb, but that's not the hand the print was from. It's the right forefinger. But how can you admit you are wrong about Shirley's print without opening yourself to doubt about the murder suspect, too?" Bayle posted a comment on Onin.com, a Web site trafficked regularly by the world's fingerprint community. "I have looked at the McKie case," he wrote. "The mark is not identical. I have shown this mark to many experts in the UK and they have come to the same conclusions."

Bayle's assertion caused a furor. He was threatened with disciplinary action, shunned by his colleagues, and, after a quarter century with the Metropolitan Police, driven from his job. But in the end McKie was acquitted, and Bayle's statement helped challenge a system that had, until then, simply been taken for granted.

For more than a century, the fingerprint has been regarded as an unassailable symbol of truth, particularly in the courtroom. When a trained expert tells a judge and jury that prints found at a crime scene match those of the accused, his testimony often decides the case. The Federal Bureau of Investigation's basic text on the subject is entitled "The Science of Fingerprints," and a science is what F.B.I. officials believe fingerprinting to be; their Web site states that "fingerprints offer an infallible means of personal identification." The Bureau maintains a database that includes the fingerprints of more than

forty-three million Americans; it can be searched from precinct houses and properly equipped police cruisers across the country. Fingerprints are regularly used to resolve disputes, prevent forgery, and certify the remains of the dead; they have helped send countless people to prison. Until this year, fingerprint evidence had never successfully been challenged in any American courtroom.

Then, on January 7th, U.S. District Court Judge Louis H. Pollak – a former dean of the law schools at Yale and at the University of Pennsylvania – issued a ruling that limited the use of fingerprint evidence in a drug-related murder case now under way in Philadelphia. He decided that there were not enough data showing that methods used by fingerprint analysts would pass the tests of scientific rigor required by the Supreme Court, and noted the "alarmingly high" error rates on periodic proficiency exams. Although Judge Pollak later decided to permit F.B.I. fingerprint experts to testify in this particular case, students of forensic science felt his skepticism was justified. "We have seen forensic disciplines which focus on bite marks, hair analysis, and handwriting increasingly questioned in the courts," Robert Epstein, who had argued for the exclusion of fingerprint testimony in the case, told me. "But we have accepted fingerprinting uncritically for a hundred years."

Epstein, an assistant federal public defender in Philadelphia, was responsible for the first major court challenge to the discipline, in 1999, in U.S. v. Byron Mitchell. In that case, Epstein showed that standards for examiners vary widely, and that errors on proficiency tests – which are given irregularly and in a variety of forms – are far from rare. The critical evidence consisted of two fingerprint marks lifted from a car used in a robbery. To prepare for the trial, F.B.I. officials had sent the prints to agencies in all fifty states; roughly twenty per cent failed to identify them correctly. "After all this time, we still have no idea how well fingerprinting really works," Epstein said. "The F.B.I. calls it a science. By what definition is it a science? Where are the data? Where are the studies? We know that fingerprint examiners are not always right. But are they usually right or are they sometimes right? That, I am afraid, we don't know. Are there a few people in prison who shouldn't be? Are there many? Nobody has ever bothered to try and find out. Look closely at the great discipline of fingerprinting. It's not only not a science –

it should not even be admitted as evidence in an American court of law."

Fingerprints have been a source of fascination for thousands of years. They were used as seals on legal contracts in ancient Babylonia, and have been found embossed on six-thousand-year-old Chinese earthenware and pressed onto walls in the tomb of Tutankhamun. Hundreds of years ago, the outline of a hand with etchings representing the ridge patterns on fingertips was scratched into slate rock beside Kejimkujik Lake, in Nova Scotia.

For most of human history, using fingerprints to establish a person's identity was unnecessary. Until the nineteenth century, people rarely left the villages in which they were born, and it was possible to live for years without setting eyes on a stranger. With the rise of the Industrial Revolution, cities throughout Europe and America filled with migrants whose names and backgrounds could not be easily verified by employers or landlords. As the sociologist Simon Cole made clear in "Suspect Identities," a recent history of fingerprinting, felons quickly learned to lie about their names, and the soaring rate of urban crime forced police to search for a more exacting way to determine and keep track of identities. The first such system was devised in 1883 by a Parisian police clerk named Alphonse Bertillon. His method, called anthropometry, relied on an elaborate set of anatomical measurements – such as head size, length of the left middle finger, face height – and features like scars and hair and eye color to distinguish one person from another. Anthropometry proved useful, but fingerprinting, which was then coming into use in Britain, held more promise. By the eighteen-sixties, Sir William J. Herschel, a British civil servant in India, had begun to keep records of fingerprints and use them to resolve common contract disputes and petty frauds.

Fingerprinting did not become indispensable, however, until 1869, when Britain stopped exiling criminals to Australia, and Parliament passed the Habitual Criminals Act. This law required judges to take past offenses into account when determining the severity of a sentence. But in order to include prior offenses in an evaluation one would need to know whether the convict had a previous record, and many criminals simply used a different alias each time they were arrested. The discovery that no two people had exactly the same pattern of ridge characteristics on their fingertips seemed to offer a solution. In 1880, Dr. Henry Faulds published the first comments, in

the scientific journal *Nature*, on the use of fingerprints to solve crimes. Soon afterward, Charles Darwin's misanthropic cousin, Sir Francis Galton, an anthropologist and the founder of eugenics, designed a system of numbering the ridges on the tips of fingers – now known as Galton points – which is still in use throughout the world. (Ultimately, though, he saw fingerprints as a way to classify people by race.)

Nobody is sure exactly how Mark Twain learned about fingerprints, but his novel "Pudd'nhead Wilson," published in 1894, planted them in the American imagination. The main character in the book, a lawyer, earned the nickname Pudd'nhead in part because he spent so much time collecting "finger-marks" – which was regarded as proof of his foolishness until he astounded his fellow-citizens by using the marks to solve a murder. If you were to walk into a courtroom today and listen to the testimony of a typical forensic expert, you might hear a recitation much like Pudd'nhead Wilson's:

Every human being carries with him from his cradle to his grave certain physical marks which do not change their character, and by which he can always be identified – and that without shade of doubt or question. These marks are his signature, his physiological autograph, so to speak, and this autograph cannot be counterfeited, nor can he disguise it or hide it away, nor can it become illegible by the wear and the mutations of time. This signature is each man's very own. There is no duplicate of it among the swarming populations of the globe!

Some things have changed since Pudd'nhead Wilson, of course. A few weeks ago, I visited the headquarters of the Integrated Automated Fingerprint Identification Systems, the F.B.I.'s billion-dollar data center, just outside Clarksburg, West Virginia – a citadel of the American forensic community. After driving past a series of shacks and double-wides and Bob Evans restaurants, you come upon a forest with a vast, futuristic complex looming above the trees. (I.A.F.I.S. moved from more crowded quarters in the Hoover Building in 1995, thanks to the influence of the state's senior senator, Robert C. Byrd.)

Clarksburg is home to the world's largest collection of fingerprints; on an average day, forty thousand are fed into the system. The I.A.F.I.S. computers, which can process three thousand searches a second, sort through the

database in a variety of ways. For example, they compare complete sets of fingerprints in the files with new arrivals – as when a suspect is held in custody and the police send his "ten-prints" to I.A.F.I.S. The computer hunts for shared characteristics, and then attempts to match the prints to a record on file. "We identify about eight thousand fugitives per month here," Billy P. Martin, the acting chief of the Identification and Investigative Services Section, told me. Martin said that eleven per cent of job applicants whose fingerprints are entered into the system – they could be day-care workers, casino staff, federal employees – turn out to have criminal records; as many as sixty per cent of the matches are repeat offenders.

The center looks like a NASA control room, with dozens of people monitoring the encrypted network of fingerprint machines sending in data from police stations throughout the country. The main computer floor is the size of two football fields and contains sixty-two purple-and-gray "jukeboxes," each filled with two hundred compact disks containing fingerprints. (There are three thousand sets on each CD.) When someone is arrested, his prints are initially searched against a state's computer files. If the search finds nothing, the information is forwarded to the federal database in Clarksburg. To make a match, the I.A.F.I.S. computer analyzes the many points on the ridges of every fingerprint it receives, starting with the thumb and working toward the pinkie; only when the data produce prints that match (or several prints that seem similar) is the original print forwarded to an analyst for comparison.

"We used to go to a file cabinet, pull out paper cards. If it was all loops – which is the most common type of print – you could spend an hour," Martin said. "Now a computer algorithm does it in seconds. The system searches the electronic image against the database and pulls up the image onto the screen. The accuracy rate on first run is 99.97 per cent." Still, this would mean that the I.A.F.I.S. computers make three hundred mistakes in every million searches. That is where trained examiners come in. The patterns on fingertips are more like topographical maps or handwriting than, say, bar codes. They can be so similar that even the most sophisticated computer program can't tell them apart; it takes a trained human eye to detect the subtle differences.

I sat with one of the examiners in a dim, nearly silent room lined with

what seemed like an endless series of cubicles. At each station, someone was staring at a monitor with two huge fingerprints on it. No two people – not even identical twins – have ever been shown to share fingerprints. The friction ridges that cover the skin on your hands and feet are formed by the seventeenth week in the womb; at birth they have become so deep that nothing can alter them, not even surgery. Look at your fingertips: the patterns resemble finely detailed maps of the bypasses and exit ramps on modern roads. Experts use the nomenclature of the highway to describe them: there are spurs, bifurcations, and crossovers. Some people have fingertips that are dominated by "loops," others by "tented arches" or small circles that examiners call "lakes," or smaller ones still, called "dots." Collectively, these details are referred to as minutiae – an average human fingerprint may contain as many as a hundred and fifty minutia points. To identify fingerprints, an expert must compare these points individually, until enough of them correspond that he or she feels confident of a match.

When fingerprints are properly recorded (inked, then rolled, finger by finger, onto a flat surface, or scanned into a machine that captures and stores each finger as a digital image), identification works almost flawlessly. The trouble is that investigators in the field rarely see the pristine prints that can be quickly analyzed by a computer; most of the prints introduced at criminal trials are fragments known as "latent prints." Crime scenes are messy, and the average fingerprint taken from them represents only a fraction of a full fingertip – about twenty per cent. They are frequently distorted and hard to read, having been lifted from a grainy table or a bloodstained floor. "It is one thing to say that fingerprints are unique and quite another to suggest that a partial latent print, often covered in blood or taken from an obscure surface, is unique, identical, or easy to identify," Barry Scheck told me. In the past decade, Scheck, who directs the Innocence Project, has used DNA evidence to exonerate more than a hundred prisoners, some of them on death row. "We have always been told that fingerprint evidence is the gold standard of forensic science. If you have a print, you have your man. But it is not an objective decision. It is inexact, partial, and open to all sorts of critics."

Police use several methods to discover latent fingerprints. First, they shine a flashlight or a laser along the clean, solid surfaces on which a print may

have been left by the perspiration and oil on a fingertip. When a print is discovered, detectives use a brush and powder to mark it, much as they did in the nineteenth century; the powder clings to the perspiration. (The method works best on smooth surfaces, like glass.) The print is then photographed and lifted with tape.

The technology for retrieving partial and obscure fingerprints keeps improving. On a recent episode of the television program "C.S.I.," you might have seen detectives using a technique called superglue fuming to reveal the outline of a face on a plastic bag – an unconventional use of a common practice. In order to find difficult prints on an irregular surface, such as the human body, crime-scene investigators blow fumes of superglue over it. As the fumes adhere to the surface, the ridges of any fingerprint left there turn white and come clearly into view. Another common method involves ninhydrin, which works like invisible ink: when you douse paper with it, the chemical brings out any sweat that may have been left by fingertips. Ninhydrin is particularly useful with old prints or those covered in blood.

F.B.I. fingerprint examiners have a variety of computer tools – a sort of specialized version of Photoshop – to help them compare rolled prints with those in their system. In front of me, an I.A.F.I.S. examiner stared at his computer screen as a training instructor, Charles W. Jones, Jr., explained the process. "He is looking for ridges that form dots," Jones said. "Bifurcations. Usually they look for six or seven of those." The examiners work around the clock, in three shifts, and are required to evaluate at least thirty prints an hour. They know nothing about the people attached to the fingers on their screens; the prints could be those of a rapist, a serial killer, Osama bin Laden, a woman applying for a job in the Secret Service, or a bus driver from Queens. ("Yesterday I did fifty-one for a couple hours in a row," an examiner told me proudly.)

At the bottom of the screen there are three buttons – "Ident," "Unable," and "Non-Ident" – and the examiner must click on one of them. If he identifies a finger, the print goes to a second analyst. If the two examiners independently reach the same conclusion, the fingerprint is considered to have been identified. If not, it gets forwarded to an analyst with more experience. "We have a pretty good fail-safe system," Jones said. "Computers help immensely.

But in the end they can't pull the trigger. That's our job."

Only a human being can make critical decisions about identity, and yet the talent, training, and experience of examiners vary widely. "The current identification system is only as genuine as the knowledge, experience, and ability of the specialist carrying out the comparison," David R. Ashbaugh, a staff sergeant with the Royal Canadian Mounted Police, writes, in "Quantitative-Qualitative Friction Ridge Analysis," which is considered the Bible of the field. And although fingerprint analysis has been in use for decades, there has never been any consensus about professional standards. How many distinct characteristics are necessary to prove that a latent fingerprint comes from a specific person? The answer is different in New York, California, and London. In certain states, and in many countries, fingerprint examiners must show that prints share a set number of Galton points before they can say they have made an identification. Australia and France require at least twelve matching Galton points; in Italy, the number is sixteen. In America, standards vary, even within a state. The F.B.I. doesn't require a minimum number of points; all such regulations were dropped fifty years ago, because, according to Stephen B. Meagher, the chief of the Bureau's latent-print unit, the F.B.I. believes that making an identification using Galton points alone can cause errors.

Meagher says that fingerprint analysis is an objective science; Robert Epstein, the Philadelphia attorney who has led the fight against presenting fingerprint evidence in court, says it is not a science at all. Neither is exactly right. Examining the many contours of a human finger is not as objective as measuring someone's temperature or weight, or developing a new vaccine. But it's not guesswork, either. It involves, inevitably, human judgment, and most people agree that when it is done well it is highly accurate. The difficulty is in determining whether it has been done well.

Scientific methodology is based on generating hypotheses and testing them to see if they make sense; in laboratories throughout the world, researchers spend at least as much time trying to disprove a theory as they do trying to prove it. Eventually, those ideas that don't prove false are accepted. But fingerprinting was developed by the police, not by scientists, and it has never been subjected to rigorous analysis – you cannot go to Harvard, Berkeley,

or Oxford and talk to the scholar working on fingerprint research. Yet by the early twentieth century fingerprinting had become so widely accepted in American courts that further research no longer seemed necessary, and none of any significance has been completed.

David L. Faigman, who teaches at the Hastings College of the Law and is an editor of the annually revised forensic text "Modern Scientific Evidence," has spent most of his career campaigning to increase the scientific literacy of judges and juries. Faigman likens the acceptance of fingerprint evidence to the way leeches were once assumed to be of great medical value. "Leeches were used for centuries," he told me. "It was especially common for the treatment of pneumonia and it was considered an effective therapy. It wasn't till late in the nineteenth century that they did the clinical tests to show that leeches did not help for pneumonia, and they may have actually hurt. Fingerprinting is like that in at least one crucial way: it is something we assume works but something we have never properly tested. Until we test our beliefs, we can't say for sure if we have leeches or we have aspirin" – an effective remedy that was used before it was understood. "One of the things that science teaches us is that you can't know the answers until you ask the questions."

The discussion of fingerprinting is only the most visible element in a much larger debate about how forensic science fits into the legal system. For years, any sophisticated attorney was certain to call upon expert witnesses – doctors, psychiatrists, Bruno Magli shoe salesmen – to assert whatever might help his case. And studies have shown that juries are in fact susceptible to the influence of such experts. Until recently, though, there were no guidelines for qualification; nearly anybody could be called an expert, which meant that, unlike other witnesses, the expert could present his "opinion" almost as if it were fact. Experts have been asked to testify about the rate at which a tire would skid, and the distance blood would splatter when a certain calibre bullet smashed into a skull. They have lectured scores of juries on the likelihood that a medicine could cause a particular side effect; they have interpreted polygraphs and handwriting, and have pronounced on whether a bite mark was made by one set of teeth to the exclusion of all others.

Although forensic evidence has proved particularly powerful with juries,

it is particularly weak as a science. By the nineteen-eighties, the kind of evidence that was routinely admitted into court without any statistical grounding or rationale had earned a name: "junk science." And junk science had become ubiquitous. With the problem growing out of control, in 1993 the Supreme Court took up a lawsuit called Daubert v. Merrell Dow Pharmaceuticals. The case involved a child who suffered from serious birth defects. His lawyers claimed that the defects were caused by Bendectin, a drug that was for many years routinely prescribed for morning sickness, which his mother took while she was pregnant. The company argued that no valid evidence existed to support the claim. The Court's decision set a new standard for scientific evidence in America: for the first time, it held that it was not permissible for expert witnesses to testify to what was "generally accepted" to be true in their field. Judges had to act as "gatekeepers," the Court said; if an expert lacked reliability he was no longer allowed in the courtroom. The ruling, and others that expanded upon it, laid down clear guidelines for the federal bench, requiring judges to consider a series of questions: Could a technique be tested or proved false? Was there a known or potential error rate? (DNA identification has provided the model, because experts have gathered enough statistical evidence to estimate the odds – which are astronomical – that one person's DNA could be traced to another.) The Court also instructed judges to consider whether a particular theory had ever been subjected to the academic rigor of peer review or publication.

The Daubert ruling forced federal judges to become more sophisticated about science, which has not been easy for them. "Daubert changed everything," Michael J. Saks, a law professor at Arizona State University, who has written widely on the subject, told me. "And it is pretty clear when you look at those criteria that fingerprinting simply doesn't satisfy any of them." Since the Daubert ruling, federal courts have judged handwriting evidence and hair identification to be unscientific. The use of polygraph data has also been curtailed. Questions have been raised about ballistics – say, whether a bullet can be traced back to a particular gun. Somehow, though, until Judge Pollak came along, challenges to fingerprinting continued to be regarded as heresy.

Relying largely on testimony presented by Robert Epstein in U.S. v. Byron

Mitchell, the first post-Daubert case involving fingerprint testimony, Judge Pollak ruled in January that an expert could say whether he thought fingerprints belonged to the people accused of the crime, but he could not say that the fingerprints he had examined were, beyond doubt, those of the defendant.

Pollak is one of the federal judiciary's most respected judges. Federal prosecutors were so concerned that any ruling he issued would carry a significance even greater than its legal weight that they asked the Judge to reconsider his precedent-shattering decision. Pollak agreed.

Late in February, Pollak held a hearing on the reliability of fingerprint evidence. For three days, several of the world's most prominent experts discussed their field in his courtroom. The F.B.I.'s Stephen B. Meagher testified that no Bureau analyst had ever misidentified a person in court, and that the Bureau's annual proficiency test was among the reasons that the Judge should be confident about admitting expert testimony. Allan Bayle, the British forensic specialist, flew in from London at the request of the defense. He had a different view. He told Pollak that the F.B.I.'s proficiency test was so easy it could be passed with no more than six weeks of training. "If I gave my experts [at Scotland Yard] these tests, they would fall about laughing," he told Pollak in court. Later, in conversation with me, he expanded on those comments. "The F.B.I. are conning themselves and they are conning everybody else," he said. "They don't even use real scene-of-crime marks for the fingerprint tests." He pointed out that the fingerprints used in the exams were so different from each other that almost anybody could tell them apart. "Let's say I asked you to look at a zebra, a giraffe, an elephant, and a lion. Then I asked you to find the zebra. How hard would that be? What the Bureau should be doing is comparing five zebras and selecting among them." Bayle and other critics stopped short of calling fingerprint evidence junk science, but they noted that there are few data showing how often latent prints are properly identified.

By February 27th, the final day of the hearing, the fissures in an old and accepted discipline had become visible, and Judge Pollak promised to issue a final ruling within a couple of weeks.

A few days after Pollak's hearing ended, I flew to Cardiff to attend the annual meeting of the Fingerprint Society. It was raining in Wales, and

the members of the society were deeply unsettled because their profession was under assault. Each year, the society gathers for a few days to listen to lectures and to talk about developments in the field. The society has always been a club – the type where you might expect to stumble upon Sherlock Holmes or G. K. Chesterton. The bar at the Thistle Hotel, where the conference was held, was filled with police officers from Sussex, Aberdeen, and most places in between. The conference was well attended by representatives of the United States Secret Service and the F.B.I. There were also a few stray academics interested in the latest obscure technology, such as magnetic nanoflake powders, which are able to capture fingerprints without disturbing whatever traces of DNA may be present. (With conventional methods, an investigator has to choose: either swab a mark to harvest the DNA or lift it to find the print.)

By the time I arrived, the society was preoccupied by two issues: the Pollak hearings and the lingering ill will from the McKie case, in Scotland. One of those in attendance was Meagher, the lead F.B.I. witness in Judge Pollak's courtroom. I introduced myself, and told him that I understood he couldn't discuss the Philadelphia case while it was under review, but asked if we could talk about the field in general. "No," he said, without a moment's hesitation. Iain McKie had also come to Cardiff that weekend, as had Allan Bayle. McKie, a tall, reedy man with a great nimbus of curly white hair, presented a lecture on the ethics of fingerprinting. He remained livid about the fact that a fingerprint had destroyed his daughter's career; although she had been acquitted of perjury, she felt unwelcome on the police force after having been strip-searched and jailed by her colleagues, and had resigned soon after her trial. She never returned to work. Today, she spends much of her time trying to force Scottish authorities to admit that what they did to her was wrong. "I believe a person made a mistake, and instead of admitting it they were prepared to send me to jail," Shirley McKie said after she was acquitted of perjury. "It ruined my life, and now I am trying to pick up the pieces."

The Scottish Criminal Record Office has never acknowledged the error, nor has the Fingerprint Society issued any statement about the incident. (David Asbury, the man convicted of the murder, was released in August of

2000, pending an appeal. As expected, the judge in the case questioned the validity of the fingerprint evidence that had led to his conviction.) In Cardiff, McKie told the Fingerprint Society that the system they represented was "incestuous, secretive, and arrogant. It has been opened to unprecedented analysis and it's sadly lacking. It pains me to say that, because I was a police officer for thirty years. You are indicted on the basis of a fingerprint. You are not innocent till proven guilty; if the police have a print, you are assumed to be guilty. We need to start a new culture. The view that the police and fingerprint evidence are always right, the rest of the world be damned, has to end."

Afterward, the corridors and conference rooms were buzzing; it was as if somebody had challenged the fundamentals of grammar at the annual meeting of the Modern Language Association. But McKie was far from the only speaker at the conference to raise questions about the field. Christophe Champod, who works for a British organization called the Forensic Science Service, has long attempted to apply rigorous statistical methods to fingerprinting. Champod spoke in an understated and academic manner, but what he had to say was even more forceful than McKie's presentation. He told the audience that they had only themselves to blame for the state of the field, that for years they had resisted any attempts to carry out large trials, which would then permit examiners to provide some guidance to juries about the value of their analysis, as is the case with DNA. "What we are trying to do in this field is reduce, reduce, reduce the population so that there is only a single individual that can possess a set of fingerprints. But we can never examine the fingerprints of the entire universe. So, based on your experience, you make an inference: the probability that there is another person in the universe that could have a good match for the mark is very small. In the end, it's like a leap of faith. It's a very small leap, but it is a leap nonetheless."

Half an hour had been allotted for questions, but there was only silence. Afterward, one of the organizers explained it to me: "He was using the terms of religion to describe our science. That's just not fair."

Allan Bayle invited me to visit him in London after the meeting. Bayle is six feet five with sandy hair and flecks of gray in his blue eyes. He had recently married and he lives with his wife, child, and mother-in-law just

steps from the M1 motorway entrance in Hendon, on the northern edge of the city. We sat in his conservatory on a cloudy day while his five-month-old boy slept in a stroller beside us.

Bayle was frustrated. For the past five years, he had worked mostly as a lecturer on fingerprints for the Metropolitan Police. "I taught advanced forensic scene examination, and I loved it. Once I said I would give evidence in the McKie case, though, I was no longer allowed to go to meetings. But that is not why I left. They did nothing about this mistake in identity. When you know something is wrong, how can you stay silent?" He told me he was particularly upset that Shirley McKie's career as a police officer had ended for no reason. Bayle's life, too, has changed. He now works as an independent consultant. Although he has been portrayed as a critic of fingerprint analysis, he is critical only of the notion that it should never be questioned. "It's a valuable craft," he said. "But is it a science like physics or biology? Well, of course not. All I have been saying is, let's admit we make errors and do what we can to limit them. It is such a subjective job. The F.B.I. want to say they are not subjective. Well, look at what David Ashbaugh – certainly among the most noted of all fingerprint analysts – said when he testified in the Mitchell case." Ashbaugh had clearly stated that fingerprint identification was "subjective," adding that the examiner's talents are his "personal knowledge, ability, and experience."

Bayle took out a large portfolio containing dozens of fingerprints, as well as gruesome pictures of crime scenes. "Look at the mess," he said. He showed me a series of photographs: jagged fingerprints – black smudges, really – recovered from the scenes of several murders he had investigated. "With all that information, you then come to your conclusions. You have to somehow match that to this clean image" – he handed me a picture of a perfect print, taken at a police booking – "and say, finally, it's one man's print. You have got to look at everything, not just points. The Bureau has not had a missed ident in all their years of working, and I applaud that. But they are not testing their experts' ability. And that is dangerous."

The following week, Stephen Meagher agreed to speak with me at the F.B.I. headquarters, on Pennsylvania Avenue in Washington. Meagher is perhaps the best known and most forceful advocate for the view that fin-

gerprint evidence is scientifically valid and that it ought to be welcome in courts.

"But is it really a science?" I asked as soon as we settled down to talk in his office. Meagher said that he didn't think of science as a term that could be easily defined or tailored to fit all disciplines in the same way. "There is academic science, legal science, and forensic science," he told me. "They are different. You can be an expert in the field and give testimony without having an academic level of scientific knowledge. It is not achievable to take pure science and move it into a legal arena." This seemed surprising, since Meagher had often argued that, when performed correctly, fingerprint analysis is an "objective" science. In 1999, when he was asked in court whether, based on the unique properties of fingerprints, he had an opinion of the error rate associated with his work, he said, "As applied to the scientific methodology, it's zero." (Scientists don't talk this way; it is an axiom among biomedical researchers that nothing in biology is true a hundred per cent of the time.)

Later, when I asked David Faigman, the Hastings law professor, whether it made sense to divide science into legal, academic, and forensic subgroups, he laughed.

"Of course it makes no sense," he said. "Mr. Meagher operates on a sixteenth-century notion – a Francis Bacon idea – of what science is all about. To me, the analogue for law is meteorology. It deals with physics and chemistry – the most basic sciences. Yet it has to make predictions and empirical statements regarding complex reality. That is because so many factors determine the weather that it's really a probabilistic science. And I think fingerprinting is the same."

"Most fields of normal science could pull from the shelf dozens or hundreds, if not thousands, of studies testing their various hypotheses and contentions, which had been conducted over the past decades or century, and hand them to the court," Michael Saks wrote in "Modern Scientific Evidence." For fingerprinting there was nothing. In 1999, the F.B.I. conducted its study in preparation for the Byron Mitchell trial. The study asked examiners to match the two actual latent prints taken from the car in the Mitchell case with the known set of fingerprints of the man on trial. Both sets of prints were sent to the crime laboratories of fifty-three law-enforcement agencies. Of

the thirty-five agencies that examined them and responded, most concluded that the latent prints matched the known prints of the accused; eight said that no match could be made for one of the latent prints, and six said that no match could be made for the other print. The F.B.I., realizing it had a problem, sent annotated enlargements of all the prints to those examiners who had said the fingerprints couldn't be matched. In these photographs, the points of similarity on the fingertips were clearly marked. This time, every lab adopted the F.B.I.'s conclusions.

When I asked Meagher about the study, he told me that the test was supposed to demonstrate the uniqueness of the prints; it was not meant to be a test of competency. He claimed opponents have used the data unfairly. At the same time, he conceded that it would not matter how clean a fingerprint was if the person examining it hadn't been trained properly. "Our system is a huge statistical-probability model, but it doesn't make identifications, because it doesn't have all the information that is needed," he said. "It's a job for human beings."

On March 13th, Judge Pollak vacated his earlier order. He issued a new opinion, in which he stated that the defense had succeeded in raising "real questions about the adequacy of the proficiency tests taken annually by certified F.B.I. fingerprint examiners." Yet he was persuaded by the F.B.I.'s record of accuracy, and wrote that "whatever may be the case for other law-enforcement agencies" the Bureau's standards seemed good enough to permit F.B.I. experts to testify in his courtroom. "In short," he concluded, "I have changed my mind." It was, naturally, a blow to the opposition – though Pollak was careful to rule only on the case before him and only with regard to the F.B.I.

I met with the Judge shortly after he issued his decision. Having arrived early for our meeting, I watched as he led the jury-selection process in the case in which Meagher will now be permitted to testify. Like most courtrooms, it was decorated with an American flag, but it was filled with art as well: prints by Matisse, Czanne, and Eakins and drawings by Victor Hugo lined the walls.

During the lunch break, we sat in his ramshackle office. The stuffing was falling out of both of our chairs. Pollak, a lively man in his late seventies,

declined to talk specifically about the case, but was happy to consider the broader issues it raised. "The most important question here, of course, is, Am I the right person to be a gatekeeper?" he said. "I, who know little of science. As society comes to rely more fully on technology, the question will become acute." Pollak said that he found it worrisome that the Supreme Court ruling in the Daubert case meant that he could rule one way on an issue like fingerprints and another federal judge in a different jurisdiction could do the opposite, and neither ruling would be reversed (the Court will hear appeals only on procedure, not on the law). He was frank about how poorly prepared most judges are for making decisions based on scientific issues.

"I want to tell you that shortly after I got into this line of work there was no more unqualified district judge" – for making such decisions – "in the United States," Judge Pollak said of himself. He told me that in the early nineteen-eighties he had met a former chief executive of Dupont at a reception. "He asked me how it can be that people like me are entrusted to make such major scientific decisions. He wasn't questioning my good faith. But by virtue of my job I have been asked to make decisions that are out of the range of any competence that I have." Pollak conceded that the DuPont chairman had a point. I asked if he felt scientifically competent to rule on the current case in Philadelphia. He laughed but didn't answer. "I knew when I decided the thing there was going to be some surprise," he said, referring to his initial opinion. "Honestly, I don't think I had anticipated the degree to which people would be startled. Other lawyers in fingerprint situations are now almost duty bound to raise these questions and challenges again. How could they in good faith act in any other way? This decision is certainly not the end. I think we can be certain of that."

### 0.0.9 Under Suspicion, Atul Gawande (*New Yorker*), January 8, 2001

January 8, 2001

Atul Gawande (*New Yorker*)

The fugitive science of criminal justice.

In 1901, a professor of criminal law at the University of Berlin was lecturing to his class when a student suddenly shouted an objection to his line of argument. Another student countered angrily, and the two exchanged insults. Fists were clenched, threats made: "If you say another word ... " Then the first student drew a gun, the second rushed at him, and the professor recklessly interposed himself between them. A struggle, a blast — then pandemonium.

Whereupon the two putative antagonists disengaged and returned to their seats. The professor swiftly restored order, explaining to his students that the incident had been staged, and for a purpose. He asked the students, as eyewitnesses, to describe exactly what they had seen. Some were to write down their account on the spot, some a day or a week later; a few even had to depose their observations under cross-examination. The results were dismal. The most accurate witness got twenty-six per cent of the significant details wrong; others up to eighty per cent. Words were put in people's mouths. Actions were described that had never taken place. Events that *had* taken place disappeared from memory.

In the century since, professors around the world have reenacted the experiment, in one form or another, thousands of times; the findings have been recounted in legal texts, courtrooms, and popular crime books. The trick has even been played on audiences of judges. The implications are not trivial. Each year, in the United States, more than seventy-five thousand people become criminal suspects based on eyewitness identification, with lineups used as a standard control measure. Studies of wrongful convictions — cases where a defendant was later exonerated by DNA testing — have shown the most common cause to be eyewitness error. In medicine, this kind of systematic misdiagnosis would receive intense scientific scrutiny. Yet the legal profes-

sion has conducted no further experiments on the reliability of eyewitness evidence, or on much else, for that matter. Science finds its way to the court house in the form of "expert testimony" – forensic analysis, ballistics, and so forth. But the law has balked at submitting its methods to scientific inquiry. Meanwhile, researchers working outside the legal establishment have discovered that surprisingly simple changes in legal procedures could substantially reduce misidentification. They suggest how scientific experimentation, which transformed medicine in the last century, could transform the justice system in the next.

For more than two decades now, the leading figure in eyewitness research has been a blond, jeans-and-tweed-wearing Midwesterner named Gary Wells. He got involved in the field by happenstance: one morning in 1974, a packet from a Cincinnati defense attorney arrived at the department of psychology at Ohio State University, in Columbus, where Wells was a twenty-three-year-old graduate student. The attorney had written to see if anyone there could help him analyze a case in which he believed his client had been wrongly identified as an armed robber. Inside the envelope was a large black-and-white photograph of the lineup from which his client had been picked out. Digging around a little in his spare time, Wells was surprised to discover that little was known about how misidentification occurs. He corresponded with the attorney several times during the next year, though he never came up with anything useful. The suspect was tried, convicted, and set to prison. Wells never did find out whether the client had been falsely identified. But the case got him thinking.

Some months later, he put together his first experiment. He asked people in a waiting room to watch a bag while he left the room. After he went out, a confederate got up and grabbed the bag. Then he dropped it and picked it up again, giving everyone a good look at him, and bolted. (One problem emerged in the initial experiment: some people gave chase. Wells had to provide his shill with a hiding place just outside the room.) Wells knew from all the previous demonstrations that people would often misidentify the perpetrator. Still, he figured, if they did it without great assurance it wouldn't matter much: under directions that the Supreme Court laid out in 1972, courts placed strong weight on an eyewitness's level of certainty. Wells

found, however, that the witnesses who picked the wrong person out of the lineup were just as confident about their choices as those who identified the right person. In a later experiment, he assembled volunteer juries and had them observe witnesses under cross-examination. The jurors, it turned out, believed inaccurate witnesses just as often as they did accurate ones.

Wells tried variations on these experiments, first at the University of Alberta and later at Iowa State, where he's now a professor of psychology; but after a time even he found the work discouraging. He did not just want to show how things go wrong, he wanted to figure out how they could be improved. His first clue came after several years, when he noticed an unexpected pattern having multiple witnesses did not insure accurate identifications. In his studies, a crime might be witnessed by dozens of people, yet they would often finger the same wrong suspect. The errors were clearly not random.

To investigate further, Wells staged another crime, over and over, until he had gathered two hundred witnesses. The subjects were seated in a room, filling out what they thought were applications for a temporary job, when a loud crash came from behind the door to an adjacent room. A stranger (a graying, middle-aged, mustached local whom Wells had hired) then burst through the door, stopped in his tracks in evident surprise at finding people in the room, and retreated through the same door. Apparently finding a dead end that way, the man rushed in again, dropped an expensive-looking camera, picked it up, and ran out through the exit at the opposite end of the room. Everyone got several good looks at him. At this point, another person dashed in and said, "What happened to my camera?" Wells tested each witness, one by one. Half the group was given a photo lineup of six people — a "six-pack" as the police call it — which included the actual perpetrator. (Police use photo lineups far more frequently than live ones.) In a group of a hundred individuals, fifty-four picked the perpetrator correctly; twenty-one said they didn't think the guy was there; and the others spread their picks across the people in the lineup.

The second group of witnesses was given the same lineup, minus the perpetrator. This time, thirty-two people picked no one. But most of the rest chose the same wrong person — the one who most resembled the perpetrator. Wells theorizes that witnesses faced with a photo spread tend to make

a relative decision, weighing one candidate against the others and against incomplete traces of memory. Studies of actual wrongful convictions lend support to the thesis. For example, in a study of sixty-three DNA exonerations of wrongfully convicted people, fifty-three involved witnesses making a mistaken identification, and almost invariably they had viewed a lineup in which the actual perpetrator was not there. "The dangerous situation is exactly what our experiments said it would be," Wells says.

Once this was established, he and others set about designing ways to limit such errors. Researchers at the State University of New York at Plattsburgh discovered that witnesses who were not explicitly warned that a lineup may not include the actual perpetrator are substantially more likely to make a false identification, under the misapprehension that they've got to pick someone. Wells found that putting more than one suspect in a lineup — something the police do routinely — also dramatically increases errors. Most provocative, however, were the experiments performed by Wells and Rod Lindsay, a colleague from Queen's University in Ontario, which played with the way lineups were structured. The convention is to show a witness a whole lineup at once. Wells and Lindsay decided to see what would happen if witnesses were shown only one person at a time, and made to decide whether he was the culprit before moving on. Now, after a staged theft, the vast majority of witnesses who were shown a lineup that did not include the culprit went through the set without picking anyone. And when the culprit was present, witnesses who viewed a sequential lineup were no less adept at identifying him than witnesses who saw a standard lineup. The innovation reduced false identifications by well over fifty per cent without sacrificing correct identifications. The results have since been replicated by others. And the technique is beautifully simple. It wouldn't cost a dime to adopt it.

It has now been fifteen years since Wells and Lindsay published their results. I asked Wells how widely the procedure has been followed. He laughed, because, aside from a scattered handful of police departments, mainly in Canada, it was not picked up at all. "In general," he told me, "the reaction before criminal-law audiences was 'Well, that's very interesting, but ... ' " A Department of Justice report released in 1999 acknowledged that scientific evidence had established the superiority of sequential-lineup procedures. Yet

the report goes on to emphasize that the department still has no preference between the two methods.

Among the inquisitive and scientifically minded, there are a few peculiar souls for whom the justice system looks the way the human body once did for eighteenth-century anatomists. They see infirmities to be understood, remedies to be invented and tested. And eyewitness identification is just one of the practices that invite empirical scrutiny. Unfortunately, only a handful of scientists have had any luck in gaining access to courtrooms and police departments. One of them is Lawrence Sherman, a sociologist at the University of Pennsylvania, who is the first person to carry out a randomized field experiment in criminal enforcement methods. In 1982, with the support of Minneapolis Police Chief Anthony Bouza, Sherman and his team of researchers completed a seventeen-month trial in which they compared three tactics for responding to non-life-threatening domestic-violence calls: arrest, mediation, and ordering the violent husband or boyfriend to leave the home for eight hours. Arrest emerged as the most effective way to prevent repeated acts of violence. The research was tremendously influential. Previously, it had been rare to arrest a violent husband, at least where the assault was considered "non-severe." Afterward, across the country, arrest became a standard police response.

Such cooperation from law enforcement has proved rare. In Broward County, Florida, researchers started a randomized study to see whether counseling for convicted wife-beaters reduced repeat violence — and prosecutors went to court to stop the study. The state of Florida had granted judges discretion in mandating such counseling, and there was a strong belief that is should be assigned broadly; to stop violence, not randomly, for the sake of study. ("No one is suggesting counseling is a panacea and will solve everyone's problems," the lead prosecutor told the local newspaper, "but I think everyone will agree, in a certain percentage of cases it works.") The researchers managed to get most of the men through the study before it was shut down, though, and they discovered not only that counseling provided no benefit but that it actually increased the likelihood of re-arrest in unemployed men. (Probably that's because the women misguidedly believed that counseling worked, and were more likely to agree to see the men again.) In the field of

law enforcement, people simply do not admit such possibilities, let alone test them.

Consider the jury box. Steven Penrod, a professor of both psychology and law at the University of Nebraska at Lincoln and another lonely pioneer in this area, is happy to rattle off a series of unexplored questions. Are there certain voting arrangements that make false convictions or mistaken acquittals less likely? (Most states require jurors to reach a unanimous verdict for a criminal conviction, but others allow conviction by as few as eight out of twelve jurors.) How would changing the number of jurors seated — say, to three or seventeen or eight — affect decisions? Do jurors understand and follow the instructions that judges give them? What instructions would be most effective in helping juries reach an accurate and just decision? Are there practical ways of getting juries to disregard inadmissible testimony that a lawyer has brought in? These are important questions, but researchers have little hope of making their way into jury rooms.

Lawrence Sherman points out that one of the most fertile areas for work is that of prosecutorial discretion. Most criminal cases are handled outside the courtroom, and no one knows how prosecutors decide whom to prosecute, how effectively they make these decisions, how often they let risky people go, and so on. But he reports that prosecutors he has approached have been "uniformly opposed" to allowing observation, let alone experimental study: "I've proposed repeatedly, and I've failed," Sherman told me. He has a difficult enough time getting cooperation from the police, he says, "but the lawyers are by far the worst." In his view, the process of bringing scientific scrutiny to the methods of the justice system has hardly begun. "We're holding a tiny little cardboard match in the middle of a huge forest at night," he told me. "We're about where surgery was a century ago."

Researchers like Sherman say that one of their problems is the scarcity of financial support. The largest source of research funding is an obscure government agency called the National Institute of Justice, which was modeled on the National Institutes of Health when it was established, in 1968, but has a budget of less than one per cent of the N.I.H.'s. (The government spends more on meat and poultry research.) The harder problem, though, is the clash of cultures between the legal and the scientific approach, which

is compounded by ignorance and suspicion. In medicine, there are hundreds of academic teaching hospitals, where innovation and testing are a routine part of what doctors do. There is no such thing as an academic police department or a teaching courthouse. The legal system takes its methods for granted: it is common sense that lineups are to be trusted, that wife-beaters are to be counseled, and that jurors are not to ask witnesses questions. Law enforcement, finally, is in thrall to a culture of precedent and convention, not of experiment and change. And science remains deeply mistrusted.

"The legal system doesn't understand science," Gary Wells told me. "I taught in law school for a year. Believe me, there's no science in there at all." When he speaks to people in the justice system about his work, he finds that most of his time is spent educating them about basic scientific methods. "To them, it seems like magic hand-waving and — boom — here's the result. So then all they want to know is whose side you're on — the prosecutor's or the defendant's." In an adversarial system, where even facts come in two versions, it's easy to view science as just another form of spin.

For a scientist, Gary Wells is a man of remarkable faith; he has spent more than twenty-five years doing research at the periphery of his own field for an audience that has barely been listening. When I point this out to him, it makes him chuckle. "It's true," he admits, and yet it does not seem to trouble him. "This may be my American optimism talking, but don't you think, in the long run, the better idea will prevail?"

Lately, he has become fascinated with the alibi. "You know," he told me in a recent conversation, "one of the strange things that pop up in DNA-exoneration cases is that innocent people often seem to be done in by weak or inconsistent alibis." And it has got him thinking. Alibis seem so straightforward. The detective asks the suspect, "Where were you last Friday around 11 P.M.?" And if the suspect can't account for his whereabouts — or worse, gives one story now and another later — we take that as evidence against him. But should we? Wells wonders. How well do people remember where they were? How often do they misremember and change their minds? What times of the day is a person likely to have a provable alibi and what time not? How much does this vary among people who are married, who live alone, who are unemployed? Are there better ways to establish whether a suspect has a

legitimate alibi? “No one knows these things,” he says.

### 0.0.10 Annals of Medicine: The Dictionary of Disorder, Alix Spiegel (*New Yorker*), January 3, 2005

January 3, 2005

Alix Spiegel (*New Yorker*)

How one man revolutionized psychiatry.

In the mid-nineteen-forties, Robert Spitzer, a mathematically minded boy of fifteen, began weekly sessions of Reichian psychotherapy. Wilhelm Reich was an Austrian psychoanalyst and a student of Sigmund Freud who, among other things, had marketed a device that he called the orgone accumulator – an iron appliance, the size of a telephone booth, that he claimed could both enhance sexual powers and cure cancer. Spitzer had asked his parents for permission to try Reichian analysis, but his parents had refused – they thought it was a sham – and so he decided to go to the sessions in secret. He paid five dollars a week to a therapist on the Lower East Side of Manhattan, a young man willing to talk frankly about the single most compelling issue Spitzer had yet encountered: women. Spitzer found this methodical approach to the enigma of attraction both soothing and invigorating. The real draw of the therapy, however, was that it greatly reduced Spitzer's anxieties about his troubled family life: his mother was a "professional patient" who cried continuously, and his father was cold and remote. Spitzer, unfortunately, had inherited his mother's unruly inner life and his father's repressed affect; though he often found himself overpowered by emotion, he was somehow unable to express his feelings. The sessions helped him, as he says, "become alive," and he always looked back on them with fondness. It was this experience that confirmed what would become his guiding principle: the best way to master the wilderness of emotion was through systematic study and analysis.

Robert Spitzer isn't widely known outside the field of mental health, but he is, without question, one of the most influential psychiatrists of the twentieth century. It was Spitzer who took the *Diagnostic and Statistical Manual of Mental Disorders* – the official listing of all mental diseases recognized by the American Psychiatric Association (A.P.A.) – and established it as a scientific instrument of enormous power. Because insurance companies now

require a DSM diagnosis for reimbursement, the manual is mandatory for any mental-health professional seeking compensation. It's also used by the court system to help determine insanity, by social-services agencies, schools, prisons, governments, and, occasionally, as a plot device on "The Sopranos." This magnitude of cultural authority, however, is a relatively recent phenomenon. Although the DSM was first published in 1952 and a second edition (DSM-II) came out in 1968, early versions of the document were largely ignored. Spitzer began work on the third version (DSM-III) in 1974, when the manual was a spiral-bound paperback of a hundred and fifty pages. It provided cursory descriptions of about a hundred mental disorders, and was sold primarily to large state mental institutions, for three dollars and fifty cents. Under Spitzer's direction – which lasted through the DSM-III, published in 1980, and the DSM-IIIR ("R" for "revision"), published in 1987 – both the girth of the DSM and its stature substantially increased. It is now nine hundred pages, defines close to three hundred mental illnesses, and sells hundreds of thousands of copies, at eighty-three dollars each. But a mere description of the physical evolution of the DSM doesn't fully capture what Spitzer was able to accomplish. In the course of defining more than a hundred mental diseases, he not only revolutionized the practice of psychiatry but also gave people all over the United States a new language with which to interpret their daily experiences and tame the anarchy of their emotional lives.

The Biometrics Department of the New York State Psychiatric Institute at Columbia Presbyterian Medical Center is situated in an imposing neo-Gothic building on West 168th Street. I met Spitzer in the lobby, a sparsely decorated and strangely silent place that doesn't seem to get much use. Spitzer, a tall, thin man with well-cut clothes and a light step, was brought up on the Upper West Side. He is in his seventies but seems much younger; his graying hair is dyed a deep shade of brown. He has worked at Columbia for more than forty years, and his office is filled with the debris of decades. Calligraphed certificates with seals of red and gold cover the walls, and his desk is overwhelmed by paper.

Spitzer first came to the university as a resident and student at the Columbia Center for Psychoanalytic Training and Research, after graduating from N.Y.U.

School of Medicine in 1957. He had had a brilliant medical-school career, publishing in professional journals a series of well-received papers about childhood schizophrenia and reading disabilities. He had also established himself outside the academy, by helping to discredit his erstwhile hero Reich. In addition to his weekly sessions on the Lower East Side, the teen-age Spitzer had persuaded another Reichian doctor to give him free access to an orgone accumulator, and he spent many hours sitting hopefully on the booth's tiny stool, absorbing healing orgone energy, to no obvious avail. In time, he became disillusioned, and in college he wrote a paper critical of the therapy, which was consulted by the Food and Drug Administration when they later prosecuted Reich for fraud.

At Columbia Psychoanalytic, however, Spitzer's career faltered. Psychoanalysis was too abstract, too theoretical, and somehow his patients rarely seemed to improve. "I was always unsure that I was being helpful, and I was uncomfortable with not knowing what to do with their messiness," he told me. "I don't think I was uncomfortable listening and empathizing – I just didn't know what the hell to do." Spitzer managed to graduate, and secured a position as an instructor in the psychiatry department (he has held some version of the job ever since), but he is a man of tremendous drive and ambition – also a devoted contrarian – and he found teaching intellectually limiting. For satisfaction, he turned to research. He worked on depression and on diagnostic interview techniques, but neither line of inquiry produced the radical innovation or epic discovery that he would need to make his name.

As Spitzer struggled to find his professional footing in the nineteen-sixties, the still young field of psychiatry was also in crisis. The central issue involved the problem of diagnosis: psychiatrists couldn't seem to agree on who was sick and what ailed them. A patient identified as a textbook hysteric by one psychiatrist might easily be classified as a hypochondriac depressive by another. Blame for this discrepancy was assigned to the DSM. Critics claimed that the manual lacked what in the world of science is known as "reliability" – the ability to produce a consistent, replicable result – and therefore also lacked scientific validity. In order for any diagnostic instrument to be considered useful, it must have both. The S.A.T., for example, is viewed as reliable because a person who takes the test on a Tuesday and gets a score of 1200

will get a similar score if he takes the test on a Thursday. It is considered valid because scores are believed to correlate with an external reality – "scholastic aptitude" – and the test is seen as predictive of success in an academic setting. Though validity is the more important measure, it is impossible to achieve validity without reliability: if you take the S.A.T. on a Tuesday and get a 1200 and repeat it on a Thursday and get a 600, the test is clearly not able to gauge academic performance. Reliability, therefore, is the threshold standard.

Problems with the reliability of psychiatric diagnosis became evident during the Second World War, when the military noticed that medical boards in different parts of the country had dramatically different rejection rates for men attempting to enlist. A draft board in Wichita, say, might have a twenty-per-cent exclusion rate, while Baltimore might find sixty per cent of its applicants unfit for service. Much of the disparity was on psychiatric grounds, and this was puzzling. It seemed implausible that the mental stability of potential recruits would vary so greatly from one area to another. A close study of the boards eventually determined that the psychiatrists responsible for making the decisions had widely divergent criteria. So a hypothesis emerged: perhaps it was not the young men but the doctors who were the problem.

In 1949, the psychologist Philip Ash published a study showing that three psychiatrists faced with a single patient, and given identical information at the same moment, were able to reach the same diagnostic conclusion only twenty per cent of the time. Aaron T. Beck, one of the founders of cognitive behavioral therapy, published a similar paper on reliability in 1962. His review of nine different studies found rates of agreement between thirty-two and forty-two per cent. These were not encouraging numbers, given that diagnostic reliability isn't merely an academic issue: if psychiatrists can't agree on a patient's condition, then they can't agree on the treatment of that condition, and, essentially, there's no relationship between diagnosis and cure. In addition, research depends on doctors' ability to form homogeneous subject groups. How can you test the effectiveness of a new drug to treat depression if you can't be sure that the person you're testing is suffering from that disorder? Allen Frances, who worked under Spitzer on the DSM-III and

who, in 1987, was appointed the director of the DSM-IV, says, "Without reliability the system is completely random, and the diagnoses mean almost nothing – maybe worse than nothing, because they're falsely labelling. You're better off not having a diagnostic system."

Spitzer had no particular interest in psychiatric diagnosis, but in 1966 he happened to share a lunch table in the Columbia cafeteria with the chairman of the DSM-II task force. The two struck up a conversation, got along well, and by the end of the meal Spitzer had been offered the job of note-taker on the DSM-II committee. He accepted it, and served ably. He was soon promoted, and when gay activists began to protest the designation of homosexuality as a pathology Spitzer brokered a compromise that eventually resulted in the removal of homosexuality from the DSM. Given the acrimony surrounding the subject, this was an impressive feat of nosological diplomacy, and in the early seventies, when another revision of the DSM came due, Spitzer was asked to be the chairman of the task force.

Today, the chair of the DSM task force is a coveted post – people work for years to position themselves as candidates – but in the early nineteen-seventies descriptive psychiatry was a backwater. Donald Klein, a panic expert at Columbia, who contributed to the DSM-III, says, "When Bob was appointed to the DSM-III, the job was of no consequence. In fact, one of the reasons Bob got the job was that it wasn't considered that important. The vast majority of psychiatrists, or for that matter the A.P.A., didn't expect anything to come from it." This attitude was particularly prevalent among Freudian psychoanalysts, who were the voice of the mental-health profession for much of the twentieth century. They saw descriptive psychiatry as narrow, bloodless, and without real significance. "Psychoanalysts dismiss symptoms as being unimportant, and they say that the real thing is the internal conflicts," Klein says. "So to be interested in descriptive diagnosis was to be superficial and a little bit stupid."

Spitzer, however, managed to turn this obscurity to his advantage. Given unlimited administrative control, he established twenty-five committees whose task it would be to come up with detailed descriptions of mental disorders, and selected a group of psychiatrists who saw themselves primarily as scientists to sit on those committees. These men and women came to be known in

the halls of Columbia as dops, for "data-oriented people." They were deeply skeptical of psychiatry's unquestioning embrace of Freud. "Rather than just appealing to authority, the authority of Freud, the appeal was: Are there studies? What evidence is there?" Spitzer says. "The people I appointed had all made a commitment to be guided by data." Like Spitzer, Jean Endicott, one of the original members of the DSM-III task force, felt frustrated with the rigid dogmatism of psychoanalysis. She says, "For us dops, it was like, Come on – let's get out of the nineteenth century! Let's move into the twentieth, maybe the twenty-first, and apply what we've learned."

There was just one problem with this utopian vision of better psychiatry through science: the "science" hadn't yet been done. "There was very little systematic research, and much of the research that existed was really a hodgepodge – scattered, inconsistent, and ambiguous," Theodore Millon, one of the members of the DSM-III task force, says. "I think the majority of us recognized that the amount of good, solid science upon which we were making our decisions was pretty modest." Members of the various committees would regularly meet and attempt to come up with more specific and comprehensive descriptions of mental disorders. David Shaffer, a British psychiatrist who worked on the DSM-III and the DSM-IIIR, told me that the sessions were often chaotic. "There would be these meetings of the so-called experts or advisers, and people would be standing and sitting and moving around," he said. "People would talk on top of each other. But Bob would be too busy typing notes to chair the meeting in an orderly way." One participant said that the haphazardness of the meetings he attended could be "disquieting." He went on, "Suddenly, these things would happen and there didn't seem to be much basis for it except that someone just decided all of a sudden to run with it." Allen Frances agrees that the loudest voices usually won out. Both he and Shaffer say, however, that the process designed by Spitzer was generally sound. "There was not another way of doing it, no extensive literature that one could turn to," Frances says. According to him, after the meetings Spitzer would retreat to his office to make sense of the information he'd collected. "The way it worked was that after a period of erosion, with different opinions being condensed in his mind, a list of criteria would come up," Frances says. "It would usually be some combination of the accepted

wisdom of the group, as interpreted by Bob, with a little added weight to the people he respected most, and a little bit to whoever got there last."

Because there are very few records of the process, it's hard to pin down exactly how Spitzer and his staff determined which mental disorders to include in the new manual and which to reject. Spitzer seems to have made many of the final decisions with minimal consultation. "He must have had some internal criteria," Shaffer says. "But I don't always know what they were." One afternoon in his office at Columbia, I asked Spitzer what factors would lead him to add a new disease. "How logical it was," he said, vaguely. "Whether it fit in. The main thing was that it had to make sense. It had to be logical." He went on, "For most of the categories, it was just the best thinking of people who seemed to have expertise in the area."

Not every mental disorder made the final cut. For instance, a group of child psychiatrists aspired to introduce a category they called "atypical child" – an idea that, according to Spitzer, didn't survive the first meeting. "I kept saying, 'O.K., how would you define "atypical child"?' And the answer was 'Well, it's very difficult to define, because these kids are all very different.'" As a general rule, though, Spitzer was more interested in including mental disorders than in excluding them. "Bob never met a new diagnosis that he didn't at least get interested in," Frances says. "Anything, however against his own leanings that might be, was a new thing to play with, a new toy." In 1974, Roger Peele and Paul Luisada, psychiatrists at St. Elizabeth's Hospital, in Washington, D.C., wrote a paper in which they used the term "hysterical psychoses" to describe the behavior of two kinds of patients they had observed: those who suffered from extremely short episodes of delusion and hallucination after a major traumatic event, and those who felt compelled to show up in an emergency room even though they had no genuine physical or psychological problems. Spitzer read the paper and asked Peele and Luisada if he could come to Washington to meet them. During a forty-minute conversation, the three decided that "hysterical psychoses" should really be divided into two disorders. Short episodes of delusion and hallucination would be labelled "brief reactive psychosis," and the tendency to show up in an emergency room without authentic cause would be called "factitious disorder." "Then Bob asked for a typewriter," Peele says. To Peele's surprise, Spitzer

drafted the definitions on the spot. "He banged out criteria sets for factitious disorder and for brief reactive psychosis, and it struck me that this was a productive fellow! He comes in to talk about an issue and walks away with diagnostic criteria for two different mental disorders!" Both factitious disorder and brief reactive psychosis were included in the DSM-III with only minor adjustments.

The process of identifying new disorders wasn't usually so improvisatory, though, and it is certain that psychiatric treatment was significantly improved by the designation of many of the new syndromes. Attention-deficit disorder, autism, anorexia nervosa, bulimia, panic disorder, and post-traumatic stress disorder are all examples of diseases added during Spitzer's tenure which now receive specialized treatment. But by far the most radical innovation in the new DSM – and certainly the one that got the most attention in the psychiatric community – was that, alongside the greatly expanded prose descriptions for each disorder, Spitzer added a checklist of symptoms that should be present in order to justify a diagnosis. For example, the current DSM describes a person with obsessive-compulsive personality disorder as someone who:

– is preoccupied with details, rules, lists, order, organization, or schedules to the extent that the major point of the activity is lost ...

– is unable to discard worn-out or worthless objects even when they have no sentimental value ...

– adopts a miserly spending style towards both self and others.

Five other criteria are listed in a box beneath the description of the disorder, and clinicians are cautioned that at least four of the eight must be present in order for the label to be applied.

Finally, Spitzer and the dops argued, here was the answer to the problem of reliability, the issue that had bedevilled psychiatry for years. As they understood it, there were two reasons that doctors couldn't agree on a diagnosis. The first was informational variance: because of rapport or interview style, different doctors get different information from the same patient. The second was interpretive variance: each doctor carries in his mind his own definition of what a specific disease looks like. One goal of the DSM-III was to reduce interpretive variance by standardizing definitions. Spitzer's team

reasoned that if a clear set of criteria were provided, diagnostic reliability would inevitably improve. They also argued that the criteria would enable mental-health professionals to communicate, and greatly facilitate psychiatric research. But the real victory was that each mental disorder could now be identified by a foolproof little recipe.

Spitzer labored over the DSM-III for six years, often working seventy or eighty hours a week. "He's kind of an idiot savant of diagnosis – in a good sense, in the sense that he never tires of it," Allen Frances says. John Talbott, a former president of the American Psychiatric Association, who has been friends with Spitzer for years, says, "I remember the first time I saw him walk into a breakfast at an A.P.A. meeting in a jogging suit, sweating, and having exercised. I was taken aback. The idea that I saw Bob Spitzer away from his suit and computer was mind-shattering." But Spitzer's dedication didn't always endear him to the people he worked with. "He was famous for walking down a crowded hallway and not looking left or right or saying anything to anyone," one colleague recalled. "He would never say hello. You could stand right next to him and be talking to him and he wouldn't even hear you. He didn't seem to recognize that anyone was there."

Despite Spitzer's genius at describing the particulars of emotional behavior, he didn't seem to grasp other people very well. Jean Endicott, his collaborator of many years, says, "He got very involved with issues, with ideas, and with questions. At times he was unaware of how people were responding to him or to the issue. He was surprised when he learned that someone was annoyed. He'd say, 'Why was he annoyed? What'd I do?'" After years of confrontations, Spitzer is now aware of this shortcoming, and says that he struggles with it in his everyday life. "I find it very hard to give presents," he says. "I never know what to give. A lot of people, they can see something and say, 'Oh, that person would like that.' But that just doesn't happen to me. It's not that I'm stingy. I'm just not able to project what they would like." Frances argues that Spitzer's emotional myopia has benefitted him in his chosen career: "He doesn't understand people's emotions. He knows he doesn't. But that's actually helpful in labelling symptoms. It provides less noise."

What may have been a professional strength had disruptive consequences

in Spitzer's personal life. In 1958, he married a doctor, and they had two children. As the demands of his project mounted, he spent less and less time with his family, and eventually fell in love with Janet Williams, an attractive, outspoken social worker he had hired to help edit the manual. In 1979, he and his wife separated, and several years later Spitzer and Williams were married. Williams became a professor at Columbia, and she and Spitzer went on to have three children. Spitzer remained close to his oldest son, but his relationship with his daughter from his first marriage was initially strained by the divorce.

The DSM was scheduled to be published in 1980, which meant that Spitzer had to have a draft prepared in the spring of 1979. Like any major American Psychiatric Association initiative, the DSM had to be ratified by the assembly of the A.P.A., a decision-making body composed of elected officials from all over the country. Spitzer's anti-Freudian ideas had caused resentment throughout the production process, and, as the date of the assembly approached, the opposition gathered strength and narrowed its focus to a single, crucial word – "neurosis" – which Spitzer wanted stricken from the DSM.

The term "neurosis" has a very long history, but over the course of the twentieth century it became inseparable from Freudian psychoanalytic philosophy. A neurosis, Freud believed, emerged from unconscious conflict. This was the bedrock psychoanalytic concept at the height of the psychoanalytic era, and both the DSM-I and the DSM-II made frequent use of the term. Spitzer and the dops, however, reasoned that, because a wide range of mental-health professionals were going to use the manual in everyday practice, the DSM could not be aligned with any single theory. They decided to restrict themselves simply to describing behaviors that were visible to the human eye: they couldn't tell you why someone developed obsessive-compulsive personality disorder, but they were happy to observe that such a person is often "over-conscientious, scrupulous, and inflexible about matters of morality."

When word of Spitzer's intention to eliminate "neurosis" from the DSM got out, Donald Klein says, "people were aghast. 'Neurosis' was the bread-and-butter term of psychiatry, and people thought that we were calling into question their livelihood." Roger Peele, of St. Elizabeth's, was sympathetic

to Spitzer's work, but, as a representative of the Washington, D.C., branch of the A.P.A., he felt a need to challenge Spitzer on behalf of his constituency. "The most common diagnosis in private practices in Washington, D.C., in the nineteen-seventies was something called depressive neurosis," Peele says. "That was what they were doing day after day." Psychoanalysts bitterly denounced the early drafts. One psychiatrist, Howard Berk, wrote a letter to Spitzer saying that "the DSM-III gets rid of the castle of neurosis and replaces it with a diagnostic Levittown."

Without the support of the psychoanalysts, it was possible that the DSM-III wouldn't pass the assembly and the entire project would come to nothing. The A.P.A. leadership got involved, instructing Spitzer and the dops to include psychoanalysts in their deliberations. After months of acrimonious debate, Spitzer and the psychoanalysts were able to reach a compromise: the word "neurosis" was retained in discreet parentheses in three or four key categories.

With this issue resolved, Spitzer presented the final draft of the DSM-III to the A.P.A. assembly in May of 1979. Roughly three hundred and fifty psychiatrists gathered in a large auditorium in Chicago. Spitzer got up onstage and reviewed the DSM process and what they were trying to accomplish, and there was a motion to pass it. "Then a rather remarkable thing happened," Peele says. "Something that you don't see in the assembly very often. People stood up and applauded." Peele remembers watching shock break over Spitzer's face. "Bob's eyes got watery. Here was a group that he was afraid would torpedo all his efforts, and instead he gets a standing ovation."

The DSM-III and the DSM-IIIR together sold more than a million copies. Sales of the DSM-IV (1994) also exceeded a million, and the DSM-IV TR (for "text revision"), the most recent iteration of the DSM, has sold four hundred and twenty thousand copies since its publication, in 2000. Its success continues to grow. Today, there are forty DSM-related products available on the Web site of the American Psychiatric Association. Stuart Kirk, a professor of public policy at U.C.L.A., and Herb Kutchins, a professor emeritus of social work at California State University, Sacramento, have studied the creation of the modern DSM for more than seventeen years, and they argue that its

financial and academic success can be attributed to Spitzer's skillful salesmanship. According to Kirk and Kutchins, immediately after the publication of the DSM-III Spitzer embarked on a P.R. campaign, touting its reliability as "far greater" and "higher than previously achieved" and "extremely good." "For the first time ... claims were made that the new manual was scientifically sound," they write in "Making Us Crazy: DSM – The Psychiatric Bible and the Creation of Mental Disorders" (1997). Gerald Klerman, a prominent psychiatrist, published an influential book in 1986 that flatly announced, "The reliability problem has been solved."

It was largely on the basis of statements like these that the new DSM was embraced by psychiatrists and psychiatric institutions all over the globe. "The DSM revolution in reliability is a revolution in rhetoric, not in reality," Kutchins and Kirk write. Kirk told me, "No one really scrutinized the science very carefully." This was owing, in part, to the manual's imposing physical appearance. "One of the objections was that it appeared to be more authoritative than it was. The way it was laid out made it seem like a textbook, as if it was a depository of all known facts," David Shaffer says. "The average reader would feel that it carried great authority and weight, which was not necessarily merited."

Almost immediately, the book started to turn up everywhere. It was translated into thirteen languages. Insurance companies, which expanded their coverage as psychotherapy became more widespread in the nineteen-seventies, welcomed the DSM-III as a standard. But it was more than that: the DSM had become a cultural phenomenon. There were splashy stories in the press, and TV news magazines showcased several of the newly identified disorders. "It was a runaway success in terms of publicity," Allen Frances says. Spitzer, Williams, and the rest of the dops were surprised and pleased by the reception. "For us it was kind of like being rock stars," Williams says. "Because everyone saw that it was the next big thing, everyone knew us and wanted to talk to us. It was like suddenly being the most popular kid on the block."

A year and a half after the publication of the DSM-III, Spitzer began work on its revision. Emboldened by his success, he became still more adamant about his opinions, and made enemies of a variety of groups. "I love con-

troversy," Spitzer admits, "so if there was something that I thought needed to be added that was controversial, so much the better." He enraged feminists when he tried to include a diagnosis termed "masochistic personality disorder," a nonsexual form of masochism which critics claimed implied that some abused wives might be responsible for their own mistreatment. He angered women's groups again when he attempted to designate PMS as a mental disorder ("pre-menstrual dysphoric disorder"). "A lot of what's in the DSM represents what Bob thinks is right," Michael First, a psychiatrist at Columbia who worked on both the DSM-IIIR and DSM-IV, says. "He really saw this as his book, and if he thought it was right he would push very hard to get it in that way." Thus, despite the success of Spitzer's two editions, and despite extensive lobbying on his part, the American Psychiatric Association gave the chairmanship of the DSM-IV task force to Allen Frances. "The American Psychiatric Association decided that they had had enough of Spitzer, and I can understand that," Spitzer says with a note of regret in his voice. "I think that there was a feeling that if the DSM was going to represent the entire profession – which obviously it has to – it would be good to have someone else." This certainly was part of the reason. But Spitzer's colleagues believe that the single-mindedness with which he transformed the DSM also contributed to his eclipse. "I think that Spitzer looked better in III than he did in IIIR," Peele says. "IIIR, for one reason or another, came across as more heavy-handed – 'Spitzer wants it this way!' "

As chair of the DSM-IV, Frances quickly set about constructing a more transparent process. Power was decentralized, there were systematic literature reviews, and the committees were put on notice that, as Frances says, "the wild growth and casual addition" of new mental disorders were to be avoided. Spitzer was made special adviser to the DSM-IV task force, but his power was dramatically reduced. He found the whole experience profoundly distressing. "I had the feeling that this wonderful thing that I created was going to be destroyed," he says.

The official position of the American Psychiatric Association is that the reliability of the DSM is sound. Darrel Regier, the director of research at the A.P.A., says, "Reliability is, of course, improved. Because you have the criteria, you're not depending on untestable theories of the cause of a diagno-

sis." He says that psychiatric practice was so radically changed by Spitzer's DSM – it was, for the first time, at least nominally evidence-based – that it's impossible to compare reliability before and after. One consequence of the addition of diagnostic criteria was the creation of long, structured interviews, which have allowed psychiatrists successfully to assemble homogeneous research populations for clinical trials. In this context, the DSM diagnoses have been found to be reliable.

But structured interviews don't always have much in common with the conversations that take place in therapists' offices, and since the publication of the DSM-III, in 1980, no major study has been able to demonstrate a substantive improvement in reliability in those less formal settings. During the production of the DSM-IV, the American Psychiatric Association received funding from the MacArthur Foundation to undertake a broad reliability study, and although the research phase of the project was completed, the findings were never published. The director of the project, Jim Thompson, says that the A.P.A. ran out of money. Another study, whose primary author was Spitzer's wife, Janet Williams, took place at six sites in the United States and one in Germany. Supervised by Williams and some of the most experienced diagnostic professionals in the world, the participating clinicians were given extensive special training before being split into pairs and asked to interview nearly six hundred prospective patients. The idea was to determine whether clinicians faced with the same client could agree on a diagnosis using the DSM. Although Williams claims that the study supported the reliability of the DSM, when the investigators wrote up their results they admitted that they "had expected higher reliability values." In fact, Kutchins and Kirk point out, the results were "not that different from those statistics achieved in the 1950s and 1960s – and in some cases were worse."

Reliability is probably lowest in the place where the most diagnoses are made: the therapist's office. As Tom Widiger, who served as head of research for the DSM-IV, points out, "There are lots of studies which show that clinicians diagnose most of their patients with one particular disorder and really don't systematically assess for other disorders. They have a bias in reference to the disorder that they are especially interested in treating and believe that most of their patients have." Unfortunately, because psychiatry and its

sister disciplines stand under the authoritative banner of science, consumers are often reluctant to challenge the labels they are given. Diagnoses are frequently liberating, helping a person to understand that what he views as a personal failing is actually a medical problem, but they can in certain cases become self-fulfilling prophecies. A child inappropriately given the label of attention-deficit/hyperactivity disorder can come to see himself as broken or limited, and act accordingly. And there are other problems with the DSM. Critics complain that it often characterizes everyday behaviors as abnormal, and that it continues to lack validity, whether or not the issue of reliability has been definitely resolved.

Even some of the manual's early advocates now think that the broad claims of reliability were exaggerated. "To my way of thinking, the reliability of the DSM – although improved – has been oversold by some people," Allen Frances says. "From a cultural standpoint, reliability was a way of authenticating the DSM as a radical innovation." He adds, "In a vacuum, to create criteria that were based on accepted wisdom as a first stab was fine, as long as you didn't take it too seriously. The processes that happened were very limited, but they were valuable in their context." And Frances believes that both psychiatry and the public have benefitted in a less tangible way from the collective fantasy that the DSM was a genuine scientific tool. "In my view, if I had been doing the DSM-III it would never have been as famous a document, because I'm a skeptic," he says. "But it was good for the world at large. Good for psychiatry, good for patients. Good for everyone at that point in time to have someone whose view may have been more simpleminded than the world really is. A more complex view of life at that point would have resulted in a ho-hum 'We have this book and maybe it will be useful in our field.' The revolution came not just from the material itself, from the substance of it, but from the passion with which it was introduced."

Spitzer, too, has grown more circumspect. "To say that we've solved the reliability problem is just not true," he told me one afternoon in his office at Columbia. "It's been improved. But if you're in a situation with a general clinician it's certainly not very good. There's still a real problem, and it's not clear how to solve the problem." His personal investment in the DSM remains intense. During one of our conversations, I asked Spitzer if he ever feels a

sense of ownership when troubled friends speak to him of their new diagnoses, or perhaps when he comes across a newspaper account that features one of the disorders to which he gave so much of his life. He admitted that he does on occasion feel a small surge of pride. “My fingers were on the typewriter that typed those. They might have been changed somewhat, but they all went through my fingers,” he said. “Every word.”

### 0.0.11 Personality Plus, Malcolm Gladwell (*New Yorker*), September 20, 2004

September 20, 2004

Malcolm Gladwell (*New Yorker*)

Employers love personality tests. But what do they really reveal?

When Alexander (Sandy) Nininger was twenty-three, and newly commissioned as a lieutenant in the United States Army, he was sent to the South Pacific to serve with the 57th Infantry of the Philippine Scouts. It was January, 1942. The Japanese had just seized Philippine ports at Vigan, Legazpi, Lamon Bay, and Lingayen, and forced the American and Philippine forces to retreat into Bataan, a rugged peninsula on the South China Sea. There, besieged and outnumbered, the Americans set to work building a defensive line, digging foxholes and constructing dikes and clearing underbrush to provide unobstructed sight lines for rifles and machine guns. Nininger's men were on the line's right flank. They labored day and night. The heat and the mosquitoes were nearly unbearable.

Quiet by nature, Nininger was tall and slender, with wavy blond hair. As Franklin M. Reck recounts in "Beyond the Call of Duty," Nininger had graduated near the top of his class at West Point, where he chaired the lecture-and-entertainment committee. He had spent many hours with a friend, discussing everything from history to the theory of relativity. He loved the theatre. In the evenings, he could often be found sitting by the fireplace in the living room of his commanding officer, sipping tea and listening to Tchaikovsky. As a boy, he once saw his father kill a hawk and had been repulsed. When he went into active service, he wrote a friend to say that he had no feelings of hate, and did not think he could ever kill anyone out of hatred. He had none of the swagger of the natural warrior. He worked hard and had a strong sense of duty.

In the second week of January, the Japanese attacked, slipping hundreds of snipers through the American lines, climbing into trees, turning the battlefield into what Reck calls a "gigantic possum hunt." On the morning of January 12th, Nininger went to his commanding officer. He wanted, he said, to be assigned to another company, one that was in the thick of the action, so he

could go hunting for Japanese snipers.

He took several grenades and ammunition belts, slung a Garand rifle over his shoulder, and grabbed a sub machine gun. Starting at the point where the fighting was heaviest – near the position of the battalion's K Company – he crawled through the jungle and shot a Japanese soldier out of a tree. He shot and killed snipers. He threw grenades into enemy positions. He was wounded in the leg, but he kept going, clearing out Japanese positions for the other members of K Company, behind him. He soon ran out of grenades and switched to his rifle, and then, when he ran out of ammunition, used only his bayonet. He was wounded a second time, but when a medic crawled toward him to help bring him back behind the lines Nininger waved him off. He saw a Japanese bunker up ahead. As he leaped out of a shell hole, he was spun around by a bullet to the shoulder, but he kept charging at the bunker, where a Japanese officer and two enlisted men were dug in. He dispatched one soldier with a double thrust of his bayonet, clubbed down the other, and bayonetted the officer. Then, with outstretched arms, he collapsed face down. For his heroism, Nininger was posthumously awarded the Medal of Honor, the first American soldier so decorated in the Second World War.

Suppose that you were a senior Army officer in the early days of the Second World War and were trying to put together a crack team of fearless and ferocious fighters. Sandy Nininger, it now appears, had exactly the right kind of personality for that assignment, but is there any way you could have known this beforehand? It clearly wouldn't have helped to ask Nininger if he was fearless and ferocious, because he didn't know that he was fearless and ferocious. Nor would it have worked to talk to people who spent time with him. His friend would have told you only that Nininger was quiet and thoughtful and loved the theatre, and his commanding officer would have talked about the evenings of tea and Tchaikovsky. With the exception, perhaps, of the Scarlet Pimpernel, a love of music, theatre, and long afternoons in front of a teapot is not a known predictor of great valor. What you need is some kind of sophisticated psychological instrument, capable of getting to the heart of his personality.

Over the course of the past century, psychology has been consumed with the search for this kind of magical instrument. Hermann Rorschach pro-

posed that great meaning lay in the way that people described inkblots. The creators of the Minnesota Multiphasic Personality Inventory believed in the revelatory power of true-false items such as "I have never had any black, tarry-looking bowel movements" or "If the money were right, I would like to work for a circus or a carnival." Today, Annie Murphy Paul tells us in her fascinating new book, "Cult of Personality," that there are twenty-five hundred kinds of personality tests. Testing is a four-hundred-million-dollar-a-year industry. A hefty percentage of American corporations use personality tests as part of the hiring and promotion process. The tests figure in custody battles and in sentencing and parole decisions. "Yet despite their prevalence – and the importance of the matters they are called upon to decide – personality tests have received surprisingly little scrutiny," Paul writes. We can call in the psychologists. We can give Sandy Nininger a battery of tests. But will any of it help?

One of the most popular personality tests in the world is the Myers-Briggs Type Indicator (M.B.T.I.), a psychological-assessment system based on Carl Jung's notion that people make sense of the world through a series of psychological frames. Some people are extroverts, some are introverts. Some process information through logical thought. Some are directed by their feelings. Some make sense of the world through intuitive leaps. Others collect data through their senses. To these three categories – (I)ntroversion/(E)xtroversion, i(N)tuition/(S)ensing, (T)hinking/(F)eeling – the Myers-Briggs test adds a fourth: (J)udging/(P)erceiving. Judgers "like to live in a planned, orderly way, seeking to regulate and manage their lives," according to an M.B.T.I. guide, whereas Perceivers "like to live in a flexible, spontaneous way, seeking to experience and understand life, rather than control it." The M.B.T.I. asks the test-taker to answer a series of "forced-choice" questions, where one choice identifies you as belonging to one of these paired traits. The basic test takes twenty minutes, and at the end you are presented with a precise, multidimensional summary of your personality–your type might be INTJ or ESFP, or some other combination. Two and a half million Americans a year take the Myers-Briggs. Eighty-nine companies out of the Fortune 100 make use of it, for things like hiring or training sessions to help employees "understand" themselves or their colleagues. Annie Murphy

Paul says that at the eminent consulting firm McKinsey, "'associates' often know their colleagues' four-letter M.B.T.I. types by heart," the way they might know their own weight or (this being McKinsey) their S.A.T. scores.

It is tempting to think, then, that we could figure out the Myers-Briggs type that corresponds best to commando work, and then test to see whether Sandy Nininger fits the profile. Unfortunately, the notion of personality type is not nearly as straightforward as it appears. For example, the Myers-Briggs poses a series of items grouped around the issue of whether you – the test-taker – are someone who likes to plan your day or evening beforehand or someone who prefers to be spontaneous. The idea is obviously to determine whether you belong to the Judger or Perceiver camp, but the basic question here is surprisingly hard to answer. I think I'm someone who likes to be spontaneous. On the other hand, I have embarked on too many spontaneous evenings that ended up with my friends and me standing on the sidewalk, looking at each other and wondering what to do next. So I guess I'm a spontaneous person who recognizes that life usually goes more smoothly if I plan first, or, rather, I'm a person who prefers to be spontaneous only if there's someone around me who isn't. Does that make me spontaneous or not? I'm not sure. I suppose it means that I'm somewhere in the middle.

This is the first problem with the Myers-Briggs. It assumes that we are either one thing or another – Intuitive or Sensing, Introverted or Extroverted. But personality doesn't fit into neat binary categories: we fall somewhere along a continuum.

Here's another question: Would you rather work under a boss (or a teacher) who is good-natured but often inconsistent, or sharp-tongued but always logical?

On the Myers-Briggs, this is one of a series of questions intended to establish whether you are a Thinker or a Feeler. But I'm not sure I know how to answer this one, either. I once had a good-natured boss whose inconsistency bothered me, because he exerted a great deal of day-to-day control over my work. Then I had a boss who was quite consistent and very sharp-tongued – but at that point I was in a job where day-to-day dealings with my boss were minimal, so his sharp tongue didn't matter that much. So what do I want in a boss? As far as I can tell, the only plausible answer is: It depends.

The Myers-Briggs assumes that who we are is consistent from one situation to another. But surely what we want in a boss, and how we behave toward our boss, is affected by what kind of job we have.

This is the gist of the now famous critique that the psychologist Walter Mischel has made of personality testing. One of Mischel's studies involved watching children interact with one another at a summer camp. Aggressiveness was among the traits that he was interested in, so he watched the children in five different situations: how they behaved when approached by a peer, when teased by a peer, when praised by an adult, when punished by an adult, and when warned by an adult. He found that how aggressively a child responded in one of those situations wasn't a good predictor of how that same child responded in another situation. Just because a boy was aggressive in the face of being teased by another boy didn't mean that he would be aggressive in the face of being warned by an adult. On the other hand, if a child responded aggressively to being teased by a peer one day, it was a pretty good indicator that he'd respond aggressively to being teased by a peer the next day. We have a personality in the sense that we have a consistent pattern of behavior. But that pattern is complex and that personality is contingent: it represents an interaction between our internal disposition and tendencies and the situations that we find ourselves in.

It's not surprising, then, that the Myers-Briggs has a large problem with consistency: according to some studies, more than half of those who take the test a second time end up with a different score than when they took it the first time. Since personality is continuous, not dichotomous, clearly some people who are borderline Introverts or Feelers one week slide over to Extroversion or Thinking the next week. And since personality is contingent, not stable, how we answer is affected by which circumstances are foremost in our minds when we take the test. If I happen to remember my first boss, then I come out as a Thinker. If my mind is on my second boss, I come out as a Feeler. When I took the Myers-Briggs, I scored as an INTJ. But, if odds are that I'm going to be something else if I take the test again, what good is it?

Once, for fun, a friend and I devised our own personality test. Like the M.B.T.I., it has four dimensions. The first is Canine/Feline. In romantic

relationships, are you the pursuer, who runs happily to the door, tail wagging? Or are you the pursued? The second is More/Different. Is it your intellectual style to gather and master as much information as you can or to make imaginative use of a discrete amount of information? The third is Insider/Outsider. Do you get along with your parents or do you define yourself outside your relationship with your mother and father? And, finally, there is Nibbler/Gobbler. Do you work steadily, in small increments, or do everything at once, in a big gulp? I'm quite pleased with the personality inventory we devised. It directly touches on four aspects of life and temperament–romance, cognition, family, and work style – that are only hinted at by Myers-Briggs. And it can be completed in under a minute, nineteen minutes faster than Myers-Briggs, an advantage not to be dismissed in today's fast-paced business environment. Of course, the four traits it measures are utterly arbitrary, based on what my friend and I came up with over the course of a phone call. But then again surely all universal dichotomous typing systems are arbitrary.

Where did the Myers-Briggs come from, after all? As Paul tells us, it began with a housewife from Washington, D.C., named Katharine Briggs, at the turn of the last century. Briggs had a daughter, Isabel, an only child for whom (as one relative put it) she did "everything but breathe." When Isabel was still in her teens, Katharine wrote a book-length manuscript about her daughter's remarkable childhood, calling her a "genius" and "a little Shakespeare." When Isabel went off to Swarthmore College, in 1915, the two exchanged letters nearly every day. Then, one day, Isabel brought home her college boyfriend and announced that they were to be married. His name was Clarence (Chief) Myers. He was tall and handsome and studying to be a lawyer, and he could not have been more different from the Briggs women. Katharine and Isabel were bold and imaginative and intuitive. Myers was practical and logical and detail-oriented. Katharine could not understand her future son-in-law. "When the blissful young couple returned to Swarthmore," Paul writes, "Katharine retreated to her study, intent on 'figuring out Chief.'" She began to read widely in psychology and philosophy. Then, in 1923, she came across the first English translation of Carl Jung's "Psychological Types." "This is it!" Katharine told her daughter. Paul recounts, "In a dramatic display of conviction she burned all her own research and adopted

Jung's book as her 'Bible,' as she gushed in a letter to the man himself. His system explained it all: Lyman [Katharine's husband], Katharine, Isabel, and Chief were introverts; the two men were thinkers, while the women were feelers; and of course the Briggses were intuitives, while Chief was a senser." Encouraged by her mother, Isabel – who was living in Swarthmore and writing mystery novels – devised a paper-and-pencil test to help people identify which of the Jungian categories they belonged to, and then spent the rest of her life tirelessly and brilliantly promoting her creation.

The problem, as Paul points out, is that Myers and her mother did not actually understand Jung at all. Jung didn't believe that types were easily identifiable, and he didn't believe that people could be permanently slotted into one category or another. "Every individual is an exception to the rule," he wrote; to "stick labels on people at first sight," in his view, was "nothing but a childish parlor game." Why is a parlor game based on my desire to entertain my friends any less valid than a parlor game based on Katharine Briggs's obsession with her son-in-law?

The problems with the Myers-Briggs suggest that we need a test that is responsive to the complexity and variability of the human personality. And that is why, not long ago, I found myself in the office of a psychologist from New Jersey named Lon Gieser. He is among the country's leading experts on what is called the Thematic Apperception Test (T.A.T.), an assessment tool developed in the nineteen-thirties by Henry Murray, one of the most influential psychologists of the twentieth century.

I sat in a chair facing Gieser, as if I were his patient. He had in his hand two dozen or so pictures – mostly black-and-white drawings – on legal-sized cards, all of which had been chosen by Murray years before. "These pictures present a series of scenes," Gieser said to me. "What I want you to do with each scene is tell a story with a beginning, a middle, and an end." He handed me the first card. It was of a young boy looking at a violin. I had imagined, as Gieser was describing the test to me, that it would be hard to come up with stories to match the pictures. As I quickly discovered, though, the exercise was relatively effortless: the stories just tumbled out.

"This is a young boy," I began. "His parents want him to take up the violin, and they've been encouraging him. I think he is uncertain whether he

wants to be a violin player, and maybe even resents the imposition of having to play this instrument, which doesn't seem to have any appeal for him. He's not excited or thrilled about this. He'd rather be somewhere else. He's just sitting there looking at it, and dreading having to fulfill this parental obligation."

I continued in that vein for a few more minutes. Gieser gave me another card, this one of a muscular man clinging to a rope and looking off into the distance. "He's climbing up, not climbing down," I said, and went on:

It's out in public. It's some kind of big square, in Europe, and there is some kind of spectacle going on. It's the seventeenth or eighteenth century. The King is coming by in a carriage, and this man is shimmying up, so he can see over everyone else and get a better view of the King. I don't get the sense that he's any kind of highborn person. I think he aspires to be more than he is. And he's kind of getting a glimpse of the King as a way of giving himself a sense of what he could be, or what his own future could be like.

We went on like this for the better part of an hour, as I responded to twelve cards – each of people in various kinds of ambiguous situations. One picture showed a woman slumped on the ground, with some small object next to her; another showed an attractive couple in a kind of angry embrace, apparently having an argument. (I said that the fight they were having was staged, that each was simply playing a role.) As I talked, Gieser took notes. Later, he called me and gave me his impressions. "What came out was the way you deal with emotion," he said. "Even when you recognized the emotion, you distanced yourself from it. The underlying motive is this desire to avoid conflict. The other thing is that when there are opportunities to go to someone else and work stuff out, your character is always going off alone. There is a real avoidance of emotion and dealing with other people, and everyone goes to their own corners and works things out on their own."

How could Gieser make such a confident reading of my personality after listening to me for such a short time? I was baffled by this, at first, because I felt that I had told a series of random and idiosyncratic stories. When I listened to the tape I had made of the session, though, I saw what Gieser had picked up on: my stories were exceedingly repetitive in just the way that he had identified. The final card that Gieser gave me was blank, and he asked me

to imagine my own picture and tell a story about it. For some reason, what came to mind was Andrew Wyeth's famous painting "Christina's World," of a woman alone in a field, her hair being blown by the wind. She was from the city, I said, and had come home to see her family in the country: "I think she is taking a walk. She is pondering some piece of important news. She has gone off from the rest of the people to think about it." Only later did I realize that in the actual painting the woman is not strolling through the field. She is crawling, desperately, on her hands and knees. How obvious could my aversion to strong emotion be?

The T.A.T. has a number of cards that are used to assess achievement – that is, how interested someone is in getting ahead and succeeding in life. One is the card of the man on the rope; another is the boy looking at his violin. Gieser, in listening to my stories, concluded that I was very low in achievement:

Some people say this kid is dreaming about being a great violinist, and he's going to make it. With you, it wasn't what he wanted to do at all. His parents were making him do it. With the rope climbing, some people do this Tarzan thing. They climb the pole and get to the top and feel this great achievement. You have him going up the rope – and why is he feeling the pleasure? Because he's seeing the King. He's still a nobody in the public square, looking at the King.

Now, this is a little strange. I consider myself quite ambitious. On a questionnaire, if you asked me to rank how important getting ahead and being successful was to me, I'd check the "very important" box. But Gieser is suggesting that the T.A.T. allowed him to glimpse another dimension of my personality.

This idea – that our personality can hold contradictory elements – is at the heart of "Strangers to Ourselves," by the social psychologist Timothy D. Wilson. He is one of the discipline's most prominent researchers, and his book is what popular psychology ought to be (and rarely is): thoughtful, beautifully written, and full of unexpected insights. Wilson's interest is in what he calls the "adaptive unconscious" (not to be confused with the Freudian unconscious). The adaptive unconscious, in Wilson's description, is a big computer in our brain which sits below the surface and evaluates,

filters, and looks for patterns in the mountain of data that come in through our senses. That system, Wilson argues, has a personality: it has a set of patterns and responses and tendencies that are laid down by our genes and our early-childhood experiences. These patterns are stable and hard to change, and we are only dimly aware of them. On top of that, in his schema we have another personality: it's the conscious identity that we create for ourselves with the choices we make, the stories we tell about ourselves, and the formal reasons we come up with to explain our motives and feelings. Yet this "constructed self" has no particular connection with the personality of our adaptive unconscious. In fact, they could easily be at odds. Wilson writes:

The adaptive unconscious is more likely to influence people's uncontrolled, implicit responses, whereas the constructed self is more likely to influence people's deliberative, explicit responses. For example, the quick, spontaneous decision of whether to argue with a co-worker is likely to be under the control of one's nonconscious needs for power and affiliation. A more thoughtful decision about whether to invite a co-worker over for dinner is more likely to be under the control of one's conscious, self-attributed motives.

When Gieser said that he thought I was low in achievement, then, he presumably saw in my stories an unconscious ambivalence toward success. The T.A.T., he believes, allowed him to go beyond the way I viewed myself and arrive at a reading with greater depth and nuance.

Even if he's right, though, does this help us pick commandos? I'm not so sure. Clearly, underneath Sandy Nininger's peaceful faade there was another Nininger capable of great bravery and ferocity, and a T.A.T. of Nininger might have given us a glimpse of that part of who he was. But let's not forget that he volunteered for the front lines: he made a conscious decision to put himself in the heat of the action. What we really need is an understanding of how those two sides of his personality interact in critical situations. When is Sandy Nininger's commitment to peacefulness more, or less, important than some unconscious ferocity? The other problem with the T.A.T., of course, is that it's a subjective instrument. You could say that my story about the man climbing the rope is evidence that I'm low in achievement or you could say that it shows a strong desire for social mobility. The climber wants to look down – not up – at the King in order to get a sense "of what he could be."

You could say that my interpretation that the couple's fighting was staged was evidence of my aversion to strong emotion. Or you could say that it was evidence of my delight in deception and role-playing. This isn't to question Gieser's skill or experience as a diagnostician. The T.A.T. is supposed to do no more than identify themes and problem areas, and I'm sure Gieser would be happy to put me on the couch for a year to explore those themes and see which of his initial hypotheses had any validity. But the reason employers want a magical instrument for measuring personality is that they don't have a year to work through the ambiguities. They need an answer now.

A larger limitation of both Myers-Briggs and the T.A.T. is that they are indirect. Tests of this kind require us first to identify a personality trait that corresponds to the behavior we're interested in, and then to figure out how to measure that trait – but by then we're two steps removed from what we're after. And each of those steps represents an opportunity for error and distortion. Shouldn't we try, instead, to test directly for the behavior we're interested in? This is the idea that lies behind what's known as the Assessment Center, and the leading practitioner of this approach is a company called Development Dimensions International, or D.D.I.

Companies trying to evaluate job applicants send them to D.D.I.'s headquarters, outside Pittsburgh, where they spend the day role-playing as business executives. When I contacted D.D.I., I was told that I was going to be Terry Turner, the head of the robotics division of a company called Global Solutions.

I arrived early in the morning, and was led to an office. On the desk was a computer, a phone, and a tape recorder. In the corner of the room was a video camera, and on my desk was an agenda for the day. I had a long telephone conversation with a business partner from France. There were labor difficulties at an overseas plant. A new product – a robot for the home – had run into a series of technical glitches. I answered e-mails. I prepared and recorded a talk for a product-launch meeting. I gave a live interview to a local television reporter. In the afternoon, I met with another senior Global Solutions manager, and presented a strategic plan for the future of the robotics division. It was a long, demanding day at the office, and when I left, a team of D.D.I. specialists combed through copies of my e-

mails, the audiotapes of my phone calls and my speech, and the videotapes of my interviews, and analyzed me across four dimensions: interpersonal skills, leadership skills, business-management skills, and personal attributes. A few weeks later, I was given my report. Some of it was positive: I was a quick learner. I had good ideas. I expressed myself well, and – I was relieved to hear – wrote clearly. But, as the assessment of my performance made plain, I was something less than top management material:

Although you did a remarkable job addressing matters, you tended to handle issues from a fairly lofty perch, pitching good ideas somewhat unilaterally while lobbing supporting rationale down to the team below. ... Had you brought your team closer to decisions by vesting them with greater accountability, responsibility and decision-making authority, they would have undoubtedly felt more engaged, satisfied and valued. ... In a somewhat similar vein, but on a slightly more interpersonal level, while you seemed to recognize the value of collaboration and building positive working relationships with people, you tended to take a purely businesslike approach to forging partnerships. You spoke of win/win solutions from a business perspective and your rationale for partnering and collaboration seemed to be based solely on business logic. Additionally, at times you did not respond to some of the softer, subtler cues that spoke to people's real frustrations, more personal feelings, or true point of view.

Ouch! Of course, when the D.D.I. analysts said that I did not respond to "some of the softer, subtler cues that spoke to people's real frustrations, more personal feelings, or true point of view," they didn't mean that I was an insensitive person. They meant that I was insensitive in the role of manager. The T.A.T. and M.B.T.I. aimed to make global assessments of the different aspects of my personality. My day as Terry Turner was meant to find out only what I'm like when I'm the head of the robotics division of Global Solutions. That's an important difference. It respects the role of situation and contingency in personality. It sidesteps the difficulty of integrating my unconscious self with my constructed self by looking at the way that my various selves interact in the real world. Most important, it offers the hope that with experience and attention I can construct a more appropriate executive "self." The Assessment Center is probably the best method that employers have for

evaluating personality.

But could an Assessment Center help us identify the Sandy Niningers of the world? The center makes a behavioral prediction, and, as solid and specific as that prediction is, people are least predictable at those critical moments when prediction would be most valuable. The answer to the question of whether my Terry Turner would be a good executive is, once again: It depends. It depends on what kind of company Global Solutions is, and on what kind of respect my co-workers have for me, and on how quickly I manage to correct my shortcomings, and on all kinds of other things that cannot be anticipated. The quality of being a good manager is, in the end, as irreducible as the quality of being a good friend. We think that a friend has to be loyal and nice and interesting – and that's certainly a good start. But people whom we don't find loyal, nice, or interesting have friends, too, because loyalty, niceness, and interestingness are emergent traits. They arise out of the interaction of two people, and all we really mean when we say that someone is interesting or nice is that they are interesting or nice to us.

All these difficulties do not mean that we should give up on the task of trying to understand and categorize one another. We could certainly send Sandy Nininger to an Assessment Center, and find out whether, in a make-believe battle, he plays the role of commando with verve and discipline. We could talk to his friends and discover his love of music and theatre. We could find out how he responded to the picture of the man on a rope. We could sit him down and have him do the Myers-Briggs and dutifully note that he is an Introverted, Intuitive, Thinking Judger, and, for good measure, take an extra minute to run him through my own favorite personality inventory and type him as a Canine, Different, Insider, Gobbler. We will know all kinds of things about him then. His personnel file will be as thick as a phone book, and we can consult our findings whenever we make decisions about his future. We just have to acknowledge that his file will tell us little about the thing we're most interested in. For that, we have to join him in the jungles of Bataan.

### 0.0.12 Head Case: Can Psychiatry Be a Science?, Louis Menand (*New Yorker*), March 1, 2010

March 1, 2010

Louis Menand (*New Yorker*)

The psychiatric literature is so confusing that even the dissidents disagree. Louis Menand looks at the contradictory ways we understand and treat depression.

You arrive for work and someone informs you that you have until five o'clock to clean out your office. You have been laid off. At first, your family is brave and supportive, and although you're in shock, you convince yourself that you were ready for something new. Then you start waking up at 3 A.M., apparently in order to stare at the ceiling. You can't stop picturing the face of the employee who was deputized to give you the bad news. He does not look like George Clooney. You have fantasies of terrible things happening to him, to your boss, to George Clooney. You find – a novel recognition – not only that you have no sex drive but that you don't care. You react irritably when friends advise you to let go and move on. After a week, you have a hard time getting out of bed in the morning. After two weeks, you have a hard time getting out of the house. You go see a doctor. The doctor hears your story and prescribes an antidepressant. Do you take it?

However you go about making this decision, do not read the psychiatric literature. Everything in it, from the science (do the meds really work?) to the metaphysics (is depression really a disease?), will confuse you. There is little agreement about what causes depression and no consensus about what cures it. Virtually no scientist subscribes to the man-in-the-waiting-room theory, which is that depression is caused by a lack of serotonin, but many people report that they feel better when they take drugs that affect serotonin and other brain chemicals.

There is suspicion that the pharmaceutical industry is cooking the studies that prove that antidepressant drugs are safe and effective, and that the industry's direct-to-consumer advertising is encouraging people to demand pills to cure conditions that are not diseases (like shyness) or to get through ordinary life problems (like being laid off). The Food and Drug Administration

has been accused of setting the bar too low for the approval of brand-name drugs. Critics claim that health-care organizations are corrupted by industry largesse, and that conflict-of-interest rules are lax or nonexistent. Within the profession, the manual that prescribes the criteria for official diagnoses, the *Diagnostic and Statistical Manual of Mental Disorders*, known as the D.S.M., has been under criticism for decades. And doctors prescribe antidepressants for patients who are not suffering from depression. People take antidepressants for eating disorders, panic attacks, premature ejaculation, and alcoholism.

These complaints are not coming just from sociologists, English professors, and other troublemakers; they are being made by people within the field of psychiatry itself. As a branch of medicine, depression seems to be a mess. Business, however, is extremely good. Between 1988, the year after Prozac was approved by the F.D.A., and 2000, adult use of antidepressants almost tripled. By 2005, one out of every ten Americans had a prescription for an antidepressant. IMS Health, a company that gathers data on health care, reports that in the United States in 2008 a hundred and sixty-four million prescriptions were written for antidepressants, and sales totalled $9.6 billion. As a depressed person might ask, What does it all mean?

Two new books, Gary Greenberg's "Manufacturing Depression" (Simon & Schuster; $27) and Irving Kirsch's "The Emperor's New Drugs" (Basic; $23.95), suggest that dissensus prevails even among the dissidents. Both authors are hostile to the current psychotherapeutic regime, but for reasons that are incompatible. Greenberg is a psychologist who has a practice in Connecticut. He is an unusually eloquent writer, and his book offers a grand tour of the history of modern medicine, as well as an up-close look at contemporary practices, including clinical drug trials, cognitive-behavioral therapy, and brain imaging. The National Institute of Mental Health estimates that more than fourteen million Americans suffer from major depression every year, and more than three million suffer from minor depression (whose symptoms are milder but last longer than two years). Greenberg thinks that numbers like these are ridiculous – not because people aren't depressed but because, in most cases, their depression is not a mental illness. It's a sane response to a crazy world.

Greenberg basically regards the pathologizing of melancholy and despair, and the invention of pills designed to relieve people of those feelings, as a vast capitalist conspiracy to paste a big smiley face over a world that we have good reason to feel sick about. The aim of the conspiracy is to convince us that it's all in our heads, or, specifically, in our brains – that our unhappiness is a chemical problem, not an existential one. Greenberg is critical of psychopharmacology, but he is even more critical of cognitive-behavioral therapy, or C.B.T., a form of talk therapy that helps patients build coping strategies, and does not rely on medication. He calls C.B.T. "a method of indoctrination into the pieties of American optimism, an ideology as much as a medical treatment."

In fact, Greenberg seems to believe that contemporary psychiatry in most of its forms except existential-humanistic talk therapy, which is an actual school of psychotherapy, and which appears to be what he practices, is mainly about getting people to accept current arrangements. And it's not even that drug companies and the psychiatric establishment have some kind of moral or political stake in these arrangements – that they're in the game in order to protect the status quo. They just see, in the world's unhappiness, a chance to make money. They invented a disease so that they could sell the cure.

Greenberg is repeating a common criticism of contemporary psychiatry, which is that the profession is creating ever more expansive criteria for mental illness that end up labeling as sick people who are just different – a phenomenon that has consequences for the insurance system, the justice system, the administration of social welfare, and the cost of health care.

Jerome Wakefield, a professor of social work at New York University, has been calling out the D.S.M. on this issue for a number of years. In "The Loss of Sadness" (2007), Wakefield and Allan Horwitz, a sociologist at Rutgers, argue that the increase in the number of people who are given a diagnosis of depression suggests that what has changed is not the number of people who are clinically depressed but the definition of depression, which has been defined in a way that includes normal sadness. In the case of a patient who exhibits the required number of symptoms, the D.S.M. specifies only one exception to a diagnosis of depression: bereavement. But, Wakefield and Horwitz point out, there are many other life problems for which intense

sadness is a natural response – being laid off, for example. There is nothing in the D.S.M. to prevent a physician from labeling someone who is living through one of these problems mentally disordered.

The conversion of stuff that people used to live with into disorders that physicians can treat is not limited to psychiatry, of course. Once, people had heartburn ("I can't believe I ate the whole thing") and bought Alka-Seltzer over the counter; now they are given a diagnosis of gastroesophageal reflux disease ("Ask your doctor whether you might be suffering from GERD") and are written a prescription for Zantac. But people tend to find the medicalization of mood and personality more distressing. It has been claimed, for example, that up to 18.7 per cent of Americans suffer from social-anxiety disorder. In "Shyness" (2007), Christopher Lane, a professor of English at Northwestern, argues that this is a blatant pathologization of a common personality trait for the financial benefit of the psychiatric profession and the pharmaceutical industry. It's a case of what David Healy, in his invaluable history "The Antidepressant Era" (1997), calls "the pharmacological scalpel": if a drug (in this case, Paxil) proves to change something in patients (shyness), then that something becomes a disorder to be treated (social anxiety). The discovery of the remedy creates the disease.

Turning shyness into a mental disorder has many downstream consequences. As Steven Hyman, a former director of the National Institute of Mental Health, argues in a recent article, once a diagnosis is ensconced in the manual, it is legitimatized as a subject of scientific research. Centers are established (there is now a Shyness Research Institute, at Indiana University Southeast) and scientists get funding to, for example, find "the gene for shyness" – even though there was never any evidence that the condition has an organic basis. A juggernaut effect is built into the system.

Irving Kirsch is an American psychologist who now works in the United Kingdom. Fifteen years ago, he began conducting meta-analyses of antidepressant drug trials. A meta-analysis is a statistical abstract of many individual drug trials, and the method is controversial. Drug trials are designed for different reasons – some are done to secure government approval for a new drug, and some are done to compare treatments – and they have different processes for everything from selecting participants to measuring outcomes.

Adjusting for these differences is complicated, and Kirsch's early work was roundly criticized on methodological grounds by Donald Klein, of Columbia University, who was one of the key figures in the transformation of psychiatry to a biologically based practice. But, as Kirsch points out, meta-analyses have since become more commonly used and accepted.

Kirsch's conclusion is that antidepressants are just fancy placebos. Obviously, this is not what the individual tests showed. If they had, then none of the drugs tested would have received approval. Drug trials normally test medications against placebos – sugar pills – which are given to a control group. What a successful test typically shows is a small but statistically significant superiority (that is, greater than could be due to chance) of the drug to the placebo. So how can Kirsch claim that the drugs have zero medicinal value?

His answer is that the statistical edge, when it turns up, is a placebo effect. Drug trials are double-blind: neither the patients (paid volunteers) nor the doctors (also paid) are told which group is getting the drug and which is getting the placebo. But antidepressants have side effects, and sugar pills don't. Commonly, side effects of antidepressants are tolerable things like nausea, restlessness, dry mouth, and so on. (Uncommonly, there is, for example, hepatitis; but patients who develop hepatitis don't complete the trial.) This means that a patient who experiences minor side effects can conclude that he is taking the drug, and start to feel better, and a patient who doesn't experience side effects can conclude that she's taking the placebo, and feel worse. On Kirsch's calculation, the placebo effect – you believe that you are taking a pill that will make you feel better; therefore, you feel better – wipes out the statistical difference.

One objection to Kirsch's argument is that response to antidepressants is extremely variable. It can take several different prescriptions to find a medication that works. Measuring a single antidepressant against a placebo is not a test of the effectiveness of antidepressants as a category. And there is a well-known study, called the Sequenced Treatment Alternatives to Relieve Depression, or STAR*D trial, in which patients were given a series of different antidepressants. Though only thirty-seven per cent recovered on the first drug, another nineteen per cent recovered on the second drug, six per cent

on the third, and five per cent after the fourth – a sixty-seven-per-cent effectiveness rate for antidepressant medication, far better than the rate achieved by a placebo.

Kirsch suggests that the result in STAR*D may be one big placebo effect. He cites a 1957 study at the University of Oklahoma in which subjects were given a drug that induced nausea and vomiting, and then another drug, which they were told prevents nausea and vomiting. After the first anti-nausea drug, the subjects were switched to a different anti-nausea drug, then a third, and so on. By the sixth switch, a hundred per cent of the subjects reported that they no longer felt nauseous – even though every one of the anti-nausea drugs was a placebo.

Kirsch concludes that since antidepressants have no more effectiveness than sugar pills, the brain-chemistry theory of depression is "a myth." But, if this is so, how should we treat depression? Kirsch has an answer: C.B.T. He says it really works.

Kirsch's claims appeared to receive a big boost from a meta-analysis published in January in the *Journal of the American Medical Association* and widely reported. The study concludes that "there is little evidence" that antidepressants are more effective than a placebo for minor to moderate depression. But, as a Cornell psychiatrist, Richard Friedman, noted in a column in the *Times*, the meta-analysis was based on just six trials, with a total of seven hundred and eighteen subjects; three of those trials tested Paxil, and three tested imipramine, one of the earliest antidepressants, first used in 1956. Since there have been hundreds of antidepressant drug trials and there are around twenty-five antidepressants on the market, this is not a large sample. The authors of the meta-analysis also assert that "for patients with very severe depression, the benefit of medications over placebo is substantial" – which suggests that antidepressants do affect mood through brain chemistry. The mystery remains unsolved.

Apart from separating us unnecessarily from our money, it's hard to see how a pill that does nothing can also be bad for us. If Kirsch is right and antidepressant drugs aren't doing anything consequential to our brains, then it can't also be the case that they are turning us into Stepford wives or Nietzsche's "last men," the sort of thing that worries Greenberg. By Kirsch's

account, we are in danger of bankrupting our health-care system by spending nearly ten billion dollars a year on worthless pills. But if Greenberg is right we're in danger of losing our ability to care. Is psychopharmacology evil, or is it useless?

The question has been around since the time of Freud. The profession has been the perennial target of critics who, like Greenberg, accuse it of turning deviance into a disorder, and of confusing health with conformity. And it has also been rocked many times by studies that, like Kirsch's, cast doubt on the scientific validity of the entire enterprise.

One of the oldest complaints is that the diagnostic categories psychiatrists use don't match up with the conditions patients have. In 1949, Philip Ash, an American psychologist, published a study in which he had fifty-two mental patients examined by three psychiatrists, two of them, according to Ash, nationally known. All the psychiatrists reached the same diagnosis only twenty per cent of the time, and two were in agreement less than half the time. Ash concluded that there was a severe lack of fit between diagnostic labels and, as he put it, "the complexities of the biodynamics of mental structure" – that is, what actually goes on in people's minds.

In 1952, a British psychologist, Hans Eysenck, published a summary of several studies assessing the effectiveness of psychotherapy. "There ... appears to be an inverse correlation between recovery and psychotherapy," Eysenck dryly noted. "The more psychotherapy, the smaller the recovery rate."

Later studies have shown that patients suffering from depression and anxiety do equally well when treated by psychoanalysts and by behavioral therapists; that there is no difference in effectiveness between C.B.T., which focusses on the way patients reason, and interpersonal therapy, which focusses on their relations with other people; and that patients who are treated by psychotherapists do no better than patients who meet with sympathetic professors with no psychiatric training. Depressed patients in psychotherapy do no better or worse than depressed patients on medication. There is little evidence to support the assumption that supplementing antidepressant medication with talk therapy improves outcomes. What a load of evidence does seem to suggest is that care works for some of the people some of the time, and it doesn't much matter what sort of care it is. Patients believe

that they are being cared for by someone who will make them feel better; therefore, they feel better. It makes no difference whether they're lying on a couch interpreting dreams or sitting in a Starbucks discussing the concept of "flow."

Psychiatry has also been damaged by some embarrassing exposés, such as David Rosenhan's famous article "On Being Sane in Insane Places" (1973), which described the inability of hospital psychiatrists to distinguish mentally ill patients from impostors. The procedure used to determine the inclusion or exclusion of diagnoses in the D.S.M. has looked somewhat unseemly from a scientific point of view. Homosexuality, originally labeled a sociopathic personality disorder, was eliminated from the D.S.M. in 1973, partly in response to lobbying by gay-rights groups. The manual then inserted the category "ego-dystonic homosexuality" – distress because of the presence of homosexual arousal or the absence of heterosexual arousal. Further lobbying eliminated this category as well. Post-traumatic stress disorder was lobbied for by veterans' organizations and resisted by the Veterans Administration, and got in, while self-defeating personality disorder was lobbied against by women's groups, and was deleted.

And there was the rapid collapse of Freudianism. The first two editions of the D.S.M. (the first was published in 1952, the second in 1968) reflected the psychoanalytic theories of Freud and of the Swiss émigré Adolf Meyer, who emphasized the importance of patients' life histories and everyday problems. But the third edition, published in 1980, began a process of scrubbing Freudianism out of the manual, and giving mental health a new language. As Healy puts it, "Where once lay people had gone to psychiatrists expecting to hear about sexual repression, they now came knowing that something might be wrong with their amines or with some brain chemical." A vocabulary that had sunk deep into the popular culture – neurotic, anal, Oedipal – was wiped out of the discipline.

Finally, there has been a blare of criticism surrounding the role of the pharmaceutical industry in research and testing. The industry funds much of the testing done for the F.D.A. Drug companies donate money to hospitals, sponsor posh conferences in exotic locations, provide inducements to physicians to prescribe their drugs, lobby the F.D.A. and Congress – for example,

successfully to prevent Medicare from using its bargaining leverage to reduce the price of medications – and generally use their profits to keep a seat at every table.

So the antidepressant business looks like a demolition derby – a collision of negative research results, questionable research and regulatory practices, and popular disenchantment with the whole pharmacological regime. And it may soon turn into something bigger, something more like a train wreck. If it does, it's worth remembering that we have seen this movie before.

The early history of psychopharmacology is characterized by serendipitous discoveries, and mephenesin was one of them. A Czech émigré named Frank Berger, working in England in the nineteen-forties, was looking for a preservative for penicillin, a drug much in demand by the military. He found that mephenesin had a tranquillizing effect on mice, and published a paper announcing this result in 1946. After the war, Berger moved to the United States and eventually took a job with the drug company that became Carter-Wallace, where he synthesized a compound related to mephenesin called meprobamate. In 1955, Carter-Wallace launched meprobamate as a drug to relieve anxiety. The brand name it invented was Miltown.

Miltown, Andrea Tone says in her cultural history of tranquillizers, "The Age of Anxiety" (2009), was "the first psychotropic blockbuster and the fastest-selling drug in U.S. history." Within a year, one out of every twenty Americans had taken Miltown; within two years, a billion tablets had been manufactured. By the end of the decade, Miltown and Equanil (the same chemical, licensed from Carter-Wallace by a bigger drug company, Wyeth) accounted for a third of all prescriptions written by American physicians. These drugs were eclipsed in the nineteen-sixties by two other wildly popular anxiolytics (anti-anxiety drugs): Librium and Valium, introduced in 1960 and 1963. Between 1968 and 1981, Valium was the most frequently prescribed medication in the Western world. In 1972, stock in its manufacturer, Hoffmann-La Roche, traded at seventy-three thousand dollars a share.

As Tone and David Herzberg, in his cultural history of psychiatric drugs, "Happy Pills in America" (2008) – the books actually complement each other nicely – both point out, the anxiolytics were enmeshed in exactly the same scientific, financial, and ethical confusion as antidepressants today. The F.D.A.

did not permit direct-to-consumer – "Ask your doctor" – advertising until 1985, but the tranquillizer manufacturers invested heavily in promotion. They sent "detail men" – that is, salesmen – to teach physicians about the wonders of their medications. Carter-Wallace was an exception to this, because Berger disliked the idea of salesmanship, but the company took out the front-cover advertisement in the *American Journal of Psychiatry* every month for ten years.

Tranquilizers later became associated with the subjugation of women – "mother's little helpers" – but Miltown was marketed to men, and male celebrities were enlisted to promote the drug. It was particularly popular in Hollywood. Anxiety was pitched as the disorder of high-functioning people, the cost of success in a competitive world. Advertisements for Equanil explained that "anxiety and tension are the commonplace of the age." People on anxiolytics reported that they had never felt this well before – much like the patients Peter Kramer describes in "Listening to Prozac" (1993) who told him that they were "better than well."

Miltown seemed to fit perfectly with the state of psychiatry in the nineteen-fifties. Freud himself had called anxiety "a riddle whose solution would be bound to throw a flood of light on our whole mental existence," and the first edition of the D.S.M. identified anxiety as the "chief characteristic" of all neuroses. The D.S.M. was not widely used in the nineteen-fifties (the situation changed dramatically after 1980), but the idea that anxiety is central to the modern psyche was the subject of two popular books by mental-health professionals, Rollo May's "The Meaning of Anxiety" (1950) and Hans Selye's "The Stress of Life" (1956). (Selye was the person who coined the term "stressor.")

There was a cultural backlash as well. People worried that tranquilizers would blunt America's competitive edge. *Business Week* wrote about the possibility of "tranquil extinction." *The Nation* suggested that tranquilizers might be more destructive than the bomb: "As we watch over the decline of the West, we see the beams – the bombs and the missiles; but perhaps we miss the motes – the pretty little pills."

The weird part of it all was that, for a long time, no one was listening to Miltown. Meprobamate carved out an area of mental functioning and fired a

chemical at it, a magic bullet, and the bullet made the condition disappear. What Miltown was saying, therefore, was that the Freudian theory that neuroses are caused by conflicts between psychic drives was no longer relevant. If you can cure your anxiety with a pill, there is no point spending six years on the couch. And yet, in the nineteen-fifties, references to Freud appeared alongside references to tranquilizers with no suggestion of a contradiction. It took landmark articles by Joseph Schildkraut, in 1965, proposing the amine theory of depression (the theory that Kirsch thinks is a myth), and by Klein (Kirsch's early critic), called "Anxiety Reconceptualized," in 1980, to expose the disjunction within the profession.

The train wreck for tranquilizers arrived in two installments. The first was the discovery that thalidomide, which was prescribed as a sedative, caused birth defects. This led to legislation giving the F.D.A. power to monitor the accuracy of drug-company promotional claims, which slowed down the marketing juggernaut. The second event was the revelation that Valium and Librium can be addictive. In 1980, the F.D.A. required that anxiety medications carry a warning stating that "anxiety or tension associated with the stress of everyday life usually does not require treatment with an anxiolytic." The anxiety era was over. This is one of the reasons that when the SSRIs, such as Prozac, came on the market they were promoted as antidepressants – even though they are commonly prescribed for anxiety. Anxiety drugs had acquired a bad name.

The position behind much of the skepticism about the state of psychiatry is that it's not really science. "Cultural, political, and economic factors, not scientific progress, underlie the triumph of diagnostic psychiatry and the current 'scientific' classification of mental illness entities," Horwitz complained in an earlier book, "Creating Mental Illness" (2002), and many people echo his charge. But is this in fact the problem? The critics who say that psychiatry is not really science are not anti-science themselves. On the contrary: they hold an exaggerated view of what science, certainly medical science, and especially the science of mental health, can be.

Progress in medical science is made by lurching around. The best that can be hoped is that we are lurching in an over-all good direction. One common criticism of contemporary psychiatry has to do with the multiplication of

mental disorders. D.S.M.-II listed a hundred and eighty-two diagnoses; the current edition, D.S.M.-IV-T.R., lists three hundred and sixty-five. There is a reason for this. The goal of biological psychiatry is to identify the organic conditions underlying the symptoms of mental distress that patients complain of. (This was Freud's goal, too, though he had a completely different idea of what the mental events were to which the organic conditions corresponded.) The hope is to establish psychiatry firmly on the disease model of medical practice. In most cases, though, the underlying conditions are either imperfectly known or not known at all. So the D.S.M. lists only disorders – clusters of symptoms, drawn from clinical experience – not diseases. Since people manifest symptoms in an enormous variety of combinations, we get a large number of disorders for what may be a single disease.

Depression is a good example of the problem this makes. A fever is not a disease; it's a symptom of disease, and the disease, not the symptom, is what medicine seeks to cure. Is depression – insomnia, irritability, lack of energy, loss of libido, and so on – like a fever or like a disease? Do patients complain of these symptoms because they have contracted the neurological equivalent of an infection? Or do the accompanying mental states (thoughts that my existence is pointless, nobody loves me, etc.) have real meaning? If people feel depressed because they have a disease in their brains, then there is no reason to pay much attention to their tales of woe, and medication is the most sensible way to cure them. Peter Kramer, in "Against Depression" (2005), describes a patient who, after she recovered from depression, accused him of taking what she had said in therapy too seriously. It was the depression talking, she told him, not her.

Depression often remits spontaneously, perhaps in as many as fifty per cent of cases; but that doesn't mean that there isn't something wrong in the brain of depressed people. Kramer claims that there is a demonstrated link between depression and ill health. Even minor depression raises the risk of death from cardiac disease by half, he says, and contracting depression once increases a patient's susceptibility later in life. Kramer thinks that the notion that depression affords us, as Greenberg puts it, "some glimpse of the way things are" is a myth standing in the way of treating a potentially dangerous disease of the brain. He compares it to the association of tuberculosis with refinement

in the nineteenth century, an association that today seems the opposite of enlightened. "Against Depression" is a plea to attack a biochemical illness with chemicals.

Is depression overdiagnosed? The disease model is no help here. If you have a fever, the doctor runs some tests in order to find out what your problem is. The tests, not the fever, identify the disease. The tests determine, in fact, that there is a disease. In the case of mood disorders, it is difficult to find a test to distinguish mental illness from normal mood changes. The brains of people who are suffering from mild depression look the same on a scan as the brains of people whose football team has just lost the Super Bowl. They even look the same as the brains of people who have been asked to think sad thoughts. As Freud pointed out, you can't distinguish mourning from melancholy just by looking. So a psychiatrist who diagnoses simply by checking off the symptoms listed in the D.S.M. will, as Wakefield and others complain, end up with a lot of false positives. The anti-Freudian bias against the relevance of life histories leaves a lot of holes. But bringing life histories back into the picture isn't going to make diagnoses any more scientific.

Science, particularly medical science, is not a skyscraper made of Lucite. It is a field strewn with black boxes. There have been many medical treatments that worked even though, for a long time, we didn't know why they worked – aspirin, for example. And drugs have often been used to carve out diseases. Malaria was "discovered" when it was learned that it responded to quinine. Someone was listening to quinine. As Nicholas Christakis, a medical sociologist, has pointed out, many commonly used remedies, such as Viagra, work less than half the time, and there are conditions, such as cardiovascular disease, that respond to placebos for which we would never contemplate not using medication, even though it proves only marginally more effective in trials. Some patients with Parkinson's respond to sham surgery. The ostensibly shaky track record of antidepressants does not place them outside the pharmacological pale.

The assumption of many critics of contemporary psychiatry seems to be that if the D.S.M. "carved nature at the joints," if its diagnoses corresponded to discrete diseases, then all those categories would be acceptable. But, as Elliot Valenstein (no friend of biochemical psychiatry) points out in "Blaming

the Brain" (1998), "at some period in history the cause of every 'legitimate' disease was unknown, and they all were at one time 'syndromes' or 'disorders' characterized by common signs and symptoms."

D.S.M.-III was created to address a problem. The problem was reliability, and the manual was an attempt to get the profession on the same page so that every psychiatrist would make the same diagnosis for a given set of symptoms. The manual did not address a different problem, which is validity – the correspondence of symptoms to organic conditions. But if we couldn't treat psychiatric patients until we were certain what the underlying pathology was, we would not be treating most patients. For some disorders, such as depression, we may never know, in any useful way, what the underlying pathology is, since we can't distinguish biologically patients who are suffering from depression from patients who are enduring a depressing life problem.

For many people, this is the most troubling aspect of contemporary psychiatry. These people worry that an easy way is now available to jump the emotional queue, that people can now receive medical enhancements who do not "deserve" them. For example, would you take an antidepressant to get over the pain of being laid off? You might, if you reasoned that since your goal is to get over it and move on, there is no point in prolonging the agony. But you might also reason that learning how to cope with difficulty without a therapeutic crutch is something that it would be good to take away from this disaster. This is not a problem we should expect science to solve for us someday. It's not even a problem that we should want science to solve for us.

Mental disorders sit at the intersection of three distinct fields. They are biological conditions, since they correspond to changes in the body. They are also psychological conditions, since they are experienced cognitively and emotionally – they are part of our conscious life. And they have moral significance, since they involve us in matters such as personal agency and responsibility, social norms and values, and character, and these all vary as cultures vary.

Many people today are infatuated with the biological determinants of things. They find compelling the idea that moods, tastes, preferences, and behaviors can be explained by genes, or by natural selection, or by brain amines (even though these explanations are almost always circular: if we do

x, it must be because we have been selected to do x). People like to be able to say, I'm just an organism, and my depression is just a chemical thing, so, of the three ways of considering my condition, I choose the biological. People do say this. The question to ask them is, Who is the "I" that is making this choice? Is that your biology talking, too?

The decision to handle mental conditions biologically is as moral a decision as any other. It is a time-honored one, too. Human beings have always tried to cure psychological disorders through the body. In the Hippocratic tradition, melancholics were advised to drink white wine, in order to counteract the black bile. (This remains an option.) Some people feel an instinctive aversion to treating psychological states with pills, but no one would think it inappropriate to advise a depressed or anxious person to try exercise or meditation.

The recommendation from people who have written about their own depression is, overwhelmingly, Take the meds! It's the position of Andrew Solomon, in "The Noonday Demon" (2001), a wise and humane book. It's the position of many of the contributors to "Unholy Ghost" (2001) and "Poets on Prozac" (2008), anthologies of essays by writers about depression. The ones who took medication say that they write much better than they did when they were depressed. William Styron, in his widely read memoir "Darkness Visible" (1990), says that his experience in talk therapy was a damaging waste of time, and that he wishes he had gone straight to the hospital when his depression became severe.

What if your sadness was grief, though? And what if there were a pill that relieved you of the physical pain of bereavement – sleeplessness, weeping, loss of appetite – without diluting your love for or memory of the dead? Assuming that bereavement "naturally" remits after six months, would you take a pill today that will allow you to feel the way you will be feeling six months from now anyway? Probably most people would say no.

Is this because of what the psychiatrist Gerald Klerman once called "pharmacological Calvinism"? Klerman was describing the view, which he thought many Americans hold, that shortcuts to happiness are sinful, that happiness is not worth anything unless you have worked for it. (Klerman misunderstood Calvinist theology, but never mind.) We are proud of our children when they

learn to manage their fears and perform in public, and we feel that we would not be so proud of them if they took a pill instead, even though the desired outcome is the same. We think that sucking it up, mastering our fears, is a sign of character. But do we think that people who are naturally fearless lack character? We usually think the opposite. Yet those people are just born lucky. Why should the rest of us have to pay a price in dread, shame, and stomach aches to achieve a state of being that they enjoy for nothing?

Or do we resist the grief pill because we believe that bereavement is doing some work for us? Maybe we think that since we appear to have been naturally selected as creatures that mourn, we shouldn't short-circuit the process. Or is it that we don't want to be the kind of person who does not experience profound sorrow when someone we love dies? Questions like these are the reason we have literature and philosophy. No science will ever answer them.

### 0.0.13 Talking Back to Prozac, Frederick C. Crews (*New York Review of Books*), December 6, 2007

December 6, 2007

Frederick C. Crews (*New York Review of Books*)

The Loss of Sadness: How Psychiatry Transformed Normal Sorrow into Depressive Disorder
by Allan V. Horwitz and Jerome C. Wakefield
Oxford University Press, 287 pp., $29.95

Shyness: How Normal Behavior Became a Sickness
by Christopher Lane
Yale University Press, 263 pp., $27.50

Let Them Eat Prozac: The Unhealthy Relationship Between the Pharmaceutical Industry and Depression
by David Healy
New York University Press, 351 pp., $18.95 (paper)

During the summer of 2002, The Oprah Winfrey Show was graced by a visit from Ricky Williams, the Heisman Trophy holder and running back extraordinaire of the Miami Dolphins. Williams was there to confess that he suffered from painful and chronic shyness. Oprah and her audience were, of course, sympathetic. If Williams, who had been anything but shy on the football field, was in private a wilting violet, how many anonymous citizens would say the same if they could only overcome their inhibition long enough to do so?

To expose one's shyness to what Thoreau once called the broad, flapping American ear would itself count, one might think, as disproof of its actual sway over oneself. But football fans knew that Ricky Williams was no voluble Joe Namath. Nevertheless, there he was before the cameras, evidently risking an anxiety attack for the greater good – namely, the cause of encouraging fellow sufferers from shyness to come out of the closet, seek one another's support, and muster hope that a cure for their disability might soon be found.

Little of what we see on television, however, is quite what it seems. Williams had an incentive – the usual one in our republic, money – for over-

mastering his bashfulness on that occasion. The pharmaceutical corporation GlaxoSmithKline (GSK), through its public relations firm, Cohn & Wolfe, was paying him a still undisclosed sum, not to tout its antidepressant Paxil but simply to declare, to both Oprah and the press, "I've always been a shy person."

To understand why this was considered a worthwhile outlay, we need to know that the drug makers earn their enormous profits from a very few market-leading products for which new applications are continually sought. If those uses don't turn up through experimentation or serendipity, they can be conjured by means of "condition branding" – that is, coaching the masses to believe that one of their usual if stressful states actually partakes of a disorder requiring medication. A closely related term is more poetical: "astroturfing," or the priming of a faux-grassroots movement from which a spontaneous-looking demand for the company's miracle cure will emanate.

In this instance Cohn & Wolfe, whose other clients have included Coca-Cola, Chevron Texaco, and Taco Bell, was using an athlete to help create a belief that shyness, a common trait that some societies associate with good manners and virtue, constitutes a deplorably neglected illness. Given the altruistic aura of the occasion, it would have been tasteless to have Ricky Williams display a vial of Paxil on the spot. But later (before he was suspended from the football league for ingesting quite different drugs), a GSK press release placed his name beneath this boilerplate declaration:

As someone who has suffered from social anxiety disorder, I am so happy that new treatment options, like Paxil CR, are available today to help people with this condition.

There is nothing out of the ordinary in this episode, but that is just why it bears mentioning. Most of us naively regard mental disturbances, like physical ones, as timeless realities that our doctors address according to up-to-date research, employing medicines whose appropriateness and safety have been tested and approved by a benignly vigilant government. Here, however, we catch a glimpse of a different world in which convictions, perceived needs, and choices regarding health care are manufactured along with the products that will match them.

The corporate giants popularly known as Big Pharma spend annually,

worldwide, some $25 billion on marketing, and they employ more Washington lobbyists than there are legislators. Their power, in relation to all of the forces that might oppose their will, is so disproportionately huge that they can dictate how they are to be (lightly) regulated, shape much of the medical research agenda, spin the findings in their favor, conceal incriminating data, co-opt their potential critics, and insidiously colonize both our doctors' minds and our own.

If we hear, for example, that an unprecedented epidemic of depression and anxiety has recently been sweeping the world, we tend not to ask ourselves whose interest is served by that impression. In their painstaking study *The Loss of Sadness*, Allan V. Horwitz and Jerome C. Wakefield cite the World Health Organization's projection: that by 2020 depression will become the second leading cause of worldwide disability, behind only heart disease, and that depression is already the single leading cause of disability for people in midlife and for women of all ages.

The WHO also ranks depression, in its degree of severity for individual victims, ahead of "Down syndrome, deafness, below-the-knee amputation, and angina." But Horwitz and Wakefield cogently argue that those judgments rest on a failure to distinguish properly between major depression, which is indeed devastating for its sufferers, and lesser episodes of sadness. If so, the WHO would appear to have bought Big Pharma's line of goods.

This isn't to say that people who experience infrequent minor depression without long-term dysfunction aren't sick enough to deserve treatment. Of course they are. But as all three of the books under consideration here attest, the pharmaceutical companies haven't so much answered a need as turbocharged it. And because self-reporting is the only means by which nonpsychotic mental ailments come to notice, a wave of induced panic may wildly inflate the epidemiological numbers, which will then drive the funding of public health campaigns to combat the chosen affliction.

This dynamic also applies to a variety of commonplace if bothersome states that the drug makers want us to regard as chemically reparable. They range from excitability and poor concentration to menstrual and menopausal effects and "female sexual dysfunction," whose signature is frustration in bed with the presumably blameless husband or lover. And the same tactic – exag-

gerate the problem but imply that medication will easily fix it – plays upon legitimate worries over cardiovascular disease, osteoporosis, irritable bowel syndrome, and other threats [1]. As patients on a prophylactic regimen, we are grateful for any risk reduction, however minuscule; but our gratitude leaves us disinclined to ask whether the progressively lowered thresholds for intervention were set without any commercial influence. In that sense our prescribed drugs do extra duty as political sedatives.

Clearly, the drug companies' publicists couldn't exercise their consciousness-shaping wiles so fruitfully without a prior disposition among the populace to strive for self-improvement through every legal means. (Neither GlaxoSmithKline nor Cohn & Wolfe invented The Oprah Winfrey Show.) For the past half-century, first with tranquilizers like Miltown and Valium and more recently with the "selective serotonin reuptake inhibitors" (SSRIs), Americans have required little prodding to believe that a medication can neutralize their social handicaps and supply them with a better personality than the one they were dealt by an inconsiderate fate. The vintage and recent advertisements reproduced in Christopher Lane's polemical *Shyness*, which features the manipulations that promoted social anxiety disorder to a national emergency, reflect Madison Avenue's grasp of this yearning to be born again without the nuisance of subscribing to a creed.

Hopes along those lines for Valium and its cousins were soon dashed; the drugs did serve as calmants but at the cost of eventually producing mental fogginess and dependency. In the 1990s, however, the SSRIs Prozac, Zoloft, Paxil, Luvox, Celexa, and Efexor seemed very different, enhancing alertness and making many users feel as if a better self were surfacing. Peter Kramer, without ironic intent, named this phenomenon "cosmetic psychopharmacology," and his best-seller *Listening to Prozac* (1993) swelled a utopian wave that was racing ahead of the drug companies' most optimistic projections.

Even Kramer, though, felt obliged to mention certain troubling effects of Prozac that were already coming to light in the early Nineties. These included, for some users, uncontrollable tremors, diminished sexual capacity, a growing tolerance that was leading to potentially noxious higher doses, and "suicidality," or self-destructive tendencies cropping up in the early weeks of treatment. But because Kramer's readers were weighing the risks not against

a discrete medical benefit but against the prospect of becoming self-assured and gregarious at last, those cautions were generally disregarded.

This point is acknowledged in Kramer's recent book *Against Depression* (2005) – which, however, outdoes even the World Health Organization in its awe before the galloping plague ("The most disabling illness! The costliest!"). Kramer may want to believe the worst about depression's ravages so that the SSRIs he once hailed will still be considered a net boon. Perhaps they are such; I am in no position to judge [2]. But one thing is certain: the antidepressant makers have exploited our gullibility, obfuscated known risks, and treated the victims of their recklessness with contempt. That history needs to be widely known, because the same bullying methods will surely be deployed again as soon as the next family of glamour drugs comes onstream.

Hence the importance of David Healy's stirring firsthand account of the SSRI wars, *Let Them Eat Prozac*. Healy is a distinguished research and practicing psychiatrist, university professor, frequent expert witness, former secretary of the British Association for Psychopharmacology, and author of three books in the field. Instead of shrinking from commercial involvement, he has consulted for, run clinical trials for, and at times even testified for most of the major drug firms. But when he pressed for answers to awkward questions about side effects, he personally felt Big Pharma's power to bring about a closing of ranks against troublemakers. That experience among others has left him well prepared to puncture any illusions about the companies' benevolence or scruples.

Healy doesn't deny that SSRIs can be effective against mood disorders, and he has prescribed them to his own patients. As a psychopharmacologist, however, he saw from the outset that the drug firms were pushing a simplistic "biobabble" myth whereby depression supposedly results straightforwardly from a shortfall of the neurotransmitter serotonin in the brain. No such causation has been established, and the proposal is no more reasonable than claiming that headaches arise from aspirin deprivation [3]. But by insistently urging this idea upon physicians and the public, Big Pharma widened its net for recruiting patients, who could be counted upon to reason as follows: "I feel bad; I must lack serotonin in my brain; these serotonin-boosting pills will surely do the trick" [4]. Thus millions of people who might have needed only

counseling were exposed to incompletely explained risks.

Those risks, Healy perceived, included horrific withdrawal symptoms, such as dizziness, anxiety, nightmares, nausea, and constant agitation, that were frightening some users out of ever terminating their regimen – an especially bitter outcome in view of the manufacturers' promise of enhancing self-sufficiency and peace of mind. The key proclaimed advantage of the new serotonin drugs over the early tranquilizers, freedom from dependency, was simply false. Moreover, the companies had to have known they were gambling wildly with public health. As early as 1984, Healy reports, Eli Lilly had in hand the conclusion pronounced by Germany's ministry of health in denying a license to fluoxetine (later Prozac): "Considering the benefit and the risk, we think this preparation totally unsuitable for the treatment of depression."

As for the frequently rocky initial weeks of treatment, a troubling record not just of "suicidality" but of actual suicides and homicides was accumulating in the early 1990s. The drug firms, Healy saw, were distancing themselves from such tragedies by blaming depression itself for major side effects. Handouts for doctors and patients urged them to persist in the face of early emotional turmoil that only proved, they were told, how vigorously the medicine was tackling the ailment. So, too, dependency symptoms during termination were said to be evidence that the long-stifled depression was now reemerging.

The most gripping portions of *Let Them Eat Prozac* narrate courtroom battles in which Big Pharma's lawyers, parrying negligence suits by the bereaved, took this line of double-talk to its limit by explaining SSRI-induced stabbings, shootings, and self-hangings by formerly peaceable individuals as manifestations of not-yet-subdued depression. As an expert witness for plaintiffs against SSRI makers in cases involving violent behavior, Healy emphasized that depressives don't commit mayhem. But he also saw that his position would be strengthened if he could cite the results of a drug experiment on undepressed, certifiably normal volunteers. If some of them, too, showed grave disturbance after taking Pfizer's Zoloft – and they did in Healy's test, with long-term consequences that have left him remorseful as well as indignant – then depression was definitively ruled out as the culprit.

Healy suspected that SSRI makers had squirreled away their own awkward findings about drug-provoked derangement in healthy subjects, and he found

such evidence after gaining access to Pfizer's clinical trial data on Zoloft. In 2001, however, just when he had begun alerting academic audiences to his forthcoming inquiry, he was abruptly denied a professorship he had already accepted in a distinguished University of Toronto research institute supported by grants from Pfizer. The company hadn't directly intervened; the academics themselves had decided that there was no place on the team for a Zoloft skeptic.

Undeterred, Healy kept exposing the drug attorneys' leading sophistry, which was that a causal link to destructive behavior could be established only through extensive double-blind randomized trials – which, cynically, the firms had no intention of conducting. In any case, such experiments could have found at best a correlation, in a large anonymous group of subjects, between SSRI use and irrational acts; and the meaning of a correlation can be endlessly debated. In contrast, Healy's own study had already isolated Zoloft as the direct source of his undepressed subjects' ominous obsessions.

Thanks partly to Healy's efforts, juries in negligence suits gradually learned to be suspicious of the "randomized trial" shell game. The plaintiffs' lawyers in some of these cases cited his research. But this David doesn't suppose that he has felled Goliath. As he explains, a decisive improvement in the legal climate surrounding SSRIs came only after Eli Lilly bought the marketing rights to a near relative of its own patent-lapsed Prozac. According to the new drug's damning patent application, it was less likely than Prozac to induce "headaches, nervousness, anxiety, insomnia, inner restlessness ..., *suicidal thoughts and self mutilation*" (emphasis added). That disclosure by Prozac's own progenitor neatly illustrates Healy's belief that the full truth about any drug will emerge only when the income it produces has fallen and its defects can be advantageously contrasted with the virtues of a successor product.

Meanwhile, Healy wonders, who will now be sufficiently strong and uncorrupted to keep the drug makers honest? The FDA, he notes, is timid, underfunded, and infiltrated by friends of industry; even the most respected medical journals hesitate to offend their pharmaceutical advertisers; professional conferences are little more than trade fairs; leading professors accept huge sums in return for serving the companies in various venal ways; and,

most disgracefully of all, many of their "research" papers are now ghostwritten outright by company-hired hacks. As Healy puts it, Big Pharma doesn't just bend the rules; it buys the rulebook.

There is, however, one rulebook that does place some constraint on what the drug makers can claim. This is the American Psychiatric Association's *Diagnostic and Statistical Manual of Mental Disorders* (DSM). Its four editions (plus interim revisions) thus far from 1952 through 1994 specify the psychological ailments that the whole mental health system has agreed to deem authentic. Although "condition branding" by publicists can make normal people feel like sufferers from a given malady, the malady itself must first be listed in the DSM in order for medical treatments to be approved.

Can we rely on this guidebook, then, for independent, objective judgment about the identification and treatment of mental complaints? The compilers of each edition have boasted that their named disorders rest mainly on research findings, and most physicians take that claim at face value, as do medical insurers, epidemiologists, and the funders of empirical studies. An acquaintance with the DSM's several versions and with the controversies that shaped them, however, suggests that they tell more about the shifting zeitgeist and the factions within the APA than they do about permanently valid syndromes [5].

Human nature has not metamorphosed several times since 1952, but each DSM has included more disorders than the last. The third edition of 1980 alone, liberally subdividing earlier categories, purported to have unearthed 112 more of them than the second edition of 1968, and by the fourth edition of 1994 there were over 350, marked by such dubiously pathognomonic symptoms as feeling low, worrying, bearing grudges, and smoking. Those stigmata, furthermore, are presented in a user-friendly checklist form that awards equal value to each symptom within a disorder's entry. In Bingo style, for example, a patient who fits five out of the nine listed criteria for depression is tagged with the disorder. It is little wonder, then, that drug makers' advertisements now urge consumers to spot their own defectiveness through reprinted DSM checklists and then to demand correction via the designated pills.

It would be a bad mistake, however, to assume that the shapers of the DSM have been deliberately tilting the manual away from humane psychotherapy

and toward biological and pharmaceutical reductionism of the sort exemplified by the serotonin-deficit theory of depression. That very assumption vitiates Christopher Lane's conspiracy-minded book *Shyness*, which begins plausibly enough as an exposé of the campaign to have shy people view themselves as mentally ill. Unfortunately, Lane couples that theme with a histrionic dismissal of the DSM that is too uncomprehending and partisan to be taken seriously.

Lane is not a psychiatrist but a psychoanalytic literary critic who aligns himself with such empirically insouciant authorities as Jacques Lacan, Elisabeth Roudinesco, and Adam Phillips. Like many another Freudian, he is still in shock over DSM-III of 1980 – the edition that consigned the "neuroses" to limbo, favored descriptive over depth-psychological accounts of disorders, and established the uniform symptom-tallying procedure for certifying a diagnosis.

For Lane, the very attempt to clarify disorders according to their detectable traits constituted a spiteful purging of "almost a century of [psychoanalytic] thought" and thus a reversion to "Victorian psychiatry." He assumes that anyone who hesitates to endorse etiologies based on quarrels between the homuncular ego and superego must be hostile to all mental complexity and hence to psychotherapy in general. That is his charge against DSM-III and DSM-IV. In fact, however, the manual has never stated or implied a preference between talk therapy and pills. If it had done so, it could hardly have served as the consensual guidebook for such a heterogeneous organization as the APA.

In *The Loss of Sadness* Horwitz and Wakefield discuss the same 1980 change of direction by the DSM that leaves Christopher Lane fuming. As these authors show, the APA leadership's intentions in the late 1970s had nothing to do with pushing drugs and everything to do with lending greater scientific respectability to the psychiatric field. What was wanted thenceforth for the DSM was improved validity and reliability, by virtue of identifying disorders more accurately and providing means of detection that would render several diagnoses of the same patient less divergent.

This remains the DSM's formal goal, however elusive, and it is plainly appropriate and irreversible. What we should really be asking is whether

the DSM has approached that goal or merely gestured toward it through the false concreteness of checklists, implying that newly minted disorders are as sharply recognizable as diabetes and tuberculosis. As Horwitz and Wakefield put it, "the reliability might just represent everybody together getting the same wrong answer."

Horwitz and Wakefield's argumentation is as understated as Lane's is melodramatic. Because these collaborators maintain a constructive, scholarly tone and display a total command of the pertinent literature, they will gain a respectful hearing from psychiatrists. Those readers will discover, however, that *The Loss of Sadness* amounts to a relentless dismantling of the DSM – one that seems confined at first to a single inadequacy, only to blossom into an exposure of the manual's top-to-bottom arbitrariness. I am not sure, in fact, that the authors themselves understand the full gravity of the challenge they have posed for American psychiatry.

At the core of their book lies a demonstration that episodic sadness has always been a socially approved means of adjusting to misfortune and that much is lost, both medically and culturally, when it is misread as a depressive disorder. Yet as Robert L. Spitzer, the chief architect of DSM-III and Christopher Lane's *bête noire*, concedes in a generous foreword, the manual has propagated that very blunder by failing to clarify the difference between environmentally prompted moods – those responding to stress or hardship – and dysfunctional states persisting long after the causes of stress have abated. In no sense, however, can that indictment be confined to just one disorder. *The Loss of Sadness* implies that nearly every nonpsychotic complaint is subject to overdiagnosis unless contextual factors – familial, cultural, relational, financial – are weighed in the balance.

As might be expected, then, Horwitz and Wakefield end by begging the compilers of DSM-V (now projected for 2012) to teach their colleagues the need for inquiry into each patient's circumstances before concluding that they are faced with a bona fide disorder. The bar for authentic pathology must be set higher. If this is done, the authors declare, the DSM will be more scientifically respectable; its users, instead of regarding disadvantaged classes as infested with mental illness, will gain an appreciation of socioeconomic reasons for unhappiness; and a brake will be placed on the expensive middle-

class hypochondria that the drug companies have so assiduously encouraged and exploited.

All of which would be wonderful, but the scenario is shadowed by Horwitz and Wakefield's own shrewd analysis of the status quo and its beneficiaries. The DSM's laxity about winnowing vague discontents from real maladies is, in financial terms, highly functional for many of its practitioners and their patients. As the product of a guild whose members seek payment for treating whatever complaints are brought to them, the manual must be biased toward overmedicalizing so that both doctors and patients can be served under managed care. As Horwitz and Wakefield themselves observe:

The DSM provides flawed criteria ...; the clinician, who cannot be faulted for applying officially sanctioned DSM diagnostic criteria, knowingly or unknowingly misclassifies some normal individuals as disordered; and these two errors lead to the patient receiving desired treatment for which the therapist is reimbursed.

What motive would the APA, as a practitioners' union, have for bringing that arrangement to an end? And wouldn't the drug makers, whose power to shape psychiatric opinion should never be discounted, add their weight on the side of continued diagnostic liberality?

Horwitz and Wakefield's admirable concern for scientific rationality points us toward some uncomfortable insights about American psychiatry and its role within a far from rational health care system. That system is too cumbersome and too driven by profit considerations to meet the whole society's medical needs; but citizens possessing full insurance, when they feel mentally troubled in any way, won't be denied medication or therapy or both. Nothing more is required than some hypocrisy all around. As for psychiatry's inability to settle on a discrete list of disorders that can remain impervious to fads and fashions, that is an embarrassment only to clear academic thinkers like these two authors. For bureaucratized psychological treatment, and for the pharmaceutical industry that is now deeply enmeshed in it, confusion has its uses and is likely to persist.

Notes

[1] These matters are discussed in Ray Moynihan and Alan Cassels's *Selling Sickness: How the World's Biggest Pharmaceutical Companies Are Turn-*

*ing Us All into Patients* (Nation Books, 2005), which also cites the Ricky Williams story.

[2] For recently unearthed considerations bearing on the prudent use of these drugs, see Robert D. Gibbons et al., "Early Evidence on the Effects of Regulators' Suicidality Warnings on SSRI Prescriptions and Suicide in Children and Adults," *American Journal of Psychiatry*, Vol. 164, No. 9 (September 2007), pp. 1356–1363, and Gonzalo Laje et al., "Genetic Markers of Suicidal Ideation Emerging During Citalopram Treatment of Major Depression," *American Journal of Psychiatry*, Vol. 164, No. 10 (October 2007), pp. 1530–1538.

[3] The serotonin etiology of depression is concisely disputed in Horwitz and Wakefield's *The Loss of Sadness*, pp. 168–170. See also Elliot S. Valenstein, *Blaming the Brain: The Truth About Drugs and Mental Health* (Free Press, 1998), and Joseph Glenmullen, *Prozac Backlash: Overcoming the Dangers of Prozac, Zoloft, Paxil, and Other Anti-depressants with Safe, Effective Alternatives* (Simon and Schuster, 2000).

[4] Healy cites Tipper Gore in 1999: "It was definitely a clinical depression. ... What I learned about it is your brain needs a certain amount of serotonin and when you run out of that, it's like running out of gas." Contrary to industry propaganda, the brain possesses no known "depression center," and about 95 percent of our serotonin is found elsewhere in the body. By raising serotonin levels, the SSRIs interfere with production of the other natural "feel-good" chemicals, adrenaline and dopamine. In that sense they are hardly as "selective" as we have been led to believe.

[5] See two important collaborative critiques by Herb Kutchins and Stuart A. Kirk: *The Selling of DSM: The Rhetoric of Science in Psychiatry* (Aldine de Gruyter, 1992) and *Making Us Crazy: DSM: The Psychiatric Bible and the Creation of Mental Disorders* (Free Press, 1997).

**The Truth About Prozac: An Exchange**

February 14, 2008

By Gerald Curzon, Evan Hughes, Peter D. Kramer, Reply by Frederick C. Crews

In response to "Talking Back to Prozac" (December 6, 2007)

To the Editors:

Regarding medication's effects on the self, Frederick Crews writes [*NYR*, December 6, 2007], "Peter Kramer, without ironic intent, named this phenomenon 'cosmetic psychopharmacology.'" Without ironic intent? Here is how I introduced the concept in *Listening to Prozac*. I had been writing about a troubled patient I called Tess, and now I asked readers to imagine someone less depressed who might nonetheless gain social competence by taking medicine. Using psychotherapy for the same end might be acceptable, I wrote,

But I wondered whether we were ready for "cosmetic psychopharmacology." It was my musings about whether it would be kosher to medicate a patient like Tess in the absence of depression that led me to coin the phrase. Some people might prefer pharmacologic to psychologic self-actualization. Psychic steroids for mental gymnastics, medicinal attacks on the humors, anti-wallflower compound – these might be hard to resist. Since you only live once, why not do it as a blonde? Why not as a peppy blonde?

I think fair-minded readers will agree that I was trying at once to capture a phenomenon and take distance from it – that is, that I was signaling ironic intent. This attitude is fully apparent in context; the paragraph continues with concerns about tweaking personality in the guise of treating depression:

Already, it seems to me, psychiatric diagnosis had been subject to a sort of "diagnostic bracket creep" – the expansion of categories to match the scope of relevant medications.

In the same vein, Crews writes:

Even Kramer, though, felt obliged to mention certain troubling effects of Prozac that were already coming to light in the early Nineties. These included, for some users, uncontrollable tremors, diminished sexual capacity, a growing tolerance that was leading to potentially noxious higher doses, and "suicidality," or self-destructive tendencies cropping up in the early weeks of treatment.

Again: Even Kramer? Listing those side effects in 1993, I was relying on scattered case reports, along with my own and trusted colleagues' clinical experience, and so was out ahead of the standard literature about the new

antidepressants – hardly the posture of a reluctant critic. I noted these negative outcomes despite a warning posted in the book's introduction. There, I told readers that since I would focus on theoretical issues related to the ethics of medication and personality change, I would give extensive attention neither to drug side effects nor to positive effects in the treatment of major mental illness.

Many of the concerns Crews shares, regarding the impulse to use drugs to enhance ordinary mood, were broached in *Listening to Prozac*. Later, I noticed that people were taking my worries about cosmetic psychopharmacology and applying them – in my view, mistakenly – to conventional uses of medication; it was this observation that led me to write *Against Depression*.

Peter D. Kramer
Clinical Professor of Psychiatry and Human Behavior
Brown University Medical School
Providence, Rhode Island

---

To the Editors:

Frederick Crews argues not only that depression severe enough to justify drug treatment is much rarer than the drug industry would have us believe but also that the serotonin deprivation hypothesis of depression has little to recommend it. The drug industry disagrees and would, no doubt, be delighted if everyone who had ever felt sad took antidepressant selective serotonin reuptake inhibitors (SSRIs) for the rest of their lives.

The truth lies between these positions. Enormous numbers of unrecognized severe depressives never receive psychiatric attention. These are mostly males, Thoreau's "mass of men [leading] lives of quiet desperation." As for the serotonin deprivation hypothesis: Crews's opinion that this is like explaining headaches by aspirin deprivation lacks substance. Aspirin is only detectable in people who have taken aspirin. Serotonin is a normal constituent of every human brain where it is released from certain neuronal tracts to various sites of action. Evidence accumulating over half a century indicates that serotonin has a causal role in depressive illness and that the antidepressant properties of SSRIs are mediated by increasing its availability to specific brain sites (see reviews in *Interactive Monoaminergic Disorders*, edited by Paloma et

al., 1999, by Cowen, Curzon, and Van Praag).

This does not imply that the brain is a kind of slot machine delivering happiness if serotonin is inserted, or as one British newspaper proclaimed when our royal family was undergoing multiple emotional crises, "Fergie and Di. It was the serotonin. The difference between happiness and misery." Neither does it imply that all depressives respond to SSRIs, nor that SSRIs only act on sites mediating their therapeutic effect. Therapeutic benefit only occurs after the first few weeks of treatment during which initial period other sites become less responsive. During this initial period, some subjects experience harmful effects of the drugs which can be serious in a small fraction of patients.

Crews's article raises important issues, e.g., the power of the drug companies, especially of their sales departments, and the temptations inherent in the high cost of drug development combined with the tremendous profits that can be gained from a few of the small fraction of drugs under development that reach the market. However, he is over-polemical and his cherry-picking for negative findings from a large and often confusing literature may produce more heat than light.

Gerald Curzon
Emeritus Professor of Neurochemistry
University of London

---

To the Editors:

Frederick Crews makes much of the notion that suicidality is a demonstrated "troubling effect" of antidepressants. He speaks of "SSRI-induced stabbings, shootings, and self-hangings" and claims that makers of these medications have "treated the victims of their recklessness with contempt." But to speak of "effects," "induced" behaviors, and "victims" here is to greatly overstep. There is some correlation between SSRIs and suicidality in the early weeks of treatment, but cause is different from correlation, as Crews takes pains to point out elsewhere in the review. To borrow his own analogy for conflating the two, saying that some suicides are caused by SSRIs is like saying that "headaches arise from aspirin deprivation."

Pfizer's failure to own up promptly to the correlation between Prozac and

the incidence of suicide attempts is typical and lamentable behavior for a pharmaceutical company; that hardly needs saying. But to read Crews, one would think Big Pharma suppressed some magic data that proved their drugs alone can cause suicide. Psychiatric experts have attributed suicide attempts among SSRI patients to a combination of a preexisting low self-regard with the energy boost provided by the medication. I'm not qualified to definitely assess that hypothesis (perhaps no one is), but suicide seems far more likely to be in part a reflection of a natural mood rather than a straightforward product of medication alone, as talk of "SSRI-induced ... self-hangings" strongly implies. We should remember that any patients who killed themselves on SSRIs were taking them because they were already deemed depressed.

Crews also writes, in laudably criticizing marketing practices, that "the pharmaceutical companies haven't so much answered a need" – really? – "as turbocharged it." I should think the importance of their doing the first outweighs by far the significance of the second.

Crews subscribes to the view that depression is overdiagnosed and antidepressants overprescribed, and that psychiatrists too often write prescriptions without an adequate familiarity with the life of the patient. The available evidence suggests that these beliefs are true, as is the notion that SSRIs have been tied to suicide. But all those claims are compatible with the widely recognized fact that antidepressants have helped a great many people and that, as John Michael Bostwick of the Mayo Clinic has written,

several recent, large nonindustry studies indicated that rates of suicide and suicidal behavior were actually reduced in children who used antidepressants, despite piteous anecdotal tales.

If Pfizer ought to have admitted to correlations inconvenient to them, a critic of Pfizer, already picking an easy target, ought to admit to these highly commendable correlations too, inconvenient as they might be to him.

Evan Hughes
New York City

---

Frederick Crews replies:

I stand corrected on the matter of Peter Kramer's tone when introducing the term "cosmetic psychopharmacology." As a social meditation, *Listening*

*to Prozac* meant to identify and analyze a trend, not to endorse it. Unfortunately, hundreds of thousands of readers overlooked Kramer's misgivings and rushed out to acquire Prozac so that they, too, like the small sample of patients described in the book, could become "better than well." Psychologically healthy people thus exposed themselves to serious risks for a benefit that now appears to have been overrated.

Gerald Curzon's point about an underdiagnosed and undertreated class of depressives is well taken. It doesn't weigh against the historical fact that depression itself, as a disease entity, was vastly promoted by the companies that saw a market in it for their pills.

As a layman, I make no claim to familiarity with brain chemistry or medicine. I can only read what experts have written and compare it to the simplifications circulated by hucksters. Three concurring books that I cited emboldened me to dispute one such simplification: that depression is well known to be caused simply by a shortfall of serotonin in the brain. Dr. Curzon's demurral from the journalistic "Fergie and Di" assertion would appear to put us in agreement on that point – but possibly not.

The physiological and medical facts that Dr. Curzon mentions, including the presence of serotonin in everyone's brain and the therapeutic effect that extra serotonin can produce, are not in dispute. They certainly offer important clues for research. But they do not tell us whether a serotonin deficit causes depression or, rather, is itself a byproduct of other influences. The fact that a given agent, even a naturally occurring one, relieves a disorder doesn't prove how that disorder came about. If Dr. Curzon believes otherwise, that is where we disagree.

It is entirely possible, as Evan Hughes suggests, that most suicide attempts and other destructive acts by a small minority of SSRI takers have resulted from an "energy boost" provided by the drug in combination with preexisting tendencies. Since, in such cases, the drug isn't the only causative factor, Hughes declares that it hasn't induced the behavior. This is, to put it mildly, a strange approach to suicide prevention. If all that's required to send a depressed patient over the edge is a chemically supplied energy boost, shouldn't a physician be cautious about prescribing it?

Hughes elsewhere shifts his ground and reverts to the drug companies'

early, now abandoned, claim that nothing more than a loose correlation, devoid of causal significance, links SSRIs to suicidality; depression itself, he says, is responsible for bad outcomes. As I related, that argument began to fail when word leaked out that, in clinical trials that Pfizer withheld from the FDA and again in a test by David Healy, undepressed patients experienced the same symptoms that had been blamed on depression. Moreover, Hughes ignores Eli Lilly's eventual statement, in a later patent application, that its own Prozac can bring on "suicidal thoughts and self-mutilation."

My article didn't deny, as Hughes insinuates, that "antidepressants have helped a great many people." Nor did it claim that SSRIs have induced more suicides among young people than they have prevented. On the contrary, I cited a recent article propounding the opposite conclusion – the same finding now advanced against me by Hughes. And I cited still another article indicating that genetic tests may someday render SSRIs safer by showing which patients would poorly tolerate a given drug. Neither of those references could have appeared in an article denying any utility to SSRIs.

Detailed evidence that SSRI makers have "treated the victims of their recklessness with contempt" can be found in Healy's *Let Them Eat Prozac*, which Hughes shows no sign of having read.

**Prozac and Sexual Desire**

March 20, 2008

By Helen E. Fisher, J. Anderson Thomson Jr.
In response to "Talking Back to Prozac" (December 6, 2007)

To the Editors:

We applaud Frederick Crews's discussion of the unappreciated problems of antidepressants and the subtle techniques used to enhance their use and sales [*NYR*, December 6, 2007]. He mentions but does not elaborate on the sexual side effects produced by serotonin-enhancing antidepressants, most commonly the SSRIs such as Prozac, Zoloft, Paxil, Celexa, Luvox, and Lexapro. Today tens of millions of people take these serotonin-enhancing antidepressants. And because Prozac, Zoloft, Paxil, and Celexa are now available in generic forms, their use will only increase.

It is well known that these medications can cause emotional blunting and dysfunction in sexual desire, arousal, and performance in upward of three of every four patients. But we are writing now to add that we believe these side effects have even more serious consequences than currently appreciated, due to their impact on several other related neural mechanisms [*].

*Homo sapiens* evolved three distinct (yet overlapping) brain systems for courtship, mating, reproduction, and parenting. The sex drive evolved to motivate men and women to initiate sexual contact with a range of partners; romantic attraction evolved to motivate individuals to focus their courtship energy on specific partners, thereby conserving mating time and metabolic energy; and the neural system associated with partner attachment evolved to motivate our forebears to maintain a stable mateship long enough to complete parenting duties.

Studies using functional Magnetic Resonance Imaging (fMRI) indicate that romantic attraction is associated with subcortical dopaminergic pathways – pathways that are suppressed by elevated central serotonin. Hence serotonin-enhancing antidepressants can jeopardize one's ability to fall in love.

Due to their impact on the sex drive, these medications can also jeopardize other brain/body mechanisms that govern mate assessment, mate choice, pair formation, and partner attachment. For example, female orgasm has many functions. Among them, it aids sperm retention and enables women to discriminate between self-centered as opposed to dedicated partners – partners who might be more likely to help them rear their young. Female orgasm may also help women choose men with better genes, as women are more orgasmic with men who are healthy and symmetrical, markers of good testosterone load. Female orgasm may also enhance feelings of attachment, because it stimulates the release of oxytocin and prolactin. As these drugs impair or eliminate female orgasm, they interfere with delicate biological mechanisms designed to aid mate choice and partner attachment. As these SSRI medications impair male orgasm, they also jeopardize a male's ability to court, inseminate, and attach to a potential partner.

In short, the sex drive operates in conjunction with many other neural systems that govern desire, mate choice, romantic love, and attachment, perhaps

even mechanisms that detect facial attractiveness, immune system compatibility, and other neural systems we unconsciously use to mate, breed, and rear our young. Yet there is, as yet, no research on the complex effects of these serotonin-enhancing drugs – medications that are likely to jeopardize mate choice, romantic love, marriage, and even fertility.

Helen E. Fisher, Ph.D.
Research Professor
Department of Anthropology
Rutgers University
New York City

J. Anderson Thomson Jr., M.D.
Charlottesville, Virginia

Notes

[*] H.E. Fisher and J.A. Thomson, "Lust, Romance, Attachment: Do the Sexual Side Effects of Serotonin-Enhancing Antidepressants Jeopardize Romantic Love, Marriage, and Fertility?," in *Evolutionary Cognitive Neuroscience*, edited by S. Platek, J. Keenan, and T. Shackelford (MIT Press, 2006).

**Branded By Pharma**

May 1, 2008

By Andrew Branch

In response to "Talking Back to Prozac" (December 6, 2007)

To the Editors:

In his description of the relationship between pharmaceutical companies and the mental health industry ["Talking Back to Prozac," *NYR*, December 6, 2007], Frederick C. Crews characterizes condition branding as an effort by pharmaceutical companies to persuade "the masses ... that ... one of their usual ... states actually partakes of a disorder requiring medication," implying that condition branding is an effort by "Big Pharma" to convince potential patients of their own illness. In doing so, he neglects to address condition branding's secondary message: not only are its audience's own personal

dissatisfactions medically treatable, but their relatives', friends', and neighbors' behaviors are also controllable by medication. While "self-reporting is the only means by which nonpsychotic mental ailments come to notice," the decision to self-report isn't a self-contained process within the patient, and Big Pharma markets accordingly.

Minor children, and preschoolers specifically, have been identified as the largest growing market for antidepressants ("Trends in the Use of Antidepressants in a National Sample of Commercially Insured Pediatric Patients, 1998 to 2002," *Psychiatric Services*, April 2004). The greatest recent growth in psychopharmaceutical consumption, *ipso facto* the greatest growth in psychopharmaceutical consumption in the era of condition branding, is occurring in a population that can hardly be described as self-diagnosing. Clearly the infamous "nag factor," where advertisers market to children in the hope that they will then persuade their parents to purchase, doesn't apply to Prozac; rather, the message of condition branding is reaching its intended targets (adults) and inspiring them to see newly diagnosable, newly treatable symptoms not in themselves but in those around them.

The tendency to identify psychological illness in others is not limited to minors and it is not an unintended side effect of condition branding aimed directly at the end user. The Roche Pharmaceuticals Web site for its Accu-Chek diabetes product contains a section for "caregivers," counseling them that if their diabetic loved one is "feeling sad or irritable" or experiencing any other of a laundry list of general symptoms, they "may have depression." Using the same techniques described in Crews's article, the Web site brands depression not as something its audience should be concerned about in themselves, but as something they should look for in others, with or without their agreement. "Your spouse may feel it's hopeless to reach out for help. Your spouse is wrong" (www.accu-chek.ca), the antidepressant manufacturer enjoins the public, encouraging readers to pursue treatment for their loved ones even if faced with spouses who are less than enthusiastic about the diagnosis.

An understanding of condition branding's secondary aspect does not change the interpretation of Big Pharma's goals in its employment – through direct or once removed means, its intention is to increase the number of consumers of its products and thereby increase its profit. If "episodic sadness" has, as

Crews relates, truly changed from "socially approved means of adjusting to misfortune" to something society defines as a disease, Big Pharma has not effected the change through force or decree. Rather it has persuaded society to shift its views, and it has done so by exploiting a preexistent human desire.

Certainly, one desire it underlines is the desire to believe our flaws are not our fault and that we can fight them through self-diagnosis. However, as much as Ricky Williams's appearance on The Oprah Winfrey Show promotes the conclusion that if a star football player suffers from "shyness," so could any of us, for a viewer seeing a difference between him/herself and a Heisman winner the message might as easily be: if he suffers from "shyness," so could any of them.

Andrew Branch
Portland, Oregon

### 0.0.14 Do We Really Know What Makes Us Healthy?, Gary Taubes (*New York Times*), September 16, 2007

September 16, 2007

Gary Taubes (*New York Times*)

Once upon a time, women took estrogen only to relieve the hot flashes, sweating, vaginal dryness and the other discomforting symptoms of menopause. In the late 1960s, thanks in part to the efforts of Robert Wilson, a Brooklyn gynecologist, and his 1966 best seller, "Feminine Forever," this began to change, and estrogen therapy evolved into a long-term remedy for the chronic ills of aging. Menopause, Wilson argued, was not a natural age-related condition; it was an illness, akin to diabetes or kidney failure, and one that could be treated by taking estrogen to replace the hormones that a woman's ovaries secreted in ever diminishing amounts. With this argument estrogen evolved into hormone-replacement therapy, or H.R.T., as it came to be called, and became one of the most popular prescription drug treatments in America.

By the mid-1990s, the American Heart Association, the American College of Physicians and the American College of Obstetricians and Gynecologists had all concluded that the beneficial effects of H.R.T. were sufficiently well established that it could be recommended to older women as a means of warding off heart disease and osteoporosis. By 2001, 15 million women were filling H.R.T. prescriptions annually; perhaps 5 million were older women, taking the drug solely with the expectation that it would allow them to lead a longer and healthier life. A year later, the tide would turn. In the summer of 2002, estrogen therapy was exposed as a hazard to health rather than a benefit, and its story became what Jerry Avorn, a Harvard epidemiologist, has called the "estrogen debacle" and a "case study waiting to be written" on the elusive search for truth in medicine.

Many explanations have been offered to make sense of the here-today-gone-tomorrow nature of medical wisdom – what we are advised with confidence one year is reversed the next – but the simplest one is that it is the natural rhythm of science. An observation leads to a hypothesis. The hypothesis (last year's advice) is tested, and it fails this year's test, which is always the most likely outcome in any scientific endeavor. There are, after all, an

infinite number of wrong hypotheses for every right one, and so the odds are always against any particular hypothesis being true, no matter how obvious or vitally important it might seem.

In the case of H.R.T., as with most issues of diet, lifestyle and disease, the hypotheses begin their transformation into public-health recommendations only after they've received the requisite support from a field of research known as epidemiology. This science evolved over the last 250 years to make sense of epidemics – hence the name – and infectious diseases. Since the 1950s, it has been used to identify, or at least to try to identify, the causes of the common chronic diseases that befall us, particularly heart disease and cancer. In the process, the perception of what epidemiologic research can legitimately accomplish – by the public, the press and perhaps by many epidemiologists themselves – may have run far ahead of the reality. The case of hormone-replacement therapy for post-menopausal women is just one of the cautionary tales in the annals of epidemiology. It's a particularly glaring example of the difficulties of trying to establish reliable knowledge in any scientific field with research tools that themselves may be unreliable.

What was considered true about estrogen therapy in the 1960s and is still the case today is that it is an effective treatment for menopausal symptoms. Take H.R.T. for a few menopausal years and it's extremely unlikely that any harm will come from it. The uncertainty involves the lifelong risks and benefits should a woman choose to continue taking H.R.T. long past menopause. In 1985, the Nurses' Health Study run out of the Harvard Medical School and the Harvard School of Public Health reported that women taking estrogen had only a third as many heart attacks as women who had never taken the drug. This appeared to confirm the belief that women were protected from heart attacks until they passed through menopause and that it was estrogen that bestowed that protection, and this became the basis of the therapeutic wisdom for the next 17 years.

Faith in the protective powers of estrogen began to erode in 1998, when a clinical trial called HERS, for Heart and Estrogen-progestin Replacement Study, concluded that estrogen therapy increased, rather than decreased, the likelihood that women who already had heart disease would suffer a heart attack. It evaporated entirely in July 2002, when a second trial, the Women's

Health Initiative, or W.H.I., concluded that H.R.T. constituted a potential health risk for all postmenopausal women. While it might protect them against osteoporosis and perhaps colorectal cancer, these benefits would be outweighed by increased risks of heart disease, stroke, blood clots, breast cancer and perhaps even dementia. And that was the final word. Or at least it was until the June 21 issue of *New England Journal of Medicine*. Now the idea is that hormone-replacement therapy may indeed protect women against heart disease if they begin taking it during menopause, but it is still decidedly deleterious for those women who begin later in life.

This latest variation does come with a caveat, however, which could have been made at any point in this history. While it is easy to find authority figures in medicine and public health who will argue that today's version of H.R.T. wisdom is assuredly the correct one, it's equally easy to find authorities who will say that surely we don't know. The one thing on which they will all agree is that the kind of experimental trial necessary to determine the truth would be excessively expensive and time-consuming and so will almost assuredly never happen. Meanwhile, the question of how many women may have died prematurely or suffered strokes or breast cancer because they were taking a pill that their physicians had prescribed to protect them against heart disease lingers unanswered. A reasonable estimate would be tens of thousands.

The Flip-Flop Rhythm of Science:

At the center of the H.R.T. story is the science of epidemiology itself and, in particular, a kind of study known as a prospective or cohort study, of which the Nurses' Health Study is among the most renowned. In these studies, the investigators monitor disease rates and lifestyle factors (diet, physical activity, prescription drug use, exposure to pollutants, etc.) in or between large populations (the 122,000 nurses of the Nurses' study, for example). They then try to infer conclusions – i.e., hypotheses – about what caused the disease variations observed. Because these studies can generate an enormous number of speculations about the causes or prevention of chronic diseases, they provide the fodder for much of the health news that appears in the media – from the potential benefits of fish oil, fruits and vegetables to the supposed dangers of sedentary lives, trans fats and electromagnetic fields. Because

these studies often provide the only available evidence outside the laboratory on critical issues of our well-being, they have come to play a significant role in generating public-health recommendations as well.

The dangerous game being played here, as David Sackett, a retired Oxford University epidemiologist, has observed, is in the presumption of preventive medicine. The goal of the endeavor is to tell those of us who are otherwise in fine health how to remain healthy longer. But this advice comes with the expectation that any prescription given – whether diet or drug or a change in lifestyle – will indeed prevent disease rather than be the agent of our disability or untimely death. With that presumption, how unambiguous does the evidence have to be before any advice is offered?

The catch with observational studies like the Nurses' Health Study, no matter how well designed and how many tens of thousands of subjects they might include, is that they have a fundamental limitation. They can distinguish associations between two events – that women who take H.R.T. have less heart disease, for instance, than women who don't. But they cannot inherently determine causation – the conclusion that one event causes the other; that H.R.T. protects against heart disease. As a result, observational studies can only provide what researchers call hypothesis-generating evidence – what a defense attorney would call circumstantial evidence.

Testing these hypotheses in any definitive way requires a randomized-controlled trial – an experiment, not an observational study – and these clinical trials typically provide the flop to the flip-flop rhythm of medical wisdom. Until August 1998, the faith that H.R.T. prevented heart disease was based primarily on observational evidence, from the Nurses' Health Study most prominently. Since then, the conventional wisdom has been based on clinical trials – first HERS, which tested H.R.T. against a placebo in 2,700 women with heart disease, and then the Women's Health Initiative, which tested the therapy against a placebo in 16,500 healthy women. When the Women's Health Initiative concluded in 2002 that H.R.T. caused far more harm than good, the lesson to be learned, wrote Sackett in *The Canadian Medical Association Journal*, was about the "disastrous inadequacy of lesser evidence" for shaping medical and public-health policy. The contentious wisdom circa mid-2007 – that estrogen benefits women who begin taking

it around the time of menopause but not women who begin substantially later – is an attempt to reconcile the discordance between the observational studies and the experimental ones. And it may be right. It may not. The only way to tell for sure would be to do yet another randomized trial, one that now focused exclusively on women given H.R.T. when they begin their menopause.

A Poor Track Record of Prevention:

No one questions the value of these epidemiologic studies when they're used to identify the unexpected side effects of prescription drugs or to study the progression of diseases or their distribution between and within populations. One reason researchers believe that heart disease and many cancers can be prevented is because of observational evidence that the incidence of these diseases differ greatly in different populations and in the same populations over time. Breast cancer is not the scourge among Japanese women that it is among American women, but it takes only two generations in the United States before Japanese-Americans have the same breast cancer rates as any other ethnic group. This tells us that something about the American lifestyle or diet is a cause of breast cancer. Over the last 20 years, some two dozen large studies, the Nurses' Health Study included, have so far failed to identify what that factor is. They may be inherently incapable of doing so. Nonetheless, we know that such a carcinogenic factor of diet or lifestyle exists, waiting to be identified.

These studies have also been invaluable for identifying predictors of disease – risk factors – and this information can then guide physicians in weighing the risks and benefits of putting a particular patient on a particular drug. The studies have repeatedly confirmed that high blood pressure is associated with an increased risk of heart disease and that obesity is associated with an increased risk of most of our common chronic diseases, but they have not told us what it is that raises blood pressure or causes obesity. Indeed, if you ask the more skeptical epidemiologists in the field what diet and lifestyle factors have been convincingly established as causes of common chronic diseases based on observational studies without clinical trials, you'll get a very short list: smoking as a cause of lung cancer and cardiovascular disease, sun exposure for skin cancer, sexual activity to spread the papilloma virus that causes

cervical cancer and perhaps alcohol for a few different cancers as well.

Richard Peto, professor of medical statistics and epidemiology at Oxford University, phrases the nature of the conflict this way: "Epidemiology is so beautiful and provides such an important perspective on human life and death, but an incredible amount of rubbish is published," by which he means the results of observational studies that appear daily in the news media and often become the basis of public-health recommendations about what we should or should not do to promote our continued good health.

In January 2001, the British epidemiologists George Davey Smith and Shah Ebrahim, co-editors of *The International Journal of Epidemiology*, discussed this issue in an editorial titled "Epidemiology – Is It Time to Call It a Day?" They noted that those few times that a randomized trial had been financed to test a hypothesis supported by results from these large observational studies, the hypothesis either failed the test or, at the very least, the test failed to confirm the hypothesis: antioxidants like vitamins E and C and beta carotene did not prevent heart disease, nor did eating copious fiber protect against colon cancer.

The Nurses' Health Study is the most influential of these cohort studies, and in the six years since the Davey Smith and Ebrahim editorial, a series of new trials have chipped away at its credibility. The Women's Health Initiative hormone-therapy trial failed to confirm the proposition that H.R.T. prevented heart disease; a W.H.I. diet trial with 49,000 women failed to confirm the notion that fruits and vegetables protected against heart disease; a 40,000-woman trial failed to confirm that a daily regimen of low-dose aspirin prevented colorectal cancer and heart attacks in women under 65. And this June, yet another clinical trial – this one of 1,000 men and women with a high risk of colon cancer – contradicted the inference from the Nurses's study that folic acid supplements reduced the risk of colon cancer. Rather, if anything, they appear to increase risk.

The implication of this track record seems hard to avoid. "Even the Nurses' Health Study, one of the biggest and best of these studies, cannot be used to reliably test small-to-moderate risks or benefits," says Charles Hennekens, a principal investigator with the Nurses' study from 1976 to 2001. "None of them can."

Proponents of the value of these studies for telling us how to prevent common diseases – including the epidemiologists who do them, and physicians, nutritionists and public-health authorities who use their findings to argue for or against the health benefits of a particular regimen – will argue that they are never relying on any single study. Instead, they base their ultimate judgments on the "totality of the data," which in theory includes all the observational evidence, any existing clinical trials and any laboratory work that might provide a biological mechanism to explain the observations.

This in turn leads to the argument that the fault is with the press, not the epidemiology. "The problem is not in the research but in the way it is interpreted for the public," as Jerome Kassirer and Marcia Angell, then the editors of *New England Journal of Medicine*, explained in a 1994 editorial titled "What Should the Public Believe?" Each study, they explained, is just a "piece of a puzzle" and so the media had to do a better job of communicating the many limitations of any single study and the caveats involved – the foremost, of course, being that "an association between two events is not the same as a cause and effect."

Stephen Pauker, a professor of medicine at Tufts University and a pioneer in the field of clinical decision making, says, "Epidemiologic studies, like diagnostic tests, are probabilistic statements." They don't tell us what the truth is, he says, but they allow both physicians and patients to "estimate the truth" so they can make informed decisions. The question the skeptics will ask, however, is how can anyone judge the value of these studies without taking into account their track record? And if they take into account the track record, suggests Sander Greenland, an epidemiologist at the University of California, Los Angeles, and an author of the textbook "Modern Epidemiology," then wouldn't they do just as well if they simply tossed a coin?

As John Bailar, an epidemiologist who is now at the National Academy of Science, once memorably phrased it, "The appropriate question is not whether there are uncertainties about epidemiologic data, rather, it is whether the uncertainties are so great that one cannot draw useful conclusions from the data."

Science vs. the Public Health:

Understanding how we got into this situation is the simple part of the story. The randomized-controlled trials needed to ascertain reliable knowledge about long-term risks and benefits of a drug, lifestyle factor or aspect of our diet are inordinately expensive and time consuming. By randomly assigning research subjects into an intervention group (who take a particular pill or eat a particular diet) or a placebo group, these trials "control" for all other possible variables, both known and unknown, that might effect the outcome: the relative health or wealth of the subjects, for instance. This is why randomized trials, particularly those known as placebo-controlled, double-blind trials, are typically considered the gold standard for establishing reliable knowledge about whether a drug, surgical intervention or diet is really safe and effective.

But clinical trials also have limitations beyond their exorbitant costs and the years or decades it takes them to provide meaningful results. They can rarely be used, for instance, to study suspected harmful effects. Randomly subjecting thousands of individuals to secondhand tobacco smoke, pollutants or potentially noxious trans fats presents obvious ethical dilemmas. And even when these trials are done to study the benefits of a particular intervention, it's rarely clear how the results apply to the public at large or to any specific patient. Clinical trials invariably enroll subjects who are relatively healthy, who are motivated to volunteer and will show up regularly for treatments and checkups. As a result, randomized trials "are very good for showing that a drug does what the pharmaceutical company says it does," David Atkins, a preventive-medicine specialist at the Agency for Healthcare Research and Quality, says, "but not very good for telling you how big the benefit really is and what are the harms in typical people. Because they don't enroll typical people."

These limitations mean that the job of establishing the long-term and relatively rare risks of drug therapies has fallen to observational studies, as has the job of determining the risks and benefits of virtually all factors of diet and lifestyle that might be related to chronic diseases. The former has been a fruitful field of research; many side effects of drugs have been discovered by these observational studies. The latter is the primary point of contention.

While the tools of epidemiology – comparisons of populations with and

without a disease – have proved effective over the centuries in establishing that a disease like cholera is caused by contaminated water, as the British physician John Snow demonstrated in the 1850s, it's a much more complicated endeavor when those same tools are employed to elucidate the more subtle causes of chronic disease.

And even the success stories taught in epidemiology classes to demonstrate the historical richness and potential of the field – that pellagra, a disease that can lead to dementia and death, is caused by a nutrient-deficient diet, for instance, as Joseph Goldberger demonstrated in the 1910s – are only known to be successes because the initial hypotheses were subjected to rigorous tests and happened to survive them. Goldberger tested the competing hypothesis, which posited that the disease was caused by an infectious agent, by holding what he called "filth parties," injecting himself and seven volunteers, his wife among them, with the blood of pellagra victims. They remained healthy, thus doing a compelling, if somewhat revolting, job of refuting the alternative hypothesis.

Smoking and lung cancer is the emblematic success story of chronic-disease epidemiology. But lung cancer was a rare disease before cigarettes became widespread, and the association between smoking and lung cancer was striking: heavy smokers had 2,000 to 3,000 percent the risk of those who had never smoked. This made smoking a "turkey shoot," says Greenland of U.C.L.A., compared with the associations epidemiologists have struggled with ever since, which fall into the tens of a percent range. The good news is that such small associations, even if causal, can be considered relatively meaningless for a single individual. If a 50-year-old woman with a small risk of breast cancer takes H.R.T. and increases her risk by 30 percent, it remains a small risk.

The compelling motivation for identifying these small effects is that their impact on the public health can be enormous if they're aggregated over an entire nation: if tens of millions of women decrease their breast cancer risk by 30 percent, tens of thousands of such cancers will be prevented each year. In fact, between 2002 and 2004, breast cancer incidence in the United States dropped by 12 percent, an effect that may have been caused by the coincident decline in the use of H.R.T. (And it may not have been. The coincident

reduction in breast cancer incidence and H.R.T. use is only an association.)

Saving tens of thousands of lives each year constitutes a powerful reason to lower the standard of evidence needed to suggest a cause-and-effect relationship – to take a leap of faith. This is the crux of the issue. From a scientific perspective, epidemiologic studies may be incapable of distinguishing a small effect from no effect at all, and so caution dictates that the scientist refrain from making any claims in that situation. From the public-health perspective, a small effect can be a very dangerous or beneficial thing, at least when aggregated over an entire nation, and so caution dictates that action be taken, even if that small effect might not be real. Hence the public-health logic that it's better to err on the side of prudence even if it means persuading us all to engage in an activity, eat a food or take a pill that does nothing for us and ignoring, for the moment, the possibility that such an action could have unforeseen harmful consequences. As Greenland says, "The combination of data, statistical methodology and motivation seems a potent anesthetic for skepticism."

The Bias of Healthy Users:

The Nurses' Health Study was founded at Harvard in 1976 by Frank Speizer, an epidemiologist who wanted to study the long-term effects of oral contraceptive use. It was expanded to include postmenopausal estrogen therapy because both treatments involved long-term hormone use by millions of women, and nobody knew the consequences. Speizer's assistants in this endeavor, who would go on to become the most influential epidemiologists in the country, were young physicians – Charles Hennekens, Walter Willett, Meir Stampfer and Graham Colditz – all interested in the laudable goal of preventing disease more than curing it after the fact.

When the Nurses' Health Study first published its observations on estrogen and heart disease in 1985, it showed that women taking estrogen therapy had only a third the risk of having a heart attack as had women who had never taken it; the association seemed compelling evidence for a cause and effect. Only 90 heart attacks had been reported among the 32,000 postmenopausal nurses in the study, and Stampfer, who had done the bulk of the analysis, and his colleagues "considered the possibility that the apparent protective effect of estrogen could be attributed to some other factor associated with its

use." They decided, though, as they have ever since, that this was unlikely. The paper's ultimate conclusion was that "further work is needed to define the optimal type, dose and duration of postmenopausal hormone use" for maximizing the protective benefit.

Only after Stampfer and his colleagues published their initial report on estrogen therapy did other investigators begin to understand the nature of the other factors that might explain the association. In 1987, Diana Petitti, an epidemiologist now at the University of Southern California, reported that she, too, had detected a reduced risk of heart-disease deaths among women taking H.R.T. in the Walnut Creek Study, a population of 16,500 women. When Petitti looked at all the data, however, she "found an even more dramatic reduction in death from homicide, suicide and accidents." With little reason to believe that estrogen would ward off homicides or accidents, Petitti concluded that something else appeared to be "confounding" the association she had observed. "The same thing causing this obvious spurious association might also be contributing to the lower risk of coronary heart disease," Petitti says today.

That mysterious something is encapsulated in what epidemiologists call the healthy-user bias, and some of the most fascinating research in observational epidemiology is now aimed at understanding this phenomenon in all its insidious subtlety. Only then can epidemiologists learn how to filter out the effect of this healthy-user bias from what might otherwise appear in their studies to be real causal relationships. One complication is that it encompasses a host of different and complex issues, many or most of which might be impossible to quantify. As Jerry Avorn of Harvard puts it, the effect of healthy-user bias has the potential for "big mischief" throughout these large epidemiologic studies.

At its simplest, the problem is that people who faithfully engage in activities that are good for them – taking a drug as prescribed, for instance, or eating what they believe is a healthy diet – are fundamentally different from those who don't. One thing epidemiologists have established with certainty, for example, is that women who take H.R.T. differ from those who don't in many ways, virtually all of which associate with lower heart-disease risk: they're thinner; they have fewer risk factors for heart disease to begin with;

they tend to be more educated and wealthier; to exercise more; and to be generally more health conscious.

Considering all these factors, is it possible to isolate one factor – hormone-replacement therapy – as the legitimate cause of the small association observed or even part of it? In one large population studied by Elizabeth Barrett-Connor, an epidemiologist at the University of California, San Diego, having gone to college was associated with a 50 percent lower risk of heart disease. So if women who take H.R.T. tend to be more educated than women who don't, this confounds the association between hormone therapy and heart disease. It can give the appearance of cause and effect where none exists.

Another thing that epidemiologic studies have established convincingly is that wealth associates with less heart disease and better health, at least in developed countries. The studies have been unable to establish why this is so, but this, too, is part of the healthy-user problem and a possible confounder of the hormone-therapy story and many of the other associations these epidemiologists try to study. George Davey Smith, who began his career studying how socioeconomic status associates with health, says one thing this research teaches is that misfortunes "cluster" together. Poverty is a misfortune, and the poor are less educated than the wealthy; they smoke more and weigh more; they're more likely to have hypertension and other heart-disease risk factors, to eat what's affordable rather than what the experts tell them is healthful, to have poor medical care and to live in environments with more pollutants, noise and stress. Ideally, epidemiologists will carefully measure the wealth and education of their subjects and then use statistical methods to adjust for the effect of these influences – multiple regression analysis, for instance, as one such method is called – but, as Avorn says, it "doesn't always work as well as we'd like it to."

The Nurses' investigators have argued that differences in socioeconomic status cannot explain the associations they observe with H.R.T. because all their subjects are registered nurses and so this "controls" for variations in wealth and education. The skeptics respond that even if all registered nurses had identical educations and income, which isn't necessarily the case, then their socioeconomic status will be determined by whether they're married, how many children they have and their husbands' income. "All you have to

do is look at nurses," Petitti says. "Some are married to C.E.O.'s of corporations and some are not married and still living with their parents. It cannot be true that there is no socioeconomic distribution among nurses." Stampfer says that since the Women's Health Initiative results came out in 2002, the Nurses' Health Study investigators went back into their data to examine socioeconomic status "to the extent that we could" – looking at measures that might indirectly reflect wealth and social class. "It doesn't seem plausible" that socioeconomic status can explain the association they observed, he says. But the Nurses' investigators never published that analysis, and so the skeptics have remained unconvinced.

The Bias of Compliance:

A still more subtle component of healthy-user bias has to be confronted. This is the compliance or adherer effect. Quite simply, people who comply with their doctors' orders when given a prescription are different and healthier than people who don't. This difference may be ultimately unquantifiable. The compliance effect is another plausible explanation for many of the beneficial associations that epidemiologists commonly report, which means this alone is a reason to wonder if much of what we hear about what constitutes a healthful diet and lifestyle is misconceived.

The lesson comes from an ambitious clinical trial called the Coronary Drug Project that set out in the 1970s to test whether any of five different drugs might prevent heart attacks. The subjects were some 8,500 middle-aged men with established heart problems. Two-thirds of them were randomly assigned to take one of the five drugs and the other third a placebo. Because one of the drugs, clofibrate, lowered cholesterol levels, the researchers had high hopes that it would ward off heart disease. But when the results were tabulated after five years, clofibrate showed no beneficial effect. The researchers then considered the possibility that clofibrate appeared to fail only because the subjects failed to faithfully take their prescriptions.

As it turned out, those men who said they took more than 80 percent of the pills prescribed fared substantially better than those who didn't. Only 15 percent of these faithful "adherers" died, compared with almost 25 percent of what the project researchers called "poor adherers." This might have been taken as reason to believe that clofibrate actually did cut heart-disease deaths

almost by half, but then the researchers looked at those men who faithfully took their placebos. And those men, too, seemed to benefit from adhering closely to their prescription: only 15 percent of them died compared with 28 percent who were less conscientious. "So faithfully taking the placebo cuts the death rate by a factor of two," says David Freedman, a professor of statistics at the University of California, Berkeley. "How can this be? Well, people who take their placebo regularly are just different than the others. The rest is a little speculative. Maybe they take better care of themselves in general. But this compliance effect is quite a big effect."

The moral of the story, says Freedman, is that whenever epidemiologists compare people who faithfully engage in some activity with those who don't – whether taking prescription pills or vitamins or exercising regularly or eating what they consider a healthful diet – the researchers need to account for this compliance effect or they will most likely infer the wrong answer. They'll conclude that this behavior, whatever it is, prevents disease and saves lives, when all they're really doing is comparing two different types of people who are, in effect, incomparable.

This phenomenon is a particularly compelling explanation for why the Nurses' Health Study and other cohort studies saw a benefit of H.R.T. in current users of the drugs, but not necessarily in past users. By distinguishing among women who never used H.R.T., those who used it but then stopped and current users (who were the only ones for which a consistent benefit appeared), these observational studies may have inadvertently focused their attention specifically on, as Jerry Avorn says, the "Girl Scouts in the group, the compliant ongoing users, who are probably doing a lot of other preventive things as well."

How Doctors Confound the Science:

Another complication to what may already appear (for good reason) to be a hopelessly confusing story is what might be called the prescriber effect. The reasons a physician will prescribe one medication to one patient and another or none at all to a different patient are complex and subtle. "Doctors go through a lot of different filters when they're thinking about what kind of drug to give to what kind of person," says Avorn, whose group at Harvard has spent much of the last decade studying this effect. "Maybe they give

the drug to their sickest patients; maybe they give it to the people for whom nothing else works."

It's this prescriber effect, combined with what Avorn calls the eager-patient effect, that is one likely explanation for why people who take cholesterol-lowering drugs called statins appear to have a greatly reduced risk of dementia and death from all causes compared with people who don't take statins. The medication itself is unlikely to be the primary cause in either case, says Avorn, because the observed associations are "so much larger than the effects that have been seen in randomized-clinical trials."

If we think like physicians, Avorn explains, then we get a plausible explanation: "A physician is not going to take somebody either dying of metastatic cancer or in a persistent vegetative state or with end-stage neurologic disease and say, 'Let's get that cholesterol down, Mrs. Jones.' The consequence of that, multiplied over tens of thousands of physicians, is that many people who end up on statins are a lot healthier than the people to whom these doctors do not give statins. Then add into that the people who come to the doctor and say, 'My brother-in-law is on this drug,' or, 'I saw it in a commercial,' or, 'I want to do everything I can to prevent heart disease, can I now have a statin, please?' Those kinds of patients are very different from the patients who don't come in. The *coup de grâce* then comes from the patients who consistently take their medications on an ongoing basis, and who are still taking them two or three years later. Those people are special and unusual and, as we know from clinical trials, even if they're taking a sugar pill they will have better outcomes."

The trick to successfully understanding what any association might really mean, Avorn adds, is "being clever." "The whole point of science is self-doubt," he says, "and asking could there be another explanation for what we're seeing."

H.R.T. and the Plausibility Problem:

Until the HERS and W.H.I. trials tested and refuted the hypothesis that hormone-replacement therapy protected women against heart disease, Stampfer, Willett and their colleagues argued that these alternative explanations could not account for what they observed. They had gathered so much information about their nurses, they said, that it allowed them to compare nurses who

took H.R.T. and engaged in health-conscious behaviors against women who didn't take H.R.T. and appeared to be equally health-conscious. Because this kind of comparison didn't substantially change the size of the association observed, it seemed reasonable to conclude that the association reflected the causal effect of H.R.T. After the W.H.I. results were published, says Stampfer, their faith was shaken, but only temporarily. Clinical trials, after all, also have limitations, and so the refutation of what was originally a simple hypothesis – that H.R.T. wards off heart disease – spurred new hypotheses, not quite so simple, to explain it.

At the moment, at least three plausible explanations exist for the discrepancy between the clinical trial results and those of the Nurses' Health Study and other observational studies. One is that the associations perceived by the epidemiologic studies were due to healthy-user and prescriber effects and not H.R.T. itself. Women who took H.R.T. had less heart disease than women who didn't, because women who took H.R.T. are different from women who didn't take H.R.T. And maybe their physicians are also different. In this case, the trials got the right answer; the observational studies got the wrong answer.

A second explanation is that the observational studies got the wrong answer, but only partly. Here, healthy-user and prescriber effects are viewed as minor issues; the question is whether observational studies can accurately determine if women were really taking H.R.T. before their heart attacks. This is a measurement problem, and one conspicuous limitation of all epidemiology is the difficulty of reliably assessing whatever it is the investigators are studying: not only determining whether or not subjects have really taken a medication or consumed the diet that they reported, but whether their subsequent diseases were correctly diagnosed. "The wonder and horror of epidemiology," Avorn says, "is that it's not enough to just measure one thing very accurately. To get the right answer, you may have to measure a great many things very accurately."

The most meaningful associations are those in which all the relevant factors can be ascertained reliably. Smoking and lung cancer, for instance. Lung cancer is an easy diagnosis to make, at least compared with heart disease. And "people sort of know whether they smoke a full pack a day or half or what

have you," says Graham Colditz, who recently left the Nurses' study and is now at Washington University School of Medicine in St. Louis. "That's one of the easier measures you can get." Epidemiologists will also say they believe in the associations between LDL cholesterol, blood pressure and heart disease, because these biological variables are measured directly. The measurements don't require that the study subjects fill out a questionnaire or accurately recall what their doctors may have told them.

Even the way epidemiologists frame the questions they ask can bias a measurement and produce an association that may be particularly misleading. If researchers believe that physical activity protects against chronic disease and they ask their subjects how much leisure-time physical activity they do each week, those who do more will tend to be wealthier and healthier, and so the result the researchers get will support their preconceptions. If the questionnaire asks how much physical activity a subject's job entails, the researchers might discover that the poor tend to be more physically active, because their jobs entail more manual labor, and they tend to have more chronic diseases. That would appear to refute the hypothesis.

The simpler the question or the more objective the measurement the more likely it is that an association may stand in the causal pathway, as these researchers put it. This is why the question of whether hormone-replacement therapy effects heart-disease risk, for instance, should be significantly easier to nail down than whether any aspect of diet does. For a measurement "as easy as this," says Jamie Robins, a Harvard epidemiologist, "where maybe the confounding is not horrible, maybe you can get it right." It's simply easier to imagine that women who have taken estrogen therapy will remember and report that correctly – it's yes or no, after all – than that they will recall and report accurately what they ate and how much of it over the last week or the last year.

But as the H.R.T. experience demonstrates, even the timing of a yes-or-no question can introduce problems. The subjects of the Nurses' Health Study were asked if they were taking H.R.T. every two years, which is how often the nurses were mailed new questionnaires about their diets, prescription drug use and whatever other factors the investigators deemed potentially relevant to health. If a nurse fills out her questionnaire a few months before she begins

taking H.R.T., as Colditz explains, and she then has a heart attack, say, six months later, the Nurses' study will classify that nurse as "not using" H.R.T. when she had the heart attack.

As it turns out, 40 percent of women who try H.R.T. stay on it for less than a year, and most of the heart attacks recorded in the W.H.I. and HERS trials occurred during the first few years that the women were prescribed the therapy. So it's a reasonable possibility that the Nurses' Health Study and other observational studies misclassified many of the heart attacks that occurred among users of hormone therapy as occurring among nonusers. This is the second plausible explanation for why these epidemiologic studies may have erroneously perceived a beneficial association of hormone use with heart disease and the clinical trials did not.

In the third explanation, the clinical trials and the observational studies both got the right answer, but they asked different questions. Here the relevant facts are that the women who took H.R.T. in the observational studies were mostly younger women going through menopause. Most of the women enrolled in the clinical trials were far beyond menopause. The average age of the women in the W.H.I. trial was 63 and in HERS it was 67. The primary goal of these clinical trials was to test the hypothesis that H.R.T. prevented heart disease. Older women have a higher risk of heart disease, and so by enrolling women in their 60s and 70s, the researchers didn't have to wait nearly as long to see if estrogen protected against heart disease as they would have if they only enrolled women in their 50s.

This means the clinical trials were asking what happens when older women were given H.R.T. years after menopause. The observational studies asked whether H.R.T. prevented heart disease when taken by younger women near the onset of menopause. A different question. The answer, according to Stampfer, Willett and their colleagues, is that estrogen protects those younger women – perhaps because their arteries are still healthy – while it induces heart attacks in the older women whose arteries are not. "It does seem clear now," Willett says, "that the observational studies got it all right. The W.H.I. also got it right for the question they asked: what happens if you start taking hormones many years after menopause? But that is not the question that most women have cared about."

This last explanation is now known as the "timing" hypothesis, and it certainly seems plausible. It has received some support from analyses of small subsets of the women enrolled in the W.H.I. trial, like the study published in June in *New England Journal of Medicine*. The dilemma at the moment is that the first two explanations are also plausible. If the compliance effect can explain why anyone faithfully following her doctor's orders will be 50 percent less likely to die over the next few years than someone who's not so inclined, then it's certainly possible that what the Nurses' Health Study and other observational studies did is observe a compliance effect and mistake it for a beneficial effect of H.R.T. itself. This would also explain why the Nurses' Health Study observed a 40 percent reduction in the yearly risk of death from all causes among women taking H.R.T. And it would explain why the Nurses' Health Study reported very similar seemingly beneficial effects for antioxidants, vitamins, low-dose aspirin and folic acid, and why these, too, were refuted by clinical trials. It's not necessarily true, but it certainly could be.

While Willett, Stampfer and their colleagues will argue confidently that they can reasonably rule out these other explanations based on everything they now know about their nurses – that they can correct or adjust for compliance and prescriber effects and still see a substantial effect of H.R.T. on heart disease – the skeptics argue that such confidence can never be justified without a clinical trial, at least not when the associations being studied are so small. "You can correct for what you can measure," says Rory Collins, an epidemiologist at Oxford University, "but you can't measure these things with precision so you will tend to under-correct for them. And you can't correct for things that you can't measure."

The investigators for the Nurses' Health Study "tend to believe everything they find," says Barrett-Connor of the University of California, San Diego. Barrett-Connor also studied hormone use and heart disease among a large group of women and observed and published the same association that the Nurses' Health Study did. She simply does not find the causal explanation as easy to accept, considering the plausibility of the alternatives. The latest variation on the therapeutic wisdom on H.R.T. is plausible, she says, but it remains untested. "Now we're back to the place where we're stuck with

observational epidemiology," she adds. "I'm back to the place where I doubt everything."

What to Believe?:

So how should we respond the next time we're asked to believe that an association implies a cause and effect, that some medication or some facet of our diet or lifestyle is either killing us or making us healthier? We can fall back on several guiding principles, these skeptical epidemiologists say. One is to assume that the first report of an association is incorrect or meaningless, no matter how big that association might be. After all, it's the first claim in any scientific endeavor that is most likely to be wrong. Only after that report is made public will the authors have the opportunity to be informed by their peers of all the many ways that they might have simply misinterpreted what they saw. The regrettable reality, of course, is that it's this first report that is most newsworthy. So be skeptical.

If the association appears consistently in study after study, population after population, but is small – in the range of tens of percent – then doubt it. For the individual, such small associations, even if real, will have only minor effects or no effect on overall health or risk of disease. They can have enormous public-health implications, but they're also small enough to be treated with suspicion until a clinical trial demonstrates their validity.

If the association involves some aspect of human behavior, which is, of course, the case with the great majority of the epidemiology that attracts our attention, then question its validity. If taking a pill, eating a diet or living in proximity to some potentially noxious aspect of the environment is associated with a particular risk of disease, then other factors of socioeconomic status, education, medical care and the whole gamut of healthy-user effects are as well. These will make the association, for all practical purposes, impossible to interpret reliably.

The exception to this rule is unexpected harm, what Avorn calls "bolt from the blue events," that no one, not the epidemiologists, the subjects or their physicians, could possibly have seen coming – higher rates of vaginal cancer, for example, among the children of women taking the drug DES to prevent miscarriage, or mesothelioma among workers exposed to asbestos. If the subjects are exposing themselves to a particular pill or a vitamin or

eating a diet with the goal of promoting health, and, lo and behold, it has no effect or a negative effect – it's associated with an increased risk of some disorder, rather than a decreased risk – then that's a bad sign and worthy of our consideration, if not some anxiety. Since healthy-user effects in these cases work toward reducing the association with disease, their failure to do so implies something unexpected is at work.

All of this suggests that the best advice is to keep in mind the law of unintended consequences. The reason clinicians test drugs with randomized trials is to establish whether the hoped-for benefits are real and, if so, whether there are unforeseen side effects that may outweigh the benefits. If the implication of an epidemiologist's study is that some drug or diet will bring us improved prosperity and health, then wonder about the unforeseen consequences. In these cases, it's never a bad idea to remain skeptical until somebody spends the time and the money to do a randomized trial and, contrary to much of the history of the endeavor to date, fails to refute it.

Gary Taubes is the author of the forthcoming book "Good Calories, Bad Calories: Challenging the Conventional Wisdom on Diet, Weight Control and Disease."

### 0.0.15 The Plastic Panic, Jerome Groopman (*New Yorker*), May 31, 2010

May 31, 2010

Jerome Groopman (*New Yorker*)

Bisphenol A, commonly known as BPA, may be among the world's most vilified chemicals. The compound, used in manufacturing polycarbonate plastic and epoxy resins, is found in plastic goggles, face shields, and helmets; baby bottles; protective coatings inside metal food containers; and composites and sealants used in dentistry. As animal studies began to show links between the chemical and breast and prostate cancer, early-onset puberty, and polycystic ovary syndrome, consumer groups pressured manufacturers of reusable plastic containers, like Nalgene, to remove BPA from their products. Warnings went out to avoid microwaving plasticware or putting it in the dishwasher. On May 6th, the President's Cancer Panel issued a report deploring the rising number of carcinogens released into the environment – including BPA – and calling for much more stringent regulation and wider awareness of their dangers. The panel advised President Obama "to use the power of your office to remove the carcinogens and other toxins from our food, water, and air that needlessly increase health care costs, cripple our Nation's productivity, and devastate American lives." Dr. LaSalle Leffall, Jr., the chairman of the panel, said in a statement, "The increasing number of known or suspected environmental carcinogens compels us to action, even though we may currently lack irrefutable proof of harm."

The narrative seems to follow a familiar path. In the nineteen-sixties, several animal studies suggested that cyclamates, a class of artificial sweetener, caused chromosomal abnormalities and cancer. Some three-quarters of Americans were estimated to consume the sweeteners. In 1969, cyclamates were banned. Later research found that there was little evidence that these substances caused cancer in humans. In the nineteen-eighties, studies suggesting a cancer risk from Alar, a chemical used to regulate the color and ripening of apples, caused a minor panic among parents and a media uproar. In that case, the cancer risk was shown to have been overstated, but still present, and the substance remains classified a "probable human carcinogen." Lead, too,

was for years thought to be safe in small doses, until further study demonstrated that, particularly for children, even slight exposure could result in intellectual delays, hearing loss, and hyperactivity.

There is an inherent uncertainty in determining which substances are safe and which are not, and when their risks outweigh their benefits. Toxicity studies are difficult, because BPA and other, similar chemicals can have multiple effects on the body. Moreover, we are exposed to scores of them in a lifetime, and their effects in combination or in sequence might be very different from what they would be in isolation. In traditional toxicology, a single chemical is tested in one cell or animal to assess its harmful effects. In studying environmental hazards, one needs to test mixtures of many chemicals, across ranges of doses, at different points in time, and at different ages, from conception to childhood to old age. Given so many variables, it is difficult to determine how harmful these chemicals might be, or if they are harmful at all, or what anyone can do to avoid their effects. In the case of BPA and other chemicals of its sort, though, their increasing prevalence and a number of human studies that associate them with developmental issues have become too worrisome to ignore. The challenge now is to decide a course of action before there is any certainty about what is truly dangerous and what is not.

In 1980, Frederica Perera, a professor at Columbia's Mailman School of Public Health and a highly regarded investigator of the effects of environmental hazards, was studying how certain chemicals in cigarette smoke might cause cancer. Dissatisfied with the research at the time, which measured toxic substances outside the body and then made inferences about their effects, she began using sophisticated molecular techniques to measure compounds called polycyclic aromatic hydrocarbons, or PAH – which are plentiful in tobacco smoke – in the body. Perera found that after entering the lungs the compounds pass into the bloodstream and damage blood cells, binding to their DNA. She hoped to compare the damaged blood cells from smokers with healthy cells, and decided to seek out those she imagined would be uncontaminated by foreign substances. "I thought that the most perfect pristine blood would come from the umbilical cord of a newborn," Perera said.

But when she analyzed her samples Perera discovered PAH attached to some of the DNA in blood taken from umbilical cords, too. "I was pretty

shocked," she said. "I realized that we did not know very much about what was happening during this early stage of development."

Perera's finding that chemicals like PAH, which can also be a component of air pollution, are passed from mother to child during pregnancy has now been replicated for more than two hundred compounds. These include PCBs, chemical coolants that were banned in the United States in 1979 but have persisted in the food chain; BPA and phthalates, used to make plastics more pliable, which leach out of containers and mix with their contents; pesticides used on crops and on insects in the home; and some flame retardants, which are often applied to upholstery, curtains, and other household items.

Fetuses and newborns lack functional enzymes in the liver and other organs that break down such chemicals, and animal studies in the past several decades have shown that these chemicals can disrupt hormones and brain development. Some scientists believe that they may promote chronic diseases seen in adulthood such as diabetes, atherosclerosis, and cancer. There is some evidence that they may have what are called epigenetic effects as well, altering gene expression in cells, including those which give rise to eggs and sperm, and allowing toxic effects to be passed on to future generations.

In 1998, Perera initiated a program at Columbia to investigate short- and long-term effects of environmental chemicals on children, and she now oversees one of the largest and longest-standing studies of a cohort of mothers and newborns in the United States. More than seven hundred mother-child pairs have been recruited from Washington Heights, Harlem, and the South Bronx; Perera is also studying pregnant women in Krakw, Poland, and two cities in China, and, since September 11, 2001, a group of three hundred and twenty-nine mothers and newborns from the downtown hospitals near the World Trade Center. In all, some two thousand mother-child pairs have been studied, many for at least a decade.

This March, I visited Columbia's Center for Children's Environmental Health, where Perera is the director, and met with a woman I'll call Renee Martin in an office overlooking the George Washington Bridge. Martin was born in Harlem, attended a community college in Queens, and then moved to 155th Street and Broadway, where she is raising her five children. She entered the study eleven years ago, when she was pregnant with her first

child. “I was asthmatic growing up,” Martin said. “And I was concerned about triggers of asthma in the environment. So when they asked me to be in the study I thought it would be a good way to get information that might tell me something about my own health and the health of my child.” She showed me a small black backpack containing a metal box with a long plastic tube. During her pregnancy, Martin would drape the tube over her shoulder, close to her chin, and a vacuum inside the device would suck in a sample of air. A filter trapped particles and vapors of ambient chemicals, like pesticides, phthalates, and PAH. “I walked around pregnant with this hose next to my mouth, but, living in New York, people hardly notice,” she said with a laugh.

The Columbia team also developed a comprehensive profile of Martin’s potential proximity to chemicals, including an environmental map that charted her apartment’s distance from gas stations, dry cleaners, fast-food restaurants, supermarkets, and major roadways. They took urine samples and, at delivery, blood samples from her and from the umbilical cord, along with samples from the placenta. Nearly a hundred per cent of the mothers in the study were found to have BPA and phthalates in their urine. Urine and blood samples are taken as the babies grow older, as well as samples of their exhaled breath. “We have a treasure trove of biological material,” Perera said. The researchers track the children’s weight and sexual development, and assess I.Q., visual spatial ability, attention, memory, and behavior. Brain imaging, using an M.R.I., is performed on selected children.

Martin was still breast-feeding her two-year-old daughter. “I bottle-fed my first child,” she told me. “But when you learn what can come out of plastic bottles and all the benefits of breast-feeding – my other children were nursed.” The Columbia group regularly convenes the families to hear results and discuss ways to reduce their exposure to potential environmental hazards. At one meeting, Martin found out that some widely used pesticides could result in impaired learning and behavior. “I told the landlord to stop spraying in the apartment” to combat a roach infestation, she said. On the advice of the Columbia researchers, Martin asked him to seal the cracks in the walls that were allowing cockroaches to enter, and Martin’s family meticulously swept up crumbs. This approach has now become the New York City

Department of Health's official recommendation for pest control. "You don't need to be out in the country and have compost," Martin said. "This has made me into an urban environmentalist."

In 2001, using data from animal studies, the E.P.A. banned the sale of the pesticide chlorpyrifos (sold under the name Dursban) for residential and indoor use. Many agricultural uses are still permitted, and farming communities continue to be exposed to the insecticide. Residues on food may affect those who live in urban areas as well. In 2004, the Columbia group published results in the journal *Environmental Health Perspectives* showing that significant exposure during the prenatal period to chlorpyrifos was associated with an average hundred-and-fifty-gram reduction in birth weight – about the same effect as if the mother had smoked all through pregnancy. Those most highly exposed to the insecticide were twice as likely to be born below the tenth percentile in size for gestational age. The researchers found that children born after 2001 had much lower exposure levels – indicating that the ban was largely effective.

For those children who were exposed to the pesticide in the womb, the effects have seemed to persist. The children with the greatest exposure were starting to fall off the developmental curve and displayed signs of attention-deficit problems by the time they were three. By seven, they showed significant deficits in working memory, which is strongly tied to problem-solving, I.Q., and reading comprehension. Another study, published this month in *Pediatrics*, using a random cross-section of American children, showed that an elevated level of a particular pesticide residue nearly doubled the likelihood that a child would have A.D.H.D.

"The size of this deficit is educationally meaningful in the early preschool years," Virginia Rauh, the leader of Columbia's research, said. "Such a decline can push whole groups of children into the developmentally delayed category."

First used in Germany, in the nineteen-thirties, bisphenol A has a chemical structure similar to that of estrogen, but was considered too weak to be developed into a contraceptive pill. Recent animal studies have shown that, even at very low levels, BPA can cause changes that may lead to cancer in the prostate gland and in breast tissue. It is also linked to disruption

in brain chemistry and, in female rodents, accelerated puberty. Japanese scientists found that high levels of BPA were associated with polycystic ovary syndrome, a leading cause of impaired fertility.

Phthalates are also ubiquitous in cosmetics, shampoos, and other personal-care products. They may have effects on older children and adults as well as on neonates. A study at Massachusetts General Hospital found an association of high levels of certain phthalates with lower sperm concentrations and impaired sperm motility; young girls in Puerto Rico who had developed breasts prematurely were more likely to have high levels of phthalates in their blood. Immigrant children in Belgium who exhibited precocious puberty also showed greater exposure to the pesticide DDT, which has estrogenlike effects and has been banned in the U.S., but is still used in Africa to help control malaria.

Long-term studies have provided the most compelling evidence that chemicals once considered safe may cause health problems in communities with consistent exposure over many years. Researchers from SUNY Albany, including Lawrence Schell, a biomedical anthropologist, have worked over the past two decades with Native Americans on the Mohawk reservation that borders the St. Lawrence River, once a major shipping thoroughfare, just east of Massena, New York. General Motors built a foundry nearby that made automobile parts, Alcoa had two manufacturing plants for aluminum, and the area was contaminated with PCBs, which were used in the three plants. Several Mohawk girls experienced signs of early puberty, which coincided with higher levels of PCBs in their blood.

The Albany researchers also observed that increased levels of PCBs correlated with altered levels of thyroid hormone and lower long-term memory functioning. Similar results have been found in an area of Slovakia near heavy industry. "Folks have complained about reproductive problems," Schell said, of the residents of the Mohawk reservation. "They talked a lot about rheumatoid arthritis, about lupus, about polycystic ovary syndrome. And, you know, you hear these things and you wonder how much of it is just a heightened sensitivity, but, when you see elevated antibodies that are often a sign of autoimmune disease of one kind or another, it could be the beginning of discovering a biological basis for their complaints about these diseases."

Beginning in 2003, Antonia Calafat, a chemist at the Centers for Disease Control and Prevention, and Russ Hauser, of the Harvard School of Public Health, set out to evaluate the exposure of premature infants to certain environmental contaminants. The researchers hypothesized that infants treated in the most intensive ways – intravenous feedings and delivery of oxygen by respirators – would receive the most exposure, since chemicals like phthalates and BPA can leach from plastic tubing. They studied forty-one infants from two Boston-area intensive-care units for BPA. Calafat told me, "We saw ten times the amounts of BPA in the neonates that we are seeing in the general population." In several children, the levels of BPA were more than a hundred times as high as in healthy Americans.

Calafat, who came to the United States from Spain on a Fulbright scholarship, developed highly accurate tests to detect BPA, phthalates, and other compounds in body fluids like blood and urine. This advance, she explained, "means that you are not simply doing an exposure assessment based on the concentration of the chemicals in the food or in the air or in the soil. You are actually measuring the concentrations in the body." With this technology, she can study each individual as if he or she were a single ecosystem. Her studies at the Centers for Disease Control show that 92.6 per cent of Americans aged six and older have detectable levels of BPA in their bodies; the levels in children between six and eleven years of age are twice as high as those in older Americans.

Critics such as Elizabeth Whelan, of the American Council on Science and Health, a consumer-education group in New York (Whelan says that about a third of its two-million-dollar annual budget comes from industry), think that the case against BPA and phthalates has more in common with those against cyclamates and Alar than with the one against lead. "The fears are irrational," she said. "People fear what they can't see and don't understand. Some environmental activists emotionally manipulate parents, making them feel that the ones they love the most, their children, are in danger." Whelan argues that the public should focus on proven health issues, such as the dangers of cigarettes and obesity and the need for bicycle helmets and other protective equipment. As for chemicals in plastics, Whelan says, "What the country needs is a national psychiatrist."

To illustrate what Whelan says is a misguided focus on manufactured chemicals, her organization has constructed a dinner menu "filled with natural foods, and you can find a carcinogen or an endocrine-disrupting chemical in every course" – for instance, tofu and soy products are filled with plant-based estrogens that could affect hormonal balance. "Just because you find something in the urine doesn't mean that it's a hazard," Whelan says. "Our understanding of risks and benefits is distorted. BPA helps protect food products from spoiling and causing botulism. Flame retardants save lives, so we don't burn up on our couch."

Several studies also contradict the conclusion that these chemicals have deleterious effects. The journal *Toxicological Sciences* recently featured a study from the E.P.A. scientist Earl Gray, a widely respected researcher, which indicated that BPA had no effect on puberty in rats. A study of military conscripts in Sweden found no connection between phthalates and depressed sperm counts, and a recent survey of newborns in New York failed to turn up an increase in a male genital malformation which might be expected if the effects from BPA seen in rodents were comparable to effects in humans. Richard Sharpe, a professor at the University of Edinburgh, and an internationally recognized pioneer on the effects of chemicals in the environment on endocrine disruption, recently wrote in *Toxicological Sciences*, "Fundamental, repetitive work on bisphenol A has sucked in tens, probably hundreds of millions of dollars from government bodies and industry, which, at a time when research money is thin on the ground, looks increasingly like an investment with a nil return."

With epidemiological studies, like those at Columbia, in which scientists observe people as they live, without a control group, the real-life nature of the project can make it difficult to distinguish between correlation and causation. Unknown factors in the environment or unreported habits might escape the notice of the researchers. Moreover, even sophisticated statistical analysis can sometimes yield specious results.

Dr. John Ioannides, an epidemiologist at the University of Ioannina, in Greece, has noted that four of the six most frequently cited epidemiological studies published in leading medical journals between 1990 and 2003 were later refuted. Demonstrating the malleability of data, Peter Austin, a med-

ical statistician at the Institute for Clinical Evaluative Sciences, in Toronto, has retrospectively analyzed medical records of the more than ten million residents of Ontario. He showed that Sagittarians are thirty-eight per cent more likely to fracture an arm than people of other astrological signs, and Leos are fifteen per cent more likely to suffer a gastrointestinal hemorrhage. (Pisces were more prone to heart failure.)

To help strengthen epidemiological analysis, Sir Austin Bradford Hill, a British medical statistician, set out certain criteria in 1965 that indicate cause and effect. Researchers must be sure that exposure to the suspected cause precedes the development of a disease; that there is a high degree of correlation between the two; that findings are replicated in different studies in various settings; that a biological explanation exists that makes the association plausible; and that increased exposure makes development of the disease more likely.

When epidemiological studies fulfill most of these criteria, they can be convincing, as when studies demonstrated a link between cigarettes and lung cancer. But, in an evolving field, dealing with chemicals that are part of daily life, the lack of long-term clinical data has made firm conclusions elusive. John Vandenbergh, a biologist who found that exposure to certain chemicals like BPA could accelerate the onset of puberty in mice, served on an expert panel that advised the National Toxicology Program, a part of the National Institute of Environmental Health Sciences, on the risks of exposure to BPA. In 2007, the panel reviewed more than three hundred scientific publications and concluded that "there is some concern" about exposure of fetuses and young children to BPA, given the research from Vandenbergh's laboratory and others.

Vandenbergh is cognizant of the difficulty of extrapolating data from rodents and lower animals to humans. "Why can't we just figure this out?" he said. "Well, one of the problems is that we would have to take half of the kids in the kindergarten and give them BPA and the other half not. Or expose half of the pregnant women to BPA in the doctor's office and the other half not. And then we have to wait thirty to fifty years to see what effects this has on their development, and whether they get more prostate cancer or breast cancer. You have to wait at least until puberty to see if there is an effect

on sexual maturation. Ethically, you are not going to go and feed people something if you think it harmful, and, second, you have this incredible time span to deal with."

The inadequacy of the current regulatory system contributes greatly to the atmosphere of uncertainty. The Toxic Substances Control Act, passed in 1976, does not require manufacturers to show that chemicals used in their products are safe before they go on the market; rather, the responsibility is placed on federal agencies, as well as on researchers in universities outside the government. The burden of proof is so onerous that bans on toxic chemicals can take years to achieve, and the government is often constrained from sharing information on specific products with the public, because manufacturers claim that such information is confidential. Several agencies split responsibility for oversight, with little cordination: the Food and Drug Administration supervises cosmetics, food, and medications, the Environmental Protection Agency regulates pesticides, and the Consumer Product Safety Commission oversees children's toys and other merchandise. The European Union, in contrast, now requires manufacturers to prove that their compounds are safe before they are sold.

According to the E.P.A., some eighty-two thousand chemicals are registered for use in commerce in the United States, with about seven hundred new chemicals introduced each year. In 1998, the E.P.A. found that, among chemicals produced in quantities of more than a million pounds per year, only seven per cent had undergone the full slate of basic toxicity studies. There is no requirement to label most consumer products for their chemical contents, and no consistent regulation throughout the country. Although the F.D.A. initially concluded that BPA was safe, some states, including Massachusetts and Connecticut, either have banned it or are considering a ban. (In January, the F.D.A. announced that it would conduct further testing.)

There has been some movement toward stricter controls: in July, 2008, Congress passed the Product Safety Improvement Act, which banned six phthalates from children's toys. But so far removal from other products has been voluntary. The President's Cancer Panel report advised people to reduce exposure with strategies that echo some of what the mothers in Frederica Perera's study have learned: choose products made with minimal

toxic substances, avoid using plastic containers to store liquids, and choose produce grown without pesticides or chemical fertilizers and meat free of antibiotics and hormones.

Mike Walls, the vice-president of regulatory affairs at the American Chemistry Council, a trade association that represents manufacturers of industrial chemicals, agrees that new laws are needed to regulate such chemicals. "Science has advanced since 1976, when the last legislation was enacted," he said. But Walls notes that some eight hundred thousand people are employed in the companies that the A.C.C. represents, and that their products are found in ninety-six per cent of all American manufactured goods. "The United States is the clear leader in chemistry," Walls said. "We have three times as many new applications for novel compounds as any other country in the world. We want to make good societal decisions but avoid regulations that will increase the burden on industry and stifle innovation."

Academic researchers have found that the enormous financial stakes – the production of BPA is a six-billion-dollar-a-year industry – have prompted extra scrutiny of their results. In 2007, according to a recent article in *Nature*, a majority of non-industry-supported studies initially deemed sound by the National Toxicology Program on the safety of BPA were dismissed as unsuitable after a representative of the A.C.C. drafted a memo critiquing their methods; experimental protocols often differ from one university lab to another. Researchers are now attempting to create a single standard protocol, and a bill introduced by Representative Louise Slaughter, of New York, would fund a centralized research facility at the National Institute of Environmental Health Sciences.

Other legislation aims to completely overhaul the 1976 law. "It's clear that the current system doesn't work at all," Ben Dunham, a staffer in the office of Senator Frank Lautenberg, of New Jersey, who crafted the bill now before the Senate, told me. Henry Waxman, of California, and Bobby Rush, of Illinois, have released a companion discussion draft in the House. Lautenberg's bill seeks to allow the E.P.A. to act quickly on chemicals that it considers dangerous; to give new power to the E.P.A. to establish safety criteria in chemical compounds; to create a database identifying chemicals in industrial products; and to set specific deadlines for approving or banning compounds.

The bill also seeks to limit the number of animals used for research. (Millions of animals are estimated to be required to perform the testing mandated under the E.U. law.) How much data would be needed to either restrict use of a chemical or mandate an outright ban is still unclear. Lautenberg's bill resisted the call of environmental groups to ban certain compounds like BPA immediately.

Dr. Gina Solomon, of the Natural Resources Defense Council, said that the Lautenberg bill is "an excellent first step," but noted several "gaps" in the bill: "There is what people call lack of a hammer, meaning no meaningful penalty for missing a deadline in evaluating a chemical if E.P.A. gets bogged down, and we know from history that it can be easily bogged down." The language setting a standard for safety is too vague, she added. "You could imagine industry driving a truck through this loophole."

Linda Birnbaum, the director of the N.I.E.H.S. and its National Toxicology Program, helps assess chemicals for the federal government and, if Slaughter's bill passes, could become responsible for much of the research surrounding these safety issues. Birnbaum's branch of the National Institutes of Health is working with the National Human Genome Research Institute and the E.P.A. to test thousands of compounds, singly and in combination, to assess their potential toxicity. Part of the difficulty, she points out, is that "what is normal for me may not be normal for you. We all have our own balance of different hormones in our different systems." When it comes to development and achievement, incremental differences – such as the drop of five to ten I.Q. points, or a lower birth weight – are significant. "We're all past the point of looking for missing arms and legs," Birnbaum said.

"I know of very little science where you will ever get hundred-per-cent certainty," Birnbaum says. "Science is constantly evolving, constantly learning new things, and at times decisions have to be made in the presence of a lot of information, but maybe not certainty. The problem is we don't always want to wait ten or twelve or twenty years to identify something that may be a problem."

Perera, who is keenly aware of the potential pitfalls of epidemiological research, told me that her team employs rigorous statistical methods to avoid falsely suggesting that one chemical or another is responsible for any

given result. And she objects to the characterization of her research as fear-mongering. "Our findings in children increasingly show real deleterious effects that can occur short-term and potentially for the rest of the child's life," Perera said. In January, the Columbia group published data from the mothers and infants it studied following September 11th. Cord-blood samples saved at the time of birth had been analyzed for the presence of flame retardants. Each year, the children were assessed for mental and motor development. As a point of reference, low-level lead poisoning results in an average loss of four to five I.Q. points. Those children in Columbia's group with the highest levels of flame retardant in their blood at birth had, by the age of four, I.Q. scores nearly six points lower than normal.

How do we go forward? Flame retardants surely serve a purpose, just as BPA and phthalates have made for better and stronger plastics. Still, while the evidence of these chemicals' health consequences may be far from conclusive, safer alternatives need to be sought. More important, policymakers must create a better system for making decisions about when to ban these types of substances, and must invest in the research that will inform those decisions. There's no guarantee that we'll always be right, but protecting those at the greatest risk shouldn't be deferred.

### 0.0.16 John Rock's Error, Malcolm Gladwell (*New Yorker*), March 10, 2000

March 10, 2000

Malcolm Gladwell (*New Yorker*)

What the co-inventor of the Pill didn't know about menstruation can endanger women's health.

John Rock was christened in 1890 at the Church of the Immaculate Conception in Marlborough, Massachusetts, and married by Cardinal William O'Connell, of Boston. He had five children and nineteen grandchildren. A crucifix hung above his desk, and nearly every day of his adult life he attended the 7 a.m. Mass at St. Mary's in Brookline. Rock, his friends would say, was in love with his church. He was also one of the inventors of the birth-control pill, and it was his conviction that his faith and his vocation were perfectly compatible. To anyone who disagreed he would simply repeat the words spoken to him as a child by his home-town priest: "John, always stick to your conscience. Never let anyone else keep it for you. And I mean anyone else." Even when Monsignor Francis W. Carney, of Cleveland, called him a "moral rapist," and when Frederick Good, the longtime head of obstetrics at Boston City Hospital, went to Boston's Cardinal Richard Cushing to have Rock excommunicated, Rock was unmoved. "You should be afraid to meet your Maker," one angry woman wrote to him, soon after the Pill was approved. "My dear madam," Rock wrote back, "in my faith, we are taught that the Lord is with us always. When my time comes, there will be no need for introductions."

In the years immediately after the Pill was approved by the F.D.A., in 1960, Rock was everywhere. He appeared in interviews and documentaries on CBS and NBC, in *Time*, *Newsweek*, *Life*, *The Saturday Evening Post*. He toured the country tirelessly. He wrote a widely discussed book, "The Time Has Come: A Catholic Doctor's Proposals to End the Battle Over Birth Control," which was translated into French, German, and Dutch. Rock was six feet three and rail-thin, with impeccable manners; he held doors open for his patients and addressed them as "Mrs." or "Miss." His mere association with the Pill helped make it seem respectable. "He was a man of great

dignity," Dr. Sheldon J. Segal, of the Population Council, recalls. "Even if the occasion called for an open collar, you'd never find him without an ascot. He had the shock of white hair to go along with that. And posture, straight as an arrow, even to his last year." At Harvard Medical School, he was a giant, teaching obstetrics for more than three decades. He was a pioneer in *in-vitro* fertilization and the freezing of sperm cells, and was the first to extract an intact fertilized egg. The Pill was his crowning achievement. His two collaborators, Gregory Pincus and Min-Cheuh Chang, worked out the mechanism. He shepherded the drug through its clinical trials. "It was his name and his reputation that gave ultimate validity to the claims that the pill would protect women against unwanted pregnancy," Loretta McLaughlin writes in her marvellous 1982 biography of Rock. Not long before the Pill's approval, Rock travelled to Washington to testify before the F.D.A. about the drug's safety. The agency examiner, Pasquale DeFelice, was a Catholic obstetrician from Georgetown University, and at one point, the story goes, DeFelice suggested the unthinkable–that the Catholic Church would never approve of the birth-control pill. "I can still see Rock standing there, his face composed, his eyes riveted on DeFelice," a colleague recalled years later, "and then, in a voice that would congeal your soul, he said, 'Young man, don't you sell my church short.' "

In the end, of course, John Rock's church disappointed him. In 1968, in the encyclical "Humanae Vitae," Pope Paul VI outlawed oral contraceptives and all other "artificial" methods of birth control. The passion and urgency that animated the birth-control debates of the sixties are now a memory. John Rock still matters, though, for the simple reason that in the course of reconciling his church and his work he made an error. It was not a deliberate error. It became manifest only after his death, and through scientific advances he could not have anticipated. But because that mistake shaped the way he thought about the Pill – about what it was, and how it worked, and most of all what it meant – and because John Rock was one of those responsible for the way the Pill came into the world, his error has colored the way people have thought about contraception ever since.

John Rock believed that the Pill was a "natural" method of birth control. By that he didn't mean that it felt natural, because it obviously didn't for

many women, particularly not in its earliest days, when the doses of hormone were many times as high as they are today. He meant that it worked by natural means. Women can get pregnant only during a certain interval each month, because after ovulation their bodies produce a surge of the hormone progesterone. Progesterone – one of a class of hormones known as progestin – prepares the uterus for implantation and stops the ovaries from releasing new eggs; it favors gestation. "It is progesterone, in the healthy woman, that prevents ovulation and establishes the pre- and post-menstrual 'safe' period," Rock wrote. When a woman is pregnant, her body produces a stream of progestin in part for the same reason, so that another egg can't be released and threaten the pregnancy already under way. Progestin, in other words, is nature's contraceptive. And what was the Pill? Progestin in tablet form. When a woman was on the Pill, of course, these hormones weren't coming in a sudden surge after ovulation and weren't limited to certain times in her cycle. They were being given in a steady dose, so that ovulation was permanently shut down. They were also being given with an additional dose of estrogen, which holds the endometrium together and – as we've come to learn – helps maintain other tissues as well. But to Rock, the timing and combination of hormones wasn't the issue. The key fact was that the Pill's ingredients duplicated what could be found in the body naturally. And in that naturalness he saw enormous theological significance.

In 1951, for example, Pope Pius XII had sanctioned the rhythm method for Catholics because he deemed it a "natural" method of regulating procreation: it didn't kill the sperm, like a spermicide, or frustrate the normal process of procreation, like a diaphragm, or mutilate the organs, like sterilization. Rock knew all about the rhythm method. In the nineteen-thirties, at the Free Hospital for Women, in Brookline, he had started the country's first rhythm clinic for educating Catholic couples in natural contraception. But how did the rhythm method work? It worked by limiting sex to the safe period that progestin created. And how did the Pill work? It worked by using progestin to extend the safe period to the entire month. It didn't mutilate the reproductive organs, or damage any natural process. "Indeed," Rock wrote, oral contraceptives "may be characterized as a 'pill-established safe period,' and would seem to carry the same moral implications" as the

rhythm method. The Pill was, to Rock, no more than "an adjunct to nature."

In 1958, Pope Pius XII approved the Pill for Catholics, so long as its contraceptive effects were "indirect" – that is, so long as it was intended only as a remedy for conditions like painful menses or "a disease of the uterus." That ruling emboldened Rock still further. Short-term use of the Pill, he knew, could regulate the cycle of women whose periods had previously been unpredictable. Since a regular menstrual cycle was necessary for the successful use of the rhythm method – and since the rhythm method was sanctioned by the Church – shouldn't it be permissible for women with an irregular menstrual cycle to use the Pill in order to facilitate the use of rhythm? And if that was true why not take the logic one step further? As the federal judge John T. Noonan writes in "Contraception," his history of the Catholic position on birth control:

If it was lawful to suppress ovulation to achieve a regularity necessary for successfully sterile intercourse, why was it not lawful to suppress ovulation without appeal to rhythm? If pregnancy could be prevented by pill plus rhythm, why not by pill alone? In each case suppression of ovulation was used as a means. How was a moral difference made by the addition of rhythm?

These arguments, as arcane as they may seem, were central to the development of oral contraception. It was John Rock and Gregory Pincus who decided that the Pill ought to be taken over a four-week cycle – a woman would spend three weeks on the Pill and the fourth week off the drug (or on a placebo), to allow for menstruation. There was and is no medical reason for this. A typical woman of childbearing age has a menstrual cycle of around twenty- eight days, determined by the cascades of hormones released by her ovaries. As first estrogen and then a combination of estrogen and progestin flood the uterus, its lining becomes thick and swollen, preparing for the implantation of a fertilized egg. If the egg is not fertilized, hormone levels plunge and cause the lining – the endometrium – to be sloughed off in a menstrual bleed. When a woman is on the Pill, however, no egg is released, because the Pill suppresses ovulation. The fluxes of estrogen and progestin that cause the lining of the uterus to grow are dramatically reduced, because the Pill slows down the ovaries. Pincus and Rock knew that the effect of the Pill's hormones on the endometrium was so modest that women

could conceivably go for months without having to menstruate. "In view of the ability of this compound to prevent menstrual bleeding as long as it is taken," Pincus acknowledged in 1958, "a cycle of any desired length could presumably be produced." But he and Rock decided to cut the hormones off after three weeks and trigger a menstrual period because they believed that women would find the continuation of their monthly bleeding reassuring. More to the point, if Rock wanted to demonstrate that the Pill was no more than a natural variant of the rhythm method, he couldn't very well do away with the monthly menses. Rhythm required "regularity," and so the Pill had to produce regularity as well.

It has often been said of the Pill that no other drug has ever been so instantly recognizable by its packaging: that small, round plastic dial pack. But what was the dial pack if not the physical embodiment of the twenty-eight-day cycle? It was, in the words of its inventor, meant to fit into a case "indistinguishable" from a woman's cosmetics compact, so that it might be carried "without giving a visual clue as to matters which are of no concern to others." Today, the Pill is still often sold in dial packs and taken in twenty-eight-day cycles. It remains, in other words, a drug shaped by the dictates of the Catholic Church – by John Rock's desire to make this new method of birth control seem as natural as possible. This was John Rock's error. He was consumed by the idea of the natural. But what he thought was natural wasn't so natural after all, and the Pill he ushered into the world turned out to be something other than what he thought it was. In John Rock's mind the dictates of religion and the principles of science got mixed up, and only now are we beginning to untangle them.

In 1986, a young scientist named Beverly Strassmann travelled to Africa to live with the Dogon tribe of Mali. Her research site was the village of Sangui in the Sahel, about a hundred and twenty miles south of Timbuktu. The Sahel is thorn savannah, green in the rainy season and semi-arid the rest of the year. The Dogon grow millet, sorghum, and onions, raise livestock, and live in adobe houses on the Bandiagara escarpment. They use no contraception. Many of them have held on to their ancestral customs and religious beliefs. Dogon farmers, in many respects, live much as people of that region have lived since antiquity. Strassmann wanted to construct a precise reproductive

profile of the women in the tribe, in order to understand what female biology might have been like in the millennia that preceded the modern age. In a way, Strassmann was trying to answer the same question about female biology that John Rock and the Catholic Church had struggled with in the early sixties: what is natural? Only, her sense of "natural" was not theological but evolutionary. In the era during which natural selection established the basic patterns of human biology – the natural history of our species – how often did women have children? How often did they menstruate? When did they reach puberty and menopause? What impact did breast-feeding have on ovulation? These questions had been studied before, but never so thoroughly that anthropologists felt they knew the answers with any certainty.

Strassmann, who teaches at the University of Michigan at Ann Arbor, is a slender, soft-spoken woman with red hair, and she recalls her time in Mali with a certain wry humor. The house she stayed in while in Sangui had been used as a shelter for sheep before she came and was turned into a pigsty after she left. A small brown snake lived in her latrine, and would curl up in a camouflaged coil on the seat she sat on while bathing. The villagers, she says, were of two minds: was it a deadly snake – Kere me jongolo, literally, "My bite cannot be healed" – or a harmless mouse snake? (It turned out to be the latter.) Once, one of her neighbors and best friends in the tribe roasted her a rat as a special treat. "I told him that white people aren't allowed to eat rat because rat is our totem," Strassmann says. "I can still see it. Bloated and charred. Stretched by its paws. Whiskers singed. To say nothing of the tail." Strassmann meant to live in Sangui for eighteen months, but her experiences there were so profound and exhilarating that she stayed for two and a half years. "I felt incredibly privileged," she says. "I just couldn't tear myself away."

Part of Strassmann's work focussed on the Dogon's practice of segregating menstruating women in special huts on the fringes of the village. In Sangui, there were two menstrual huts – dark, cramped, one-room adobe structures, with boards for beds. Each accommodated three women, and when the rooms were full, latecomers were forced to stay outside on the rocks. "It's not a place where people kick back and enjoy themselves," Strassmann says. "It's simply a nighttime hangout. They get there at dusk, and get up early in

the morning and draw their water." Strassmann took urine samples from the women using the hut, to confirm that they were menstruating. Then she made a list of all the women in the village, and for her entire time in Mali – seven hundred and thirty-six consecutive nights – she kept track of everyone who visited the hut. Among the Dogon, she found, a woman, on average, has her first period at the age of sixteen and gives birth eight or nine times. From menarche, the onset of menstruation, to the age of twenty, she averages seven periods a year. Over the next decade and a half, from the age of twenty to the age of thirty-four, she spends so much time either pregnant or breast-feeding (which, among the Dogon, suppresses ovulation for an average of twenty months) that she averages only slightly more than one period per year. Then, from the age of thirty-five until menopause, at around fifty, as her fertility rapidly declines, she averages four menses a year. All told, Dogon women menstruate about a hundred times in their lives. (Those who survive early childhood typically live into their seventh or eighth decade.) By contrast, the average for contemporary Western women is somewhere between three hundred and fifty and four hundred times.

Strassmann's office is in the basement of a converted stable next to the Natural History Museum on the University of Michigan campus. Behind her desk is a row of battered filing cabinets, and as she was talking she turned and pulled out a series of yellowed charts. Each page listed, on the left, the first names and identification numbers of the Sangui women. Across the top was a time line, broken into thirty-day blocks. Every menses of every woman was marked with an X. In the village, Strassmann explained, there were two women who were sterile, and, because they couldn't get pregnant, they were regulars at the menstrual hut. She flipped through the pages until she found them. "Look, she had twenty-nine menses over two years, and the other had twenty- three." Next to each of their names was a solid line of x's. "Here's a woman approaching menopause," Strassmann went on, running her finger down the page. "She's cycling but is a little bit erratic. Here's another woman of prime childbearing age. Two periods. Then pregnant. I never saw her again at the menstrual hut. This woman here didn't go to the menstrual hut for twenty months after giving birth, because she was breast-feeding. Two periods. Got pregnant. Then she miscarried, had a

few periods, then got pregnant again. This woman had three menses in the study period." There weren't a lot of x's on Strassmann's sheets. Most of the boxes were blank. She flipped back through her sheets to the two anomalous women who were menstruating every month. "If this were a menstrual chart of undergraduates here at the University of Michigan, all the rows would be like this."

Strassmann does not claim that her statistics apply to every preindustrial society. But she believes – and other anthropological work backs her up – that the number of lifetime menses isn't greatly affected by differences in diet or climate or method of subsistence (foraging versus agriculture, say). The more significant factors, Strassmann says, are things like the prevalence of wet-nursing or sterility. But over all she believes that the basic pattern of late menarche, many pregnancies, and long menstrual-free stretches caused by intensive breast-feeding was virtually universal up until the "demographic transition" of a hundred years ago from high to low fertility. In other words, what we think of as normal – frequent menses – is in evolutionary terms abnormal. "It's a pity that gynecologists think that women have to menstruate every month," Strassmann went on. "They just don't understand the real biology of menstruation."

To Strassmann and others in the field of evolutionary medicine, this shift from a hundred to four hundred lifetime menses is enormously significant. It means that women's bodies are being subjected to changes and stresses that they were not necessarily designed by evolution to handle. In a brilliant and provocative book, "Is Menstruation Obsolete?," Drs. Elsimar Coutinho and Sheldon S. Segal, two of the world's most prominent contraceptive researchers, argue that this recent move to what they call "incessant ovulation" has become a serious problem for women's health. It doesn't mean that women are always better off the less they menstruate. There are times – particularly in the context of certain medical conditions – when women ought to be concerned if they aren't menstruating: In obese women, a failure to menstruate can signal an increased risk of uterine cancer. In female athletes, a failure to menstruate can signal an increased risk of osteoporosis. But for most women, Coutinho and Segal say, incessant ovulation serves no purpose except to increase the occurence of abdominal pain, mood shifts, migraines,

endometriosis, fibroids, and anemia – the last of which, they point out, is "one of the most serious health problems in the world."

Most serious of all is the greatly increased risk of some cancers. Cancer, after all, occurs because as cells divide and reproduce they sometimes make mistakes that cripple the cells' defenses against runaway growth. That's one of the reasons that our risk of cancer generally increases as we age: our cells have more time to make mistakes. But this also means that any change promoting cell division has the potential to increase cancer risk, and ovulation appears to be one of those changes. Whenever a woman ovulates, an egg literally bursts through the walls of her ovaries. To heal that puncture, the cells of the ovary wall have to divide and reproduce. Every time a woman gets pregnant and bears a child, her lifetime risk of ovarian cancer drops ten per cent. Why? Possibly because, between nine months of pregnancy and the suppression of ovulation associated with breast-feeding, she stops ovulating for twelve months – and saves her ovarian walls from twelve bouts of cell division. The argument is similar for endometrial cancer. When a woman is menstruating, the estrogen that flows through her uterus stimulates the growth of the uterine lining, causing a flurry of potentially dangerous cell division. Women who do not menstruate frequently spare the endometrium that risk. Ovarian and endometrial cancer are characteristically modern diseases, consequences, in part, of a century in which women have come to menstruate four hundred times in a lifetime.

In this sense, the Pill really does have a "natural" effect. By blocking the release of new eggs, the progestin in oral contraceptives reduces the rounds of ovarian cell division. Progestin also counters the surges of estrogen in the endometrium, restraining cell division there. A woman who takes the Pill for ten years cuts her ovarian-cancer risk by around seventy per cent and her endometrial-cancer risk by around sixty per cent. But here "natural" means something different from what Rock meant. He assumed that the Pill was natural because it was an unobtrusive variant of the body's own processes. In fact, as more recent research suggests, the Pill is really only natural in so far as it's radical – rescuing the ovaries and endometrium from modernity. That Rock insisted on a twenty-eight-day cycle for his pill is evidence of just how deep his misunderstanding was: the real promise of the Pill was not that

it could preserve the menstrual rhythms of the twentieth century but that it could disrupt them.

Today, a growing movement of reproductive specialists has begun to campaign loudly against the standard twenty-eight-day pill regimen. The drug company Organon has come out with a new oral contraceptive, called Mircette, that cuts the seven-day placebo interval to two days. Patricia Sulak, a medical researcher at Texas A. & M. University, has shown that most women can probably stay on the Pill, straight through, for six to twelve weeks before they experience breakthrough bleeding or spotting. More recently, Sulak has documented precisely what the cost of the Pill's monthly "off" week is. In a paper in the February issue of the journal *Obstetrics and Gynecology*, she and her colleagues documented something that will come as no surprise to most women on the Pill: during the placebo week, the number of users experiencing pelvic pain, bloating, and swelling more than triples, breast tenderness more than doubles, and headaches increase by almost fifty per cent. In other words, some women on the Pill continue to experience the kinds of side effects associated with normal menstruation. Sulak's paper is a short, dry, academic work, of the sort intended for a narrow professional audience. But it is impossible to read it without being struck by the consequences of John Rock's desire to please his church. In the past forty years, millions of women around the world have been given the Pill in such a way as to maximize their pain and suffering. And to what end? To pretend that the Pill was no more than a pharmaceutical version of the rhythm method?

In 1980 and 1981, Malcolm Pike, a medical statistician at the University of Southern California, travelled to Japan for six months to study at the Atomic Bomb Casualties Commission. Pike wasn't interested in the effects of the bomb. He wanted to examine the medical records that the commission had been painstakingly assembling on the survivors of Hiroshima and Nagasaki. He was investigating a question that would ultimately do as much to complicate our understanding of the Pill as Strassmann's research would a decade later: why did Japanese women have breast-cancer rates six times lower than American women?

In the late forties, the World Health Organization began to collect and publish comparative health statistics from around the world, and the breast-

cancer disparity between Japan and America had come to obsess cancer specialists. The obvious answer – that Japanese women were somehow genetically protected against breast cancer – didn't make sense, because once Japanese women moved to the United States they began to get breast cancer almost as often as American women did. As a result, many experts at the time assumed that the culprit had to be some unknown toxic chemical or virus unique to the West. Brian Henderson, a colleague of Pike's at U.S.C. and his regular collaborator, says that when he entered the field, in 1970, "the whole viral – and chemical – carcinogenesis idea was huge – it dominated the literature." As he recalls, "Breast cancer fell into this large, unknown box that said it was something to do with the environment – and that word 'environment' meant a lot of different things to a lot of different people. They might be talking about diet or smoking or pesticides."

Henderson and Pike, however, became fascinated by a number of statistical pecularities. For one thing, the rate of increase in breast-cancer risk rises sharply throughout women's thirties and forties and then, at menopause, it starts to slow down. If a cancer is caused by some toxic outside agent, you'd expect that rate to rise steadily with each advancing year, as the number of mutations and genetic mistakes steadily accumulates. Breast cancer, by contrast, looked as if it were being driven by something specific to a woman's reproductive years. What was more, younger women who had had their ovaries removed had a markedly lower risk of breast cancer; when their bodies weren't producing estrogen and progestin every month, they got far fewer tumors. Pike and Henderson became convinced that breast cancer was linked to a process of cell division similar to that of ovarian and endometrial cancer. The female breast, after all, is just as sensitive to the level of hormones in a woman's body as the reproductive system. When the breast is exposed to estrogen, the cells of the terminal-duct lobular unit – where most breast cancer arises – undergo a flurry of division. And during the mid-to-late stage of the menstrual cycle, when the ovaries start producing large amounts of progestin, the pace of cell division in that region doubles.

It made intuitive sense, then, that a woman's risk of breast cancer would be linked to the amount of estrogen and progestin her breasts have been exposed to during her lifetime. How old a woman is at menarche should make

a big difference, because the beginning of puberty results in a hormonal surge through a woman's body, and the breast cells of an adolescent appear to be highly susceptible to the errors that result in cancer. (For more complicated reasons, bearing children turns out to be protective against breast cancer, perhaps because in the last two trimesters of pregnancy the cells of the breast mature and become much more resistant to mutations.) How old a woman is at menopause should matter, and so should how much estrogen and progestin her ovaries actually produce, and even how much she weighs after menopause, because fat cells turn other hormones into estrogen.

Pike went to Hiroshima to test the cell-division theory. With other researchers at the medical archive, he looked first at the age when Japanese women got their period. A Japanese woman born at the turn of the century had her first period at sixteen and a half. American women born at the same time had their first period at fourteen. That difference alone, by their calculation, was sufficient to explain forty per cent of the gap between American and Japanese breast-cancer rates. "They had collected amazing records from the women of that area," Pike said. "You could follow precisely the change in age of menarche over the century. You could even see the effects of the Second World War. The age of menarche of Japanese girls went up right at that point because of poor nutrition and other hardships. And then it started to go back down after the war. That's what convinced me that the data were wonderful."

Pike, Henderson, and their colleagues then folded in the other risk factors. Age at menopause, age at first pregnancy, and number of children weren't sufficiently different between the two countries to matter. But weight was. The average post-menopausal Japanese woman weighed a hundred pounds; the average American woman weighed a hundred and forty-five pounds. That fact explained another twenty-five per cent of the difference. Finally, the researchers analyzed blood samples from women in rural Japan and China, and found that their ovaries – possibly because of their extremely low-fat diet – were producing about seventy-five per cent the amount of estrogen that American women were producing. Those three factors, added together, seemed to explain the breast-cancer gap. They also appeared to explain why the rates of breast cancer among Asian women began to increase when they

came to America: on an American diet, they started to menstruate earlier, gained more weight, and produced more estrogen. The talk of chemicals and toxins and power lines and smog was set aside. "When people say that what we understand about breast cancer explains only a small amount of the problem, that it is somehow a mystery, it's absolute nonsense," Pike says flatly. He is a South African in his sixties, with graying hair and a salt-and-pepper beard. Along with Henderson, he is an eminent figure in cancer research, but no one would ever accuse him of being tentative in his pronouncements. "We understand breast cancer extraordinarily well. We understand it as well as we understand cigarettes and lung cancer."

What Pike discovered in Japan led him to think about the Pill, because a tablet that suppressed ovulation – and the monthly tides of estrogen and progestin that come with it – obviously had the potential to be a powerful anti-breast-cancer drug. But the breast was a little different from the reproductive organs. Progestin prevented ovarian cancer because it suppressed ovulation. It was good for preventing endometrial cancer because it countered the stimulating effects of estrogen. But in breast cells, Pike believed, progestin wasn't the solution; it was one of the hormones that caused cell division. This is one explanation for why, after years of studying the Pill, researchers have concluded that it has no effect one way or the other on breast cancer: whatever beneficial effect results from what the Pill does is canceled out by how it does it. John Rock touted the fact that the Pill used progestin, because progestin was the body's own contraceptive. But Pike saw nothing "natural" about subjecting the breast to that heavy a dose of progestin. In his view, the amount of progestin and estrogen needed to make an effective contraceptive was much greater than the amount needed to keep the reproductive system healthy – and that excess was unnecessarily raising the risk of breast cancer. A truly natural Pill might be one that found a way to suppress ovulation without using progestin. Throughout the nineteen-eighties, Pike recalls, this was his obsession. "We were all trying to work out how the hell we could fix the Pill. We thought about it day and night."

Pike's proposed solution is a class of drugs known as GnRHAs, which has been around for many years. GnRHAs disrupt the signals that the pituitary gland sends when it is attempting to order the manufacture of sex hormones.

It's a circuit breaker. "We've got substantial experience with this drug," Pike says. Men suffering from prostate cancer are sometimes given a GnRHA to temporarily halt the production of testosterone, which can exacerbate their tumors. Girls suffering from what's called precocious puberty – puberty at seven or eight, or even younger – are sometimes given the drug to forestall sexual maturity. If you give GnRHA to women of childbearing age, it stops their ovaries from producing estrogen and progestin. If the conventional Pill works by convincing the body that it is, well, a little bit pregnant, Pike's pill would work by convincing the body that it was menopausal.

In the form Pike wants to use it, GnRHA will come in a clear glass bottle the size of a saltshaker, with a white plastic mister on top. It will be inhaled nasally. It breaks down in the body very quickly. A morning dose simply makes a woman menopausal for a while. Menopause, of course, has its risks. Women need estrogen to keep their hearts and bones strong. They also need progestin to keep the uterus healthy. So Pike intends to add back just enough of each hormone to solve these problems, but much less than women now receive on the Pill. Ideally, Pike says, the estrogen dose would be adjustable: women would try various levels until they found one that suited them. The progestin would come in four twelve-day stretches a year. When someone on Pike's regimen stopped the progestin, she would have one of four annual menses.

Pike and an oncologist named Darcy Spicer have joined forces with another oncologist, John Daniels, in a startup called Balance Pharmaceuticals. The firm operates out of a small white industrial strip mall next to the freeway in Santa Monica. One of the tenants is a paint store, another looks like some sort of export company. Balance's offices are housed in an oversized garage with a big overhead door and concrete floors. There is a tiny reception area, a little coffee table and a couch, and a warren of desks, bookshelves, filing cabinets, and computers. Balance is testing its formulation on a small group of women at high risk for breast cancer, and if the results continue to be encouraging, it will one day file for F.D.A. approval.

"When I met Darcy Spicer a couple of years ago," Pike said recently, as he sat at a conference table deep in the Balance garage, "he said, 'Why don't we just try it out? By taking mammograms, we should be able to see

changes in the breasts of women on this drug, even if we add back a little estrogen to avoid side effects.' So we did a study, and we found that there were huge changes." Pike pulled out a paper he and Spicer had published in the *Journal of the National Cancer Institute*, showing breast X-rays of three young women. "These are the mammograms of the women before they start," he said. Amid the grainy black outlines of the breast were large white fibrous clumps – clumps that Pike and Spicer believe are indicators of the kind of relentless cell division that increases breast-cancer risk. Next to those x-rays were three mammograms of the same women taken after a year on the GnRHA regimen. The clumps were almost entirely gone. "This to us represents that we have actually stopped the activity inside the breasts," Pike went on. "White is a proxy for cell proliferation. We're slowing down the breast."

Pike stood up from the table and turned to a sketch pad on an easel behind him. He quickly wrote a series of numbers on the paper. "Suppose a woman reaches menarche at fifteen and menopause at fifty. That's thirty-five years of stimulating the breast. If you cut that time in half, you will change her risk not by half but by half raised to the power of 4.5." He was working with a statistical model he had developed to calculate breast-cancer risk. "That's one-twenty-third. Your risk of breast cancer will be one-twenty-third of what it would be otherwise. It won't be zero. You can't get to zero. If you use this for ten years, your risk will be cut by at least half. If you use it for five years, your risk will be cut by at least a third. It's as if your breast were to be five years younger, or ten years younger – forever." The regimen, he says, should also provide protection against ovarian cancer.

Pike gave the sense that he had made this little speech many times before, to colleagues, to his family and friends – and to investors. He knew by now how strange and unbelievable what he was saying sounded. Here he was, in a cold, cramped garage in the industrial section of Santa Monica, arguing that he knew how to save the lives of hundreds of thousands of women around the world. And he wanted to do that by making young women menopausal through a chemical regimen sniffed every morning out of a bottle. This was, to say the least, a bold idea. Could he strike the right balance between the hormone levels women need to stay healthy and those that ultimately make

them sick? Was progestin really so important in breast cancer? There are cancer specialists who remain skeptical. And, most of all, what would women think? John Rock, at least, had lent the cause of birth control his Old World manners and distinguished white hair and appeals from theology; he took pains to make the Pill seem like the least radical of interventions – nature's contraceptive, something that could be slipped inside a woman's purse and pass without notice. Pike was going to take the whole forty-year mythology of "natural" and sweep it aside. "Women are going to think, I'm being manipulated here. And it's a perfectly reasonable thing to think." Pike's South African accent gets a little stronger as he becomes more animated. "But the modern way of living represents an extraordinary change in female biology. Women are going out and becoming lawyers, doctors, presidents of countries. They need to understand that what we are trying to do isn't abnormal. It's just as normal as when someone hundreds of years ago had menarche at seventeen and had five babies and had three hundred fewer menstrual cycles than most women have today. The world is not the world it was. And some of the risks that go with the benefits of a woman getting educated and not getting pregnant all the time are breast cancer and ovarian cancer, and we need to deal with it. I have three daughters. The earliest grandchild I had was when one of them was thirty-one. That's the way many women are now. They ovulate from twelve or thirteen until their early thirties. Twenty years of uninterrupted ovulation before their first child! That's a brand-new phenomenon!"

John Rock's long battle on behalf of his birth-control pill forced the Church to take notice. In the spring of 1963, just after Rock's book was published, a meeting was held at the Vatican between high officials of the Catholic Church and Donald B. Straus, the chairman of Planned Parenthood. That summit was followed by another, on the campus of the University of Notre Dame. In the summer of 1964, on the eve of the feast of St. John the Baptist, Pope Paul VI announced that he would ask a committee of church officials to rexamine the Vatican's position on contraception. The group met first at the Collegio San Jose, in Rome, and it was clear that a majority of the committee were in favor of approving the Pill. Committee reports leaked to the *National Catholic Register* confirmed that Rock's case appeared to be winning. Rock

was elated. *Newsweek* put him on its cover, and ran a picture of the Pope inside. "Not since the Copernicans suggested in the sixteenth century that the sun was the center of the planetary system has the Roman Catholic Church found itself on such a perilous collision course with a new body of knowledge," the article concluded. Paul VI, however, was unmoved. He stalled, delaying a verdict for months, and then years. Some said he fell under the sway of conservative elements within the Vatican. In the interim, theologians began exposing the holes in Rock's arguments. The rhythm method "'prevents' conception by abstinence, that is, by the non-performance of the conjugal act during the fertile period," the Catholic journal *America* concluded in a 1964 editorial. "The pill prevents conception by suppressing ovulation and by thus abolishing the fertile period. No amount of word juggling can make abstinence from sexual relations and the suppression of ovulation one and the same thing." On July 29, 1968, in the "Humanae Vitae" encyclical, the Pope broke his silence, declaring all "artificial" methods of contraception to be against the teachings of the Church.

In hindsight, it is possible to see the opportunity that Rock missed. If he had known what we know now and had talked about the Pill not as a contraceptive but as a cancer drug – not as a drug to prevent life but as one that would save life – the church might well have said yes. Hadn't Pius XII already approved the Pill for therapeutic purposes? Rock would only have had to think of the Pill as Pike thinks of it: as a drug whose contraceptive aspects are merely a means of attracting users, of getting, as Pike put it, "people who are young to take a lot of stuff they wouldn't otherwise take."

But Rock did not live long enough to understand how things might have been. What he witnessed, instead, was the terrible time at the end of the sixties when the Pill suddenly stood accused – wrongly – of causing blood clots, strokes, and heart attacks. Between the mid-seventies and the early eighties, the number of women in the United States using the Pill fell by half. Harvard Medical School, meanwhile, took over Rock's Reproductive Clinic and pushed him out. His Harvard pension paid him only seventy-five dollars a year. He had almost no money in the bank and had to sell his house in Brookline. In 1971, Rock left Boston and retreated to a farmhouse in the hills of New Hampshire. He swam in the stream behind the house. He listened to

John Philip Sousa marches. In the evening, he would sit in the living room with a pitcher of martinis. In 1983, he gave his last public interview, and it was as if the memory of his achievements was now so painful that he had blotted it out.

He was asked what the most gratifying time of his life was. "Right now," the inventor of the Pill answered, incredibly. He was sitting by the fire in a crisp white shirt and tie, reading "The Origin," Irving Stone's fictional account of the life of Darwin. "It frequently occurs to me, gosh, what a lucky guy I am. I have no responsibilities, and I have everything I want. I take a dose of equanimity every twenty minutes. I will not be disturbed about things."

Once, John Rock had gone to seven-o'clock Mass every morning and kept a crucifix above his desk. His interviewer, the writer Sara Davidson, moved her chair closer to his and asked him whether he still believed in an afterlife.

"Of course I don't," Rock answered abruptly. Though he didn't explain why, his reasons aren't hard to imagine. The church could not square the requirements of its faith with the results of his science, and if the church couldn't reconcile them how could Rock be expected to? John Rock always stuck to his conscience, and in the end his conscience forced him away from the thing he loved most. This was not John Rock's error. Nor was it his church's. It was the fault of the haphazard nature of science, which all too often produces progress in advance of understanding. If the order of events in the discovery of what was natural had been reversed, his world, and our world, too, would have been a different place.

"Heaven and Hell, Rome, all the Church stuff – that's for the solace of the multitude," Rock said. He had only a year to live. "I was an ardent practicing Catholic for a long time, and I really believed it all then, you see."

### 0.0.17 The Truth Wears Off, Jonah Lehrer (*New Yorker*), December 13, 2010

Is there something wrong with the scientific method?

Many results that are rigorously proved and accepted start shrinking in later studies. On September 18, 2007, a few dozen neuroscientists, psychiatrists, and drug-company executives gathered in a hotel conference room in Brussels to hear some startling news. It had to do with a class of drugs known as atypical or second-generation antipsychotics, which came on the market in the early nineties. The drugs, sold under brand names such as Abilify, Seroquel, and Zyprexa, had been tested on schizophrenics in several large clinical trials, all of which had demonstrated a dramatic decrease in the subjects's psychiatric symptoms. As a result, second generation antipsychotics had become one of the fastest-growing and most profitable pharmaceutical classes. By 2001, Eli Lilly's Zyprexa was generating more revenue than Prozac. It remains the company's top-selling drug.

But the data presented at the Brussels meeting made it clear that something strange was happening: the therapeutic power of the drugs appeared to be steadily waning. A recent study showed an effect that was less than half of that documented in the first trials, in the early nineteen-nineties. Many researchers began to argue that the expensive pharmaceuticals weren't any better than first-generation antipsychotics, which have been in use since the fifties. "In fact, sometimes they now look even worse," John Davis, a professor of psychiatry at the University of Illinois at Chicago, told me.

Before the effectiveness of a drug can be confirmed, it must be tested and tested again. Different scientists in different labs need to repeat the protocols and publish their results. The test of replicability, as it's known, is the foundation of modern research. Replicability is how the community enforces itself. It's a safeguard for the creep of subjectivity. Most of the time, scientists know what results they want, and that can influence the results they get. The premise of replicability is that the scientific community can correct for these flaws.

But now all sorts of well-established, multiply confirmed findings have started to look increasingly uncertain. It's as if our facts were losing their

truth: claims that have been enshrined in textbooks are suddenly unprovable. This phenomenon doesn't yet have an official name, but it's occurring across a wide range of fields, from psychology to ecology. In the field of medicine, the phenomenon seems extremely widespread, affecting not only antipsychotics but also therapies ranging from cardiac stents to Vitamin E and antidepressants: Davis has a forthcoming analysis demonstrating that the efficacy of antidepressants has gone down as much as threefold in recent decades.

For many scientists, the effect is especially troubling because of what it exposes about the scientific process. If replication is what separates the rigor of science from the squishiness of pseudoscience, where do we put all these rigorously validated findings that can no longer be proved? Which results should we believe? Francis Bacon, the early-modern philosopher and pioneer of the scientific method, once declared that experiments were essential, because they allowed us to "put nature to the question." But it appears that nature often gives us different answers.

Jonathan Schooler was a young graduate student at the University of Washington in the nineteen-eighties when he discovered a surprising new fact about language and memory. At the time, it was widely believed that the act of describing our memories improved them. But, in a series of clever experiments, Schooler demonstrated that subjects shown a face and asked to describe it were much less likely to recognize the face when shown it later than those who had simply looked at it. Schooler called the phenomenon "verbal overshadowing."

The study turned him into an academic star. Since its initial publication, in 1990, it has been cited more than four hundred times. Before long, Schooler had extended the model to a variety of other tasks, such as remembering the taste of a wine, identifying the best strawberry jam, and solving difficult creative puzzles. In each instance, asking people to put their perceptions into words led to dramatic decreases in performance.

But while Schooler was publishing these results in highly reputable journals, a secret worry gnawed at him: it was proving difficult to replicate his earlier findings. "I'd often still see an effect, but the effect just wouldn't be as strong," he told me. "It was as if verbal overshadowing, my big new idea, was

getting weaker." At first, he assumed that he'd made an error in experimental design or a statistical miscalculation. But he couldn't find anything wrong with his research. He then concluded that his initial batch of research subjects must have been unusually susceptible to verbal overshadowing. (John Davis, similarly, has speculated that part of the dropoff in the effectiveness of antipsychotics can be attributed to using subjects who suffer from milder forms of psychosis which are less likely to show dramatic improvement.) "It wasn't a very satisfying explanation," Schooler says. "One of my mentors told me that my real mistake was trying to replicate my work. He told me doing that was just setting myself up for disappointment."

Schooler tried to put the problem out of his mind; his colleagues assured him that such things happened all the time. Over the next few years, he found new research questions, got married and had kids. But his replication problem kept on getting worse. His first attempt at replicating the 1990 study, in 1995, resulted in an effect that was thirty per cent smaller. The next year, the size of the effect shrank another thirty per cent. When other labs repeated Schooler's experiments, they got a similar spread of data, with a distinct downward trend. "This was profoundly frustrating," he says. "It was as if nature gave me this great result and then tried to take it back." In private, Schooler began referring to the problem as "cosmic habituation," by analogy to the decrease in response that occurs when individuals habituate to particular stimuli. "Habituation is why you don't notice the stuff that's always there," Schooler says. "It's an inevitable process of adjustment, a ratcheting down of excitement. I started joking that it was like the cosmos was habituating to my ideas. I took it very personally."

Schooler is now a tenured professor at the University of California at Santa Barbara. He has curly black hair, pale-green eyes, and the relaxed demeanor of someone who lives five minutes away from his favorite beach. When he speaks, he tends to get distracted by his own digressions. He might begin with a point about memory, which reminds him of a favorite William James quote, which inspires a long soliloquy on the importance of introspection. Before long, we're looking at pictures from Burning Man on his iPhone, which leads us back to the fragile nature of memory.

Although verbal overshadowing remains a widely accepted theory – it's

often invoked in the context of eyewitness testimony, for instance – Schooler is still a little peeved at the cosmos. "I know I should just move on already," he says. "I really should stop talking about this. But I can't." That's because he is convinced that he has stumbled on a serious problem, one that afflicts many of the most exciting new ideas in psychology.

One of the first demonstrations of this mysterious phenomenon came in the early nineteen-thirties. Joseph Banks Rhine, a psychologist at Duke, had developed an interest in the possibility of extrasensory perception, or E.S.P. Rhine devised an experiment featuring Zener cards, a special deck of twenty-five cards printed with one of five different symbols: a card was drawn from the deck and the subject was asked to guess the symbol. Most of Rhine's subjects guessed about twenty per cent of the cards correctly, as you'd expect, but an undergraduate named Adam Linzmayer averaged nearly fifty per cent during his initial sessions, and pulled off several uncanny streaks, such as guessing nine cards in a row. The odds of this happening by chance are about one in two million. Linzmayer did it three times.

Rhine documented these stunning results in his notebook and prepared several papers for publication. But then, just as he began to believe in the possibility of extrasensory perception, the student lost his spooky talent. Between 1931 and 1933, Linzmayer guessed at the identity of another several thousand cards, but his success rate was now barely above chance. Rhine was forced to conclude that the student's "extra-sensory perception ability has gone through a marked decline." And Linzmayer wasn't the only subject to experience such a drop-off: in nearly every case in which Rhine and others documented E.S.P. the effect dramatically diminished over time. Rhine called this trend the "decline effect."

Schooler was fascinated by Rhine's experimental struggles. Here was a scientist who had repeatedly documented the decline of his data; he seemed to have a talent for finding results that fell apart. In 2004, Schooler embarked on an ironic imitation of Rhine's research: he tried to replicate this failure to replicate. In homage to Rhine's interests, he decided to test for a para-psychological phenomenon known as precognition. The experiment itself was straightforward: he flashed a set of images to a subject and asked him or her to identify each one. Most of the time, the response was negative—the images

were displayed too quickly to register. Then Schooler randomly selected half of the images to be shown again. What he wanted to know was whether the images that got a second showing were more likely to have been identified the first time around. Could subsequent exposure have somehow influenced the initial results? Could the effect become the cause?

The craziness of the hypothesis was the point: Schooler knows that precognition lacks a scientific explanation. But he wasn't testing extrasensory powers; he was testing the decline effect. "At first, the data looked amazing, just as we'd expected," Schooler says. "I couldn't believe the amount of precognition we were finding. But then, as we kept on running subjects, the effect size" – a standard statistical measure – "kept on getting smaller and smaller." The scientists eventually tested more than two thousand undergraduates. "In the end, our results looked just like Rhine's," Schooler said. "We found this strong paranormal effect, but it disappeared on us."

The most likely explanation for the decline is an obvious one: regression to the mean. As the experiment is repeated, that is, an early statistical fluke gets canceled out. The extrasensory powers of Schooler's subjects didn't decline—they were simply an illusion that vanished over time. And yet Schooler has noticed that many of the data sets that end up declining seem statistically solid—that is, they contain enough data that any regression to the mean shouldn't be dramatic. "These are the results that pass all the tests," he says. "The odds of them being random are typically quite remote, like one in a million. This means that the decline effect should almost never happen. But it happens all the time! Hell, it's happened to me multiple times." And this is why Schooler believes that the decline effect deserves more attention: its ubiquity seems to violate the laws of statistics. "Whenever I start talking about this, scientists get very nervous," he says. "But I still want to know what happened to my results. Like most scientists, I assumed that it would get easier to document my effect over time. I'd get better at doing the experiments, at zeroing in on the conditions that produce verbal overshadowing. So why did the opposite happen? I'm convinced that we can use the tools of science to figure this out. First, though, we have to admit that we've got a problem."

In 1991, the Danish zoologist Anders Moller, at Uppsala University, in

Sweden, made a remarkable discovery about sex, barn swallows, and symmetry. It had long been known that the asymmetrical appearance of a creature was directly linked to the amount of mutation in its genome, so that more mutations led to more "fluctuating asymmetry." (An easy way to measure asymmetry in humans is to compare the length of the fingers on each hand.) What Moller discovered is that female barn swallows were far more likely to mate with male birds that had long, symmetrical feathers. This suggested that the picky females were using symmetry as a proxy for the quality of male genes. Moller's paper, which was published in *Nature*, set off a frenzy of research. Here was an easily measured, widely applicable indicator of genetic quality, and females could be shown to gravitate toward it. Aesthetics was really about genetics.

In the three years following, there were ten independent tests of the role of fluctuating asymmetry in sexual selection, and nine of them found a relationship between symmetry and male reproductive success. It didn't matter if scientists were looking at the hairs on fruit flies or replicating the swallow studies, females seemed to prefer males with mirrored halves. Before long, the theory was applied to humans. Researchers found, for instance, that women preferred the smell of symmetrical men, but only during the fertile phase of the menstrual cycle. Other studies claimed that females had more orgasms when their partners were symmetrical, while a paper by anthropologists at Rutgers analyzed forty Jamaican dance routines and discovered that symmetrical men were consistently rated as better dancers.

Then the theory started to fall apart. In 1994, there were fourteen published tests of symmetry and sexual selection, and only eight found a correlation. In 1995, there were eight papers on the subject, and only four got a positive result. By 1998, when there were twelve additional investigations of fluctuating asymmetry, only a third of them confirmed the theory. Worse still, even the studies that yielded some positive result showed a steadily declining effect size. Between 1992 and 1997, the average effect size shrank by eighty per cent.

And it's not just fluctuating asymmetry. In 2001, Michael Jennions, a biologist at the Australian National University, set out to analyze "temporal trends" across a wide range of subjects in ecology and evolutionary biol-

ogy. He looked at hundreds of papers and forty-four meta-analyses (that is, statistical syntheses of related studies), and discovered a consistent decline effect over time, as many of the theories seemed to fade into irrelevance. In fact, even when numerous variables were controlled for—Jennions knew, for instance, that the same author might publish several critical papers, which could distort his analysis—there was still a significant decrease in the validity of the hypothesis, often within a year of publication. Jennions admits that his findings are troubling, but expresses a reluctance to talk about them publicly. "This is a very sensitive issue for scientists," he says. "You know, we're supposed to be dealing with hard facts, the stuff that's supposed to stand the test of time. But when you see these trends you become a little more skeptical of things."

What happened? Leigh Simmons, a biologist at the University of Western Australia, suggested one explanation when he told me about his initial enthusiasm for the theory: "I was really excited by fluctuating asymmetry. The early studies made the effect look very robust." He decided to conduct a few experiments of his own, investigating symmetry in male horned beetles. "Unfortunately, I couldn't find the effect," he said. "But the worst part was that when I submitted these null results I had difficulty getting them published. The journals only wanted confirming data. It was too exciting an idea to disprove, at least back then." For Simmons, the steep rise and slow fall of fluctuating asymmetry is a clear example of a scientific paradigm, one of those intellectual fads that both guide and constrain research: after a new paradigm is proposed, the peer-review process is tilted toward positive results. But then, after a few years, the academic incentives shift.the paradigm has become entrenched—so that the most notable results are now those that disprove the theory.

Jennions, similarly, argues that the decline effect is largely a product of publication bias, or the tendency of scientists and scientific journals to prefer positive data over null results, which is what happens when no effect is found. The bias was first identified by the statistician Theodore Sterling, in 1959, after he noticed that ninety-seven per cent of all published psychological studies with statistically significant data found the effect they were looking for. A "significant" result is defined as any data point that would be produced

by chance less than five per cent of the time. This ubiquitous test was invented in 1922 by the English mathematician Ronald Fisher, who picked five per cent as the boundary line, somewhat arbitrarily, because it made pencil and slide-rule calculations easier. Sterling saw that if ninety-seven per cent of psychology studies were proving their hypotheses, either psychologists were extraordinarily lucky or they published only the outcomes of successful experiments. In recent years, publication bias has mostly been seen as a problem for clinical trials, since pharmaceutical companies are less interested in publishing results that aren't favorable. But it's becoming increasingly clear that publication bias also produces major distortions in fields without large corporate incentives, such as psychology and ecology.

While publication bias almost certainly plays a role in the decline effect, it remains an incomplete explanation. For one thing, it fails to account for the initial prevalence of positive results among studies that never even get submitted to journals. It also fails to explain the experience of people like Schooler, who have been unable to replicate their initial data despite their best efforts. Richard Palmer, a biologist at the University of Alberta, who has studied the problems surrounding fluctuating asymmetry, suspects that an equally significant issue is the selective reporting of results—the data that scientists choose to document in the first place. Palmer's most convincing evidence relies on a statistical tool known as a funnel graph. When a large number of studies have been done on a single subject, the data should follow a pattern: studies with a large sample size should all cluster around a common value—the true result—whereas those with a smaller sample size should exhibit a random scattering, since they're subject to greater sampling error. This pattern gives the graph its name, since the distribution resembles a funnel.

The funnel graph visually captures the distortions of selective reporting. For instance, after Palmer plotted every study of fluctuating asymmetry, he noticed that the distribution of results with smaller sample sizes wasn't random at all but instead skewed heavily toward positive results. Palmer has since documented a similar problem in several other contested subject areas. "Once I realized that selective reporting is everywhere in science, I got quite depressed," Palmer told me. "As a researcher, you're always

aware that there might be some nonrandom patterns, but I had no idea how widespread it is." In a recent review article, Palmer summarized the impact of selective reporting on his field: "We cannot escape the troubling conclusion that some—perhaps many—cherished generalities are at best exaggerated in their biological significance and at worst a collective illusion nurtured by strong a-priori beliefs often repeated."

Palmer emphasizes that selective reporting is not the same as scientific fraud. Rather, the problem seems to be one of subtle omissions and unconscious misperceptions, as researchers struggle to make sense of their results. Stephen Jay Gould referred to this as the "shoehorning" process. "A lot of scientific measurement is really hard," Simmons told me. "If you're talking about fluctuating asymmetry, then it's a matter of minuscule differences between the right and left sides of an animal. It's millimetres of a tail feather. And so maybe a researcher knows that he's measuring a good male"—an animal that has successfully mated."and he knows that it's supposed to be symmetrical. Well, that act of measurement is going to be vulnerable to all sorts of perception biases. That's not a cynical statement. That's just the way human beings work."

One of the classic examples of selective reporting concerns the testing of acupuncture in different countries. While acupuncture is widely accepted as a medical treatment in various Asian countries, its use is much more contested in the West. These cultural differences have profoundly influenced the results of clinical trials. Between 1966 and 1995, there were fortyseven studies of acupuncture in China, Taiwan, and Japan, and every single trial concluded that acupuncture was an effective treatment. During the same period, there were ninety-four clinical trials of acupuncture in the United States, Sweden, and the U.K., and only fifty-six per cent of these studies found any therapeutic benefits. As Palmer notes, this wide discrepancy suggests that scientists find ways to confirm their preferred hypothesis, disregarding what they don't want to see. Our beliefs are a form of blindness.

John Ioannidis, an epidemiologist at Stanford University, argues that such distortions are a serious issue in biomedical research. "These exaggerations are why the decline has become so common," he says. "It'd be really great if the initial studies gave us an accurate summary of things. But they don't.

And so what happens is we waste a lot of money treating millions of patients and doing lots of follow-up studies on other themes based on results that are misleading." In 2005, Ioannidis published an article in the *Journal of the American Medical Association* that looked at the forty-nine most cited clinical-research studies in three major medical journals. Forty-five of these studies reported positive results, suggesting that the intervention being tested was effective. Because most of these studies were randomized controlled trials—the "gold standard" of medical evidence—they tended to have a significant impact on clinical practice, and led to the spread of treatments such as hormone replacement therapy for menopausal women and daily low-dose aspirin to prevent heart attacks and strokes. Nevertheless, the data Ioannidis found were disturbing: of the thirty-four claims that had been subject to replication, forty-one per cent had either been directly contradicted or had their effect sizes significantly downgraded.

The situation is even worse when a subject is fashionable. In recent years, for instance, there have been hundreds of studies on the various genes that control the differences in disease risk between men and women. These findings have included everything from the mutations responsible for the increased risk of schizophrenia to the genes underlying hypertension. Ioannidis and his colleagues looked at four hundred and thirty-two of these claims. They quickly discovered that the vast majority had serious flaws. But the most troubling fact emerged when he looked at the test of replication: out of four hundred and thirty-two claims, only a single one was consistently replicable. "This doesn't mean that none of these claims will turn out to be true," he says. "But, given that most of them were done badly, I wouldn't hold my breath."

According to Ioannidis, the main problem is that too many researchers engage in what he calls "significance chasing," or finding ways to interpret the data so that it passes the statistical test of significance—the ninety-five-per-cent boundary invented by Ronald Fisher. "The scientists are so eager to pass this magical test that they start playing around with the numbers, trying to find anything that seems worthy," Ioannidis says. In recent years, Ioannidis has become increasingly blunt about the pervasiveness of the problem. One of his most cited papers has a deliberately provocative title: "Why Most

Published Research Findings Are False."

The problem of selective reporting is rooted in a fundamental cognitive flaw, which is that we like proving ourselves right and hate being wrong. "It feels good to validate a hypothesis," Ioannidis said. "It feels even better when you've got a financial interest in the idea or your career depends upon it. And that's why, even after a claim has been systematically disproven"—he cites, for instance, the early work on hormone replacement therapy, or claims involving various vitamins."you still see some stubborn researchers citing the first few studies that show a strong effect. They really want to believe that it's true."

That's why Schooler argues that scientists need to become more rigorous about data collection before they publish. "We're wasting too much time chasing after bad studies and underpowered experiments," he says. The current "obsession" with replicability distracts from the real problem, which is faulty design. He notes that nobody even tries to replicate most science papers—there are simply too many. (According to *Nature*, a third of all studies never even get cited, let alone repeated.) "I've learned the hard way to be exceedingly careful," Schooler says. "Every researcher should have to spell out, in advance, how many subjects they're going to use, and what exactly they're testing, and what constitutes a sufficient level of proof. We have the tools to be much more transparent about our experiments."

In a forthcoming paper, Schooler recommends the establishment of an open-source database, in which researchers are required to outline their planned investigations and document all their results. "I think this would provide a huge increase in access to scientific work and give us a much better way to judge the quality of an experiment," Schooler says. "It would help us finally deal with all these issues that the decline effect is exposing."

Although such reforms would mitigate the dangers of publication bias and selective reporting, they still wouldn't erase the decline effect. This is largely because scientific research will always be shadowed by a force that can't be curbed, only contained: sheer randomness. Although little research has been done on the experimental dangers of chance and happenstance, the research that exists isn't encouraging.

In the late nineteen-nineties, John Crabbe, a neuroscientist at the Oregon

Health and Science University, conducted an experiment that showed how unknowable chance events can skew tests of replicability. He performed a series of experiments on mouse behavior in three different science labs: in Albany, New York; Edmonton, Alberta; and Portland, Oregon. Before he conducted the experiments, he tried to standardize every variable he could think of. The same strains of mice were used in each lab, shipped on the same day from the same supplier. The animals were raised in the same kind of enclosure, with the same brand of sawdust bedding. They had been exposed to the same amount of incandescent light, were living with the same number of littermates, and were fed the exact same type of chow pellets. When the mice were handled, it was with the same kind of surgical glove, and when they were tested it was on the same equipment, at the same time in the morning.

The premise of this test of replicability, of course, is that each of the labs should have generated the same pattern of results. "If any set of experiments should have passed the test, it should have been ours," Crabbe says. "But that's not the way it turned out." In one experiment, Crabbe injected a particular strain of mouse with cocaine. In Portland the mice given the drug moved, on average, six hundred centimetres more than they normally did; in Albany they moved seven hundred and one additional centimetres. But in the Edmonton lab they moved more than five thousand additional centimetres. Similar deviations were observed in a test of anxiety. Furthermore, these inconsistencies didn't follow any detectable pattern. In Portland one strain of mouse proved most anxious, while in Albany another strain won that distinction.

The disturbing implication of the Crabbe study is that a lot of extraordinary scientific data are nothing but noise. The hyperactivity of those coked-up Edmonton mice wasn't an interesting new fact—it was a meaningless outlier, a by-product of invisible variables we don't understand. The problem, of course, is that such dramatic findings are also the most likely to get published in prestigious journals, since the data are both statistically significant and entirely unexpected. Grants get written, follow-up studies are conducted. The end result is a scientific accident that can take years to unravel.

This suggests that the decline effect is actually a decline of illusion. While

Karl Popper imagined falsification occurring with a single, definitive experiment—Galileo refuted Aristotelian mechanics in an afternoon—the process turns out to be much messier than that. Many scientific theories continue to be considered true even after failing numerous experimental tests. Verbal overshadowing might exhibit the decline effect, but it remains extensively relied upon within the field. The same holds for any number of phenomena, from the disappearing benefits of second generation antipsychotics to the weak coupling ratio exhibited by decaying neutrons, which appears to have fallen by more than ten standard deviations between 1969 and 2001. Even the law of gravity hasn't always been perfect at predicting real-world phenomena. (In one test, physicists measuring gravity by means of deep boreholes in the Nevada desert found a two-and-a-half-per-cent discrepancy between the theoretical predictions and the actual data.) Despite these findings, second-generation antipsychotics are still widely prescribed, and our model of the neutron hasn't changed. The law of gravity remains the same.

Such anomalies demonstrate the slipperiness of empiricism. Although many scientific ideas generate conflicting results and suffer from falling effect sizes, they continue to get cited in the textbooks and drive standard medical practice. Why? Because these ideas seem true. Because they make sense. Because we can't bear to let them go. And this is why the decline effect is so troubling. Not because it reveals the human fallibility of science, in which data are tweaked and beliefs shape perceptions. (Such shortcomings aren't surprising, at least for scientists.) And not because it reveals that many of our most exciting theories are fleeting fads and will soon be rejected. (That idea has been around since Thomas Kuhn.) The decline effect is troubling because it reminds us how difficult it is to prove anything. We like to pretend that our experiments define the truth for us. But that's often not the case. Just because an idea is true doesn't mean it can be proved. And just because an idea can be proved doesn't mean it's true. When the experiments are done, we still have to choose what to believe.

———————-

Postscript
January 3, 2011
More Thoughts on the Decline Effect

Posted by Jonah Lehrer

In "The Truth Wears Off," I wanted to explore the human side of the scientific enterprise. My focus was on a troubling phenomenon often referred to as the "decline effect," which is the tendency of many exciting scientific results to fade over time. This empirical hiccup afflicts fields from pharmacology to evolutionary biology to social psychology. There is no simple explanation for the decline effect, but the article explores several possibilities, from the publication biases of peer-reviewed journals to the "selective reporting" of scientists who sift through data.

This week, the magazine published four very thoughtful letters in response to the piece. The first letter, like many of the e-mails, tweets, and comments I've received directly, argues that the decline effect is ultimately a minor worry, since "in the long run, science prevails over human bias." The letter, from Howard Stuart, cites the famous 1909 oil-drop experiment performed by Robert Millikan and Harvey Fletcher, which sought to measure the charge of the electron. It's a fascinating experimental tale, as subsequent measurements gradually corrected the data, steadily nudging the charge upwards. In his 1974 commencement address at Caltech, Richard Feynman described why the initial measurement was off, and why it took so long to fix:

Millikan measured the charge on an electron by an experiment with falling oil drops, and got an answer which we now know not to be quite right. It's a little bit off, because he had the incorrect value for the viscosity of air. It's interesting to look at the history of measurements of the charge of the electron, after Millikan. If you plot them as a function of time, you find that one is a little bigger than Millikan's, and the next one's a little bit bigger than that, and the next one's a little bit bigger than that, until finally they settle down to a number which is higher.

Why didn't they discover that the new number was higher right away? It's a thing that scientists are ashamed of—this history—because it's apparent that people did things like this: When they got a number that was too high above Millikan's, they thought something must be wrong—and they would look for and find a reason why something might be wrong. When they got a number closer to Millikan's value they didn't look so hard. And so they eliminated the numbers that were too far off, and did other things like that.

That's a pretty perfect example of selective reporting in science. One optimistic takeaway from the oil-drop experiment is that our errors get corrected, and that the truth will always win out. Like Mr. Stuart, this was the moral Feynman preferred, as he warned the Caltech undergrads to be rigorous scientists, because their lack of rigor would be quickly exposed by the scientific process. "Other experimenters will repeat your experiment and find out whether you were wrong or right," Feynman said. "Nature's phenomena will agree or they'll disagree with your theory." But that's not always the case. For one thing, a third of scientific papers never get cited, let alone repeated, which means that many errors are never exposed. But even those theories that do get replicated are shadowed by uncertainty. After all, one of the more disturbing aspects of the decline effect is that many results we now believe to be false have been replicated numerous times. To take but one example I cited in the article: After fluctuating asymmetry, a widely publicized theory in evolutionary biology, was proposed in the early nineteen-nineties, nine of the first ten independent tests confirmed the theory. In fact, it took several years before an overwhelming majority of published papers began rejecting it. This raises the obvious problem: If false results can get replicated, then how do we demarcate science from pseudoscience? And how can we be sure that anything—even a multiply confirmed finding—is true?

These questions have no easy answers. However, I think the decline effect is an important reminder that we shouldn't simply reassure ourselves with platitudes about the rigors of replication or the inevitable corrections of peer review. Although we often pretend that experiments settle the truth for us—that we are mere passive observers, dutifully recording the facts—the reality of science is a lot messier. It is an intensely human process, shaped by all of our usual talents, tendencies, and flaws.

Many letters chastised me for critiquing science in such a public venue. Here's an example, from Dr. Robert Johnson of Wayne State Medical School:

Creationism and skepticism of climate change are popularly-held opinions; Lehrer's closing words play into the hands of those who want to deny evolution, global warming, and other realities. I fear that those who wish to persuade Americans that science is just one more pressure group, and that the scientific method is a matter of opinion, will be eager to use his conclusion

to advance their cause.

This was a concern I wrestled with while writing the piece. One of the sad ironies of scientific denialism is that we tend to be skeptical of precisely the wrong kind of scientific claims. Natural selection and climate change have been verified in thousands of different ways by thousands of different scientists working in many different fields. (This doesn't mean, of course, that such theories won't change or get modified—the strength of science is that nothing is settled.) Instead of wasting public debate on solid theories, I wish we'd spend more time considering the value of second-generation antipsychotics or the verity of the latest gene-association study.

Nevertheless, I think the institutions and mechanisms of the scientific process demand investigation, even if the inside view isn't flattering. We know science works. But can it work better? There is too much at stake to not ask that question. Furthermore, the public funds a vast majority of basic research.it deserves to know about any problems.

And this brings me to another category of letters, which proposed new ways of minimizing the decline effect. Some readers suggested reducing the acceptable level of $p$-values or starting a *Journal of Negative Results*. Andrew Gelman, a professor of statistics at Columbia University, proposed the use of "retrospective power analyses," in which experimenters are forced to calculate their effect size using "real prior information," and not just the data distilled from their small sample size.

### 0.0.18 Lies, Damned Lies, and Medical Science, David H. Freedman (*The Atlantic*), November 2010

Much of what medical researchers conclude in their studies is misleading, exaggerated, or flat-out wrong. So why are doctors—to a striking extent—still drawing upon misinformation in their everyday practice? Dr. John Ioannidis has spent his career challenging his peers by exposing their bad science.

In 2001, rumors were circulating in Greek hospitals that surgery residents, eager to rack up scalpel time, were falsely diagnosing hapless Albanian immigrants with appendicitis. At the University of Ioannina medical school's teaching hospital, a newly minted doctor named Athina Tatsioni was discussing the rumors with colleagues when a professor who had overheard asked her if she'd like to try to prove whether they were true—he seemed to be almost daring her. She accepted the challenge and, with the professor's and other colleagues' help, eventually produced a formal study showing that, for whatever reason, the appendices removed from patients with Albanian names in six Greek hospitals were more than three times as likely to be perfectly healthy as those removed from patients with Greek names. "It was hard to find a journal willing to publish it, but we did," recalls Tatsioni. "I also discovered that I really liked research." Good thing, because the study had actually been a sort of audition. The professor, it turned out, had been putting together a team of exceptionally brash and curious young clinicians and Ph.D.s to join him in tackling an unusual and controversial agenda.

Last spring, I sat in on one of the team's weekly meetings on the medical school's campus, which is plunked crazily across a series of sharp hills. The building in which we met, like most at the school, had the look of a barracks and was festooned with political graffiti. But the group convened in a spacious conference room that would have been at home at a Silicon Valley start-up. Sprawled around a large table were Tatsioni and eight other youngish Greek researchers and physicians who, in contrast to the pasty younger staff frequently seen in U.S. hospitals, looked like the casually glamorous cast of a television medical drama. The professor, a dapper and soft-spoken man named John Ioannidis, loosely presided.

One of the researchers, a biostatistician named Georgia Salanti, fired up

a laptop and projector and started to take the group through a study she and a few colleagues were completing that asked this question: were drug companies manipulating published research to make their drugs look good? Salanti ticked off data that seemed to indicate they were, but the other team members almost immediately started interrupting. One noted that Salanti's study didn't address the fact that drug-company research wasn't measuring critically important "hard" outcomes for patients, such as survival versus death, and instead tended to measure "softer" outcomes, such as self-reported symptoms ("my chest doesn't hurt as much today"). Another pointed out that Salanti's study ignored the fact that when drug-company data seemed to show patients' health improving, the data often failed to show that the drug was responsible, or that the improvement was more than marginal.

Salanti remained poised, as if the grilling were par for the course, and gamely acknowledged that the suggestions were all good—but a single study can't prove everything, she said. Just as I was getting the sense that the data in drug studies were endlessly malleable, Ioannidis, who had mostly been listening, delivered what felt like a coup de grâce: wasn't it possible, he asked, that drug companies were carefully selecting the topics of their studies—for example, comparing their new drugs against those already known to be inferior to others on the market—so that they were ahead of the game even before the data juggling began? "Maybe sometimes it's the questions that are biased, not the answers," he said, flashing a friendly smile. Everyone nodded. Though the results of drug studies often make newspaper headlines, you have to wonder whether they prove anything at all. Indeed, given the breadth of the potential problems raised at the meeting, can any medical-research studies be trusted?

That question has been central to Ioannidis's career. He's what's known as a meta-researcher, and he's become one of the world's foremost experts on the credibility of medical research. He and his team have shown, again and again, and in many different ways, that much of what biomedical researchers conclude in published studies—conclusions that doctors keep in mind when they prescribe antibiotics or blood-pressure medication, or when they advise us to consume more fiber or less meat, or when they recommend surgery for heart disease or back pain—is misleading, exaggerated, and often flat-

out wrong. He charges that as much as 90 percent of the published medical information that doctors rely on is flawed. His work has been widely accepted by the medical community; it has been published in the field's top journals, where it is heavily cited; and he is a big draw at conferences. Given this exposure, and the fact that his work broadly targets everyone else's work in medicine, as well as everything that physicians do and all the health advice we get, Ioannidis may be one of the most influential scientists alive. Yet for all his influence, he worries that the field of medical research is so pervasively flawed, and so riddled with conflicts of interest, that it might be chronically resistant to change—or even to publicly admitting that there's a problem.

The city of Ioannina is a big college town a short drive from the ruins of a 20,000-seat amphitheater and a Zeusian sanctuary built at the site of the Dodona oracle. The oracle was said to have issued pronouncements to priests through the rustling of a sacred oak tree. Today, a different oak tree at the site provides visitors with a chance to try their own hands at extracting a prophecy. "I take all the researchers who visit me here, and almost every single one of them asks the tree the same question," Ioannidis tells me, as we contemplate the tree the day after the team's meeting. " 'Will my research grant be approved?" ' He chuckles, but Ioannidis (pronounced yo-NEE-dees) tends to laugh not so much in mirth as to soften the sting of his attack. And sure enough, he goes on to suggest that an obsession with winning funding has gone a long way toward weakening the reliability of medical research.

He first stumbled on the sorts of problems plaguing the field, he explains, as a young physician-researcher in the early 1990s at Harvard. At the time, he was interested in diagnosing rare diseases, for which a lack of case data can leave doctors with little to go on other than intuition and rules of thumb. But he noticed that doctors seemed to proceed in much the same manner even when it came to cancer, heart disease, and other common ailments. Where were the hard data that would back up their treatment decisions? There was plenty of published research, but much of it was remarkably unscientific, based largely on observations of a small number of cases. A new "evidence-based medicine" movement was just starting to gather force, and Ioannidis decided to throw himself into it, working first with prominent researchers at Tufts University and then taking positions at Johns Hopkins University

and the National Institutes of Health. He was unusually well armed: he had been a math prodigy of near-celebrity status in high school in Greece, and had followed his parents, who were both physician-researchers, into medicine. Now he'd have a chance to combine math and medicine by applying rigorous statistical analysis to what seemed a surprisingly sloppy field. "I assumed that everything we physicians did was basically right, but now I was going to help verify it," he says. "All we'd have to do was systematically review the evidence, trust what it told us, and then everything would be perfect."

It didn't turn out that way. In poring over medical journals, he was struck by how many findings of all types were refuted by later findings. Of course, medical-science "never minds" are hardly secret. And they sometimes make headlines, as when in recent years large studies or growing consensuses of researchers concluded that mammograms, colonoscopies, and PSA tests are far less useful cancer-detection tools than we had been told; or when widely prescribed antidepressants such as Prozac, Zoloft, and Paxil were revealed to be no more effective than a placebo for most cases of depression; or when we learned that staying out of the sun entirely can actually increase cancer risks; or when we were told that the advice to drink lots of water during intense exercise was potentially fatal; or when, last April, we were informed that taking fish oil, exercising, and doing puzzles doesn't really help fend off Alzheimer's disease, as long claimed. Peer-reviewed studies have come to opposite conclusions on whether using cell phones can cause brain cancer, whether sleeping more than eight hours a night is healthful or dangerous, whether taking aspirin every day is more likely to save your life or cut it short, and whether routine angioplasty works better than pills to unclog heart arteries.

But beyond the headlines, Ioannidis was shocked at the range and reach of the reversals he was seeing in everyday medical research. "Randomized controlled trials," which compare how one group responds to a treatment against how an identical group fares without the treatment, had long been considered nearly unshakable evidence, but they, too, ended up being wrong some of the time. "I realized even our gold-standard research had a lot of problems," he says. Baffled, he started looking for the specific ways in which studies were going wrong. And before long he discovered that the range

of errors being committed was astonishing: from what questions researchers posed, to how they set up the studies, to which patients they recruited for the studies, to which measurements they took, to how they analyzed the data, to how they presented their results, to how particular studies came to be published in medical journals.

This array suggested a bigger, underlying dysfunction, and Ioannidis thought he knew what it was. "The studies were biased," he says. "Sometimes they were overtly biased. Sometimes it was difficult to see the bias, but it was there." Researchers headed into their studies wanting certain results—and, lo and behold, they were getting them. We think of the scientific process as being objective, rigorous, and even ruthless in separating out what is true from what we merely wish to be true, but in fact it's easy to manipulate results, even unintentionally or unconsciously. "At every step in the process, there is room to distort results, a way to make a stronger claim or to select what is going to be concluded," says Ioannidis. "There is an intellectual conflict of interest that pressures researchers to find whatever it is that is most likely to get them funded."

Perhaps only a minority of researchers were succumbing to this bias, but their distorted findings were having an outsize effect on published research. To get funding and tenured positions, and often merely to stay afloat, researchers have to get their work published in well-regarded journals, where rejection rates can climb above 90 percent. Not surprisingly, the studies that tend to make the grade are those with eye-catching findings. But while coming up with eye-catching theories is relatively easy, getting reality to bear them out is another matter. The great majority collapse under the weight of contradictory data when studied rigorously. Imagine, though, that five different research teams test an interesting theory that's making the rounds, and four of the groups correctly prove the idea false, while the one less cautious group incorrectly "proves" it true through some combination of error, fluke, and clever selection of data. Guess whose findings your doctor ends up reading about in the journal, and you end up hearing about on the evening news? Researchers can sometimes win attention by refuting a prominent finding, which can help to at least raise doubts about results, but in general it is far more rewarding to add a new insight or exciting-sounding twist to existing re-

search than to retest its basic premises—after all, simply re-proving someone else's results is unlikely to get you published, and attempting to undermine the work of respected colleagues can have ugly professional repercussions.

In the late 1990s, Ioannidis set up a base at the University of Ioannina. He pulled together his team, which remains largely intact today, and started chipping away at the problem in a series of papers that pointed out specific ways certain studies were getting misleading results. Other meta-researchers were also starting to spotlight disturbingly high rates of error in the medical literature. But Ioannidis wanted to get the big picture across, and to do so with solid data, clear reasoning, and good statistical analysis. The project dragged on, until finally he retreated to the tiny island of Sikinos in the Aegean Sea, where he drew inspiration from the relatively primitive surroundings and the intellectual traditions they recalled. "A pervasive theme of ancient Greek literature is that you need to pursue the truth, no matter what the truth might be," he says. In 2005, he unleashed two papers that challenged the foundations of medical research.

He chose to publish one paper, fittingly, in the online journal PLoS Medicine, which is committed to running any methodologically sound article without regard to how "interesting" the results may be. In the paper, Ioannidis laid out a detailed mathematical proof that, assuming modest levels of researcher bias, typically imperfect research techniques, and the well-known tendency to focus on exciting rather than highly plausible theories, researchers will come up with wrong findings most of the time. Simply put, if you're attracted to ideas that have a good chance of being wrong, and if you're motivated to prove them right, and if you have a little wiggle room in how you assemble the evidence, you'll probably succeed in proving wrong theories right. His model predicted, in different fields of medical research, rates of wrongness roughly corresponding to the observed rates at which findings were later convincingly refuted: 80 percent of non-randomized studies (by far the most common type) turn out to be wrong, as do 25 percent of supposedly gold-standard randomized trials, and as much as 10 percent of the platinum-standard large randomized trials. The article spelled out his belief that researchers were frequently manipulating data analyses, chasing career-advancing findings rather than good science, and even using the peer-review process—in which journals

ask researchers to help decide which studies to publish—to suppress opposing views. "You can question some of the details of John's calculations, but it's hard to argue that the essential ideas aren't absolutely correct," says Doug Altman, an Oxford University researcher who directs the Centre for Statistics in Medicine.

Still, Ioannidis anticipated that the community might shrug off his findings: sure, a lot of dubious research makes it into journals, but we researchers and physicians know to ignore it and focus on the good stuff, so what's the big deal? The other paper headed off that claim. He zoomed in on 49 of the most highly regarded research findings in medicine over the previous 13 years, as judged by the science community's two standard measures: the papers had appeared in the journals most widely cited in research articles, and the 49 articles themselves were the most widely cited articles in these journals. These were articles that helped lead to the widespread popularity of treatments such as the use of hormone-replacement therapy for menopausal women, vitamin E to reduce the risk of heart disease, coronary stents to ward off heart attacks, and daily low-dose aspirin to control blood pressure and prevent heart attacks and strokes. Ioannidis was putting his contentions to the test not against run-of-the-mill research, or even merely well-accepted research, but against the absolute tip of the research pyramid. Of the 49 articles, 45 claimed to have uncovered effective interventions. Thirty-four of these claims had been retested, and 14 of these, or 41 percent, had been convincingly shown to be wrong or significantly exaggerated. If between a third and a half of the most acclaimed research in medicine was proving untrustworthy, the scope and impact of the problem were undeniable. That article was published in the *Journal of the American Medical Association.*

Driving me back to campus in his smallish SUV—after insisting, as he apparently does with all his visitors, on showing me a nearby lake and the six monasteries situated on an islet within it—Ioannidis apologized profusely for running a yellow light, explaining with a laugh that he didn't trust the truck behind him to stop. Considering his willingness, even eagerness, to slap the face of the medical-research community, Ioannidis comes off as thoughtful, upbeat, and deeply civil. He's a careful listener, and his frequent grin and semi-apologetic chuckle can make the sharp prodding of his arguments seem

almost good-natured. He is as quick, if not quicker, to question his own motives and competence as anyone else's. A neat and compact 45-year-old with a trim mustache, he presents as a sort of dashing nerd—Giancarlo Giannini with a bit of Mr. Bean.

The humility and graciousness seem to serve him well in getting across a message that is not easy to digest or, for that matter, believe: that even highly regarded researchers at prestigious institutions sometimes churn out attention-grabbing findings rather than findings likely to be right. But Ioannidis points out that obviously questionable findings cram the pages of top medical journals, not to mention the morning headlines. Consider, he says, the endless stream of results from nutritional studies in which researchers follow thousands of people for some number of years, tracking what they eat and what supplements they take, and how their health changes over the course of the study. "Then the researchers start asking, 'What did vitamin E do? What did vitamin C or D or A do? What changed with calorie intake, or protein or fat intake? What happened to cholesterol levels? Who got what type of cancer?" ' he says. "They run everything through the mill, one at a time, and they start finding associations, and eventually conclude that vitamin X lowers the risk of cancer Y, or this food helps with the risk of that disease." In a single week this fall, Google's news page offered these headlines: "More Omega-3 Fats Didn't Aid Heart Patients"; "Fruits, Vegetables Cut Cancer Risk for Smokers"; "Soy May Ease Sleep Problems in Older Women"; and dozens of similar stories.

When a five-year study of 10,000 people finds that those who take more vitamin X are less likely to get cancer Y, you'd think you have pretty good reason to take more vitamin X, and physicians routinely pass these recommendations on to patients. But these studies often sharply conflict with one another. Studies have gone back and forth on the cancer-preventing powers of vitamins A, D, and E; on the heart-health benefits of eating fat and carbs; and even on the question of whether being overweight is more likely to extend or shorten your life. How should we choose among these dueling, high-profile nutritional findings? Ioannidis suggests a simple approach: ignore them all.

For starters, he explains, the odds are that in any large database of many nutritional and health factors, there will be a few apparent connections that

are in fact merely flukes, not real health effects—it's a bit like combing through long, random strings of letters and claiming there's an important message in any words that happen to turn up. But even if a study managed to highlight a genuine health connection to some nutrient, you're unlikely to benefit much from taking more of it, because we consume thousands of nutrients that act together as a sort of network, and changing intake of just one of them is bound to cause ripples throughout the network that are far too complex for these studies to detect, and that may be as likely to harm you as help you. Even if changing that one factor does bring on the claimed improvement, there's still a good chance that it won't do you much good in the long run, because these studies rarely go on long enough to track the decades-long course of disease and ultimately death. Instead, they track easily measurable health "markers" such as cholesterol levels, blood pressure, and blood-sugar levels, and meta-experts have shown that changes in these markers often don't correlate as well with long-term health as we have been led to believe.

On the relatively rare occasions when a study does go on long enough to track mortality, the findings frequently upend those of the shorter studies. (For example, though the vast majority of studies of overweight individuals link excess weight to ill health, the longest of them haven't convincingly shown that overweight people are likely to die sooner, and a few of them have seemingly demonstrated that moderately overweight people are likely to live longer.) And these problems are aside from ubiquitous measurement errors (for example, people habitually misreport their diets in studies), routine misanalysis (researchers rely on complex software capable of juggling results in ways they don't always understand), and the less common, but serious, problem of outright fraud (which has been revealed, in confidential surveys, to be much more widespread than scientists like to acknowledge).

If a study somehow avoids every one of these problems and finds a real connection to long-term changes in health, you're still not guaranteed to benefit, because studies report average results that typically represent a vast range of individual outcomes. Should you be among the lucky minority that stands to benefit, don't expect a noticeable improvement in your health, because studies usually detect only modest effects that merely tend to whittle

your chances of succumbing to a particular disease from small to somewhat smaller. "The odds that anything useful will survive from any of these studies are poor," says Ioannidis—dismissing in a breath a good chunk of the research into which we sink about $100 billion a year in the United States alone.

And so it goes for all medical studies, he says. Indeed, nutritional studies aren't the worst. Drug studies have the added corruptive force of financial conflict of interest. The exciting links between genes and various diseases and traits that are relentlessly hyped in the press for heralding miraculous around-the-corner treatments for everything from colon cancer to schizophrenia have in the past proved so vulnerable to error and distortion, Ioannidis has found, that in some cases you'd have done about as well by throwing darts at a chart of the genome. (These studies seem to have improved somewhat in recent years, but whether they will hold up or be useful in treatment are still open questions.) Vioxx, Zelnorm, and Baycol were among the widely prescribed drugs found to be safe and effective in large randomized controlled trials before the drugs were yanked from the market as unsafe or not so effective, or both.

"Often the claims made by studies are so extravagant that you can immediately cross them out without needing to know much about the specific problems with the studies," Ioannidis says. But of course it's that very extravagance of claim (one large randomized controlled trial even proved that secret prayer by unknown parties can save the lives of heart-surgery patients, while another proved that secret prayer can harm them) that helps gets these findings into journals and then into our treatments and lifestyles, especially when the claim builds on impressive-sounding evidence. "Even when the evidence shows that a particular research idea is wrong, if you have thousands of scientists who have invested their careers in it, they'll continue to publish papers on it," he says. "It's like an epidemic, in the sense that they're infected with these wrong ideas, and they're spreading it to other researchers through journals."

Though scientists and science journalists are constantly talking up the value of the peer-review process, researchers admit among themselves that biased, erroneous, and even blatantly fraudulent studies easily slip through it. *Nature*, the grande dame of science journals, stated in a 2006 editorial,

"Scientists understand that peer review per se provides only a minimal assurance of quality, and that the public conception of peer review as a stamp of authentication is far from the truth." What's more, the peer-review process often pressures researchers to shy away from striking out in genuinely new directions, and instead to build on the findings of their colleagues (that is, their potential reviewers) in ways that only seem like breakthroughs—as with the exciting-sounding gene linkages (autism genes identified!) and nutritional findings (olive oil lowers blood pressure!) that are really just dubious and conflicting variations on a theme.

Most journal editors don't even claim to protect against the problems that plague these studies. University and government research overseers rarely step in to directly enforce research quality, and when they do, the science community goes ballistic over the outside interference. The ultimate protection against research error and bias is supposed to come from the way scientists constantly retest each other's results—except they don't. Only the most prominent findings are likely to be put to the test, because there's likely to be publication payoff in firming up the proof, or contradicting it.

But even for medicine's most influential studies, the evidence sometimes remains surprisingly narrow. Of those 45 super-cited studies that Ioannidis focused on, 11 had never been retested. Perhaps worse, Ioannidis found that even when a research error is outed, it typically persists for years or even decades. He looked at three prominent health studies from the 1980s and 1990s that were each later soundly refuted, and discovered that researchers continued to cite the original results as correct more often than as flawed—in one case for at least 12 years after the results were discredited.

Doctors may notice that their patients don't seem to fare as well with certain treatments as the literature would lead them to expect, but the field is appropriately conditioned to subjugate such anecdotal evidence to study findings. Yet much, perhaps even most, of what doctors do has never been formally put to the test in credible studies, given that the need to do so became obvious to the field only in the 1990s, leaving it playing catch-up with a century or more of non-evidence-based medicine, and contributing to Ioannidis's shockingly high estimate of the degree to which medical knowledge is flawed. That we're not routinely made seriously ill by this shortfall, he

argues, is due largely to the fact that most medical interventions and advice don't address life-and-death situations, but rather aim to leave us marginally healthier or less unhealthy, so we usually neither gain nor risk all that much.

Medical research is not especially plagued with wrongness. Other meta-research experts have confirmed that similar issues distort research in all fields of science, from physics to economics (where the highly regarded economists J. Bradford DeLong and Kevin Lang once showed how a remarkably consistent paucity of strong evidence in published economics studies made it unlikely that any of them were right). And needless to say, things only get worse when it comes to the pop expertise that endlessly spews at us from diet, relationship, investment, and parenting gurus and pundits. But we expect more of scientists, and especially of medical scientists, given that we believe we are staking our lives on their results. The public hardly recognizes how bad a bet this is. The medical community itself might still be largely oblivious to the scope of the problem, if Ioannidis hadn't forced a confrontation when he published his studies in 2005.

Ioannidis initially thought the community might come out fighting. Instead, it seemed relieved, as if it had been guiltily waiting for someone to blow the whistle, and eager to hear more. David Gorski, a surgeon and researcher at Detroit's Barbara Ann Karmanos Cancer Institute, noted in his prominent medical blog that when he presented Ioannidis's paper on highly cited research at a professional meeting, "not a single one of my surgical colleagues was the least bit surprised or disturbed by its findings." Ioannidis offers a theory for the relatively calm reception. "I think that people didn't feel I was only trying to provoke them, because I showed that it was a community problem, instead of pointing fingers at individual examples of bad research," he says. In a sense, he gave scientists an opportunity to cluck about the wrongness without having to acknowledge that they themselves succumb to it—it was something everyone else did.

To say that Ioannidis's work has been embraced would be an understatement. His PLoS Medicine paper is the most downloaded in the journal's history, and it's not even Ioannidis's most-cited work—that would be a paper he published in *Nature Genetics* on the problems with gene-link studies. Other researchers are eager to work with him: he has published papers with

1,328 different co-authors at 538 institutions in 43 countries, he says. Last year he received, by his estimate, invitations to speak at 1,000 conferences and institutions around the world, and he was accepting an average of about five invitations a month until a case last year of excessive-travel-induced vertigo led him to cut back. Even so, in the weeks before I visited him he had addressed an AIDS conference in San Francisco, the European Society for Clinical Investigation, Harvard's School of Public Health, and the medical schools at Stanford and Tufts.

The irony of his having achieved this sort of success by accusing the medical-research community of chasing after success is not lost on him, and he notes that it ought to raise the question of whether he himself might be pumping up his findings. "If I did a study and the results showed that in fact there wasn't really much bias in research, would I be willing to publish it?" he asks. "That would create a real psychological conflict for me." But his bigger worry, he says, is that while his fellow researchers seem to be getting the message, he hasn't necessarily forced anyone to do a better job. He fears he won't in the end have done much to improve anyone's health. "There may not be fierce objections to what I'm saying," he explains. "But it's difficult to change the way that everyday doctors, patients, and healthy people think and behave."

As helter-skelter as the University of Ioannina Medical School campus looks, the hospital abutting it looks reassuringly stolid. Athina Tatsioni has offered to take me on a tour of the facility, but we make it only as far as the entrance when she is greeted—accosted, really—by a worried-looking older woman. Tatsioni, normally a bit reserved, is warm and animated with the woman, and the two have a brief but intense conversation before embracing and saying goodbye. Tatsioni explains to me that the woman and her husband were patients of hers years ago; now the husband has been admitted to the hospital with abdominal pains, and Tatsioni has promised she'll stop by his room later to say hello. Recalling the appendicitis story, I prod a bit, and she confesses she plans to do her own exam. She needs to be circumspect, though, so she won't appear to be second-guessing the other doctors.

Tatsioni doesn't so much fear that someone will carve out the man's healthy appendix. Rather, she's concerned that, like many patients, he'll

end up with prescriptions for multiple drugs that will do little to help him, and may well harm him. "Usually what happens is that the doctor will ask for a suite of biochemical tests—liver fat, pancreas function, and so on," she tells me. "The tests could turn up something, but they're probably irrelevant. Just having a good talk with the patient and getting a close history is much more likely to tell me what's wrong." Of course, the doctors have all been trained to order these tests, she notes, and doing so is a lot quicker than a long bedside chat. They're also trained to ply the patient with whatever drugs might help whack any errant test numbers back into line. What they're not trained to do is to go back and look at the research papers that helped make these drugs the standard of care. "When you look the papers up, you often find the drugs didn't even work better than a placebo. And no one tested how they worked in combination with the other drugs," she says. "Just taking the patient off everything can improve their health right away." But not only is checking out the research another time-consuming task, patients often don't even like it when they're taken off their drugs, she explains; they find their prescriptions reassuring.

Later, Ioannidis tells me he makes a point of having several clinicians on his team. "Researchers and physicians often don't understand each other; they speak different languages," he says. Knowing that some of his researchers are spending more than half their time seeing patients makes him feel the team is better positioned to bridge that gap; their experience informs the team's research with firsthand knowledge, and helps the team shape its papers in a way more likely to hit home with physicians. It's not that he envisions doctors making all their decisions based solely on solid evidence—there's simply too much complexity in patient treatment to pin down every situation with a great study. "Doctors need to rely on instinct and judgment to make choices," he says. "But these choices should be as informed as possible by the evidence. And if the evidence isn't good, doctors should know that, too. And so should patients."

In fact, the question of whether the problems with medical research should be broadcast to the public is a sticky one in the meta-research community. Already feeling that they're fighting to keep patients from turning to alternative medical treatments such as homeopathy, or misdiagnosing themselves on the

Internet, or simply neglecting medical treatment altogether, many researchers and physicians aren't eager to provide even more reason to be skeptical of what doctors do—not to mention how public disenchantment with medicine could affect research funding. Ioannidis dismisses these concerns. "If we don't tell the public about these problems, then we're no better than nonscientists who falsely claim they can heal," he says. "If the drugs don't work and we're not sure how to treat something, why should we claim differently? Some fear that there may be less funding because we stop claiming we can prove we have miraculous treatments. But if we can't really provide those miracles, how long will we be able to fool the public anyway? The scientific enterprise is probably the most fantastic achievement in human history, but that doesn't mean we have a right to overstate what we're accomplishing."

We could solve much of the wrongness problem, Ioannidis says, if the world simply stopped expecting scientists to be right. That's because being wrong in science is fine, and even necessary—as long as scientists recognize that they blew it, report their mistake openly instead of disguising it as a success, and then move on to the next thing, until they come up with the very occasional genuine breakthrough. But as long as careers remain contingent on producing a stream of research that's dressed up to seem more right than it is, scientists will keep delivering exactly that.

"Science is a noble endeavor, but it's also a low-yield endeavor," he says. "I'm not sure that more than a very small percentage of medical research is ever likely to lead to major improvements in clinical outcomes and quality of life. We should be very comfortable with that fact."

### 0.0.19 Meta-Analysis at 25; Gene V Glass, January 2000

Gene V Glass
College of Education
Arizona State University

My topic is meta-analysis. It has been nearly 25 years since meta-analysis, under that name and in its current guise made its first appearance. I wish to avoid the weary references to the new century or millenium – depending on how apocalyptic you're feeling (besides, it's 5759 on my calendar anyway) – and simply point out that meta-analysis is at the age when most things graduate from college, so it's not too soon to ask what accounting can be made of it. I have refrained from publishing anything on the topic of the methods of meta-analysis since about 1980 out of a reluctance to lay some heavy hand on other people's enthusiasms and a wish to hide my cynicism from public view. Others have eagerly advanced its development and I'll get to their contributions shortly (Cooper and Hedges, 1994; Hedges and Olkin, 1985; Hunter, Schmidt and Jackson, 1982).

Autobiography may be the truest, most honest narrative, even if it risks self aggrandizement, or worse, self-deception. Forgive me if I risk the latter for the sake of the former. For some reason it is increasingly difficult these days to speak in any other way.

In the span of this rather conventional paper, I wish to review the brief history of the form of quantitative research synthesis that is now generally known as "meta-analysis" (though I can't possibly recount this history as well as has Morton Hunt (1997) in his new book *How Science Takes Stock: The Story of Meta-Analysis*), tell where it came from, why it happened when it did, what was wrong with it and what remains to be done to make the findings of research in the social and behavioral sciences more understandable and useful.

Meta-analysis Beginnings:

In 25 years, meta-analysis has grown from an unheard of preoccupation of a very small group of statisticians working on problems of research integration in education and psychotherapy to a minor academic industry, as well as a commercial endeavor. A keyword web search – the contemporary

measure of visibility and impact – (Excite, January 28, 2000) on the word "meta-analysis" brings 2,200 "hits" of varying degrees of relevance, of course.[1] About 25% of the articles in the *Psychological Bulletin* in the past several years have the term "meta-analysis" in the title. Its popularity in the social sciences and education is nothing compared to its influence in medicine, where literally hundreds of meta-analyses have been published in the past 20 years. (In fact, my internist quotes findings of what he identifies as published meta-analyses during my physical exams.) An ERIC search shows well over 1,500 articles on meta-analyses written since 1975.

Surely it is true that as far as meta-analysis is concerned, necessity was the mother of invention, and if it hadn't been invented – so to speak – in the early 1970s it would have been invented soon thereafter since the volume of research in many fields was growing at such a rate that traditional narrative approaches to summarizing and integrating research were beginning to break down. But still, the combination of circumstances that brought about meta-analysis in about 1975 may itself be interesting and revealing. There were three circumstances that influenced me.

The first was personal. I left the University of Wisconsin in 1965 with a brand new PhD in psychometrics and statistics and a major league neurosis – years in the making – that was increasingly making my life miserable. Luckily, I found my way into psychotherapy that year while on the faculty of the University of Illinois and never left it until eight years later while teaching at the University of Colorado. I was so impressed with the power of psychotherapy as a means of changing my life and making it better that by 1970 I was studying clinical psychology (with the help of a good friend and colleague Vic Raimy at Boulder) and looking for opportunities to gain experience doing therapy. In spite of my personal enthusiasm for psychotherapy, the weight of academic opinion at that time derived from Hans Eysenck's frequent and tendentious reviews of the psychotherapy outcome research that proclaimed psychotherapy as worthless – a mere placebo, if that. I found this conclusion personally threatening – it called into question not only the preoccupation of about a decade of my life but my scholarly judgment (and the wisdom of having dropped a fair chunk of change) as well. I read Eysenck's literature

[1] A Google search on the one word "meta-analysis" on April 1, 2010, brought 10,400,000 hits.

reviews and was impressed primarily with their arbitrariness, idiosyncrasy and high-handed dismissiveness. I wanted to take on Eysenck and show that he was wrong: psychotherapy does change lives and make them better.

The second circumstance that prompted meta-analysis to come out when it did had to do with an obligation to give a speech. In 1974, I was elected President of the American Educational Research Association, in a peculiar miscarriage of the democratic process. This position is largely an honorific title that involves little more than chairing a few Association Council meetings and delivering a "presidential address" at the Annual Meeting. It's the "presidential address" that is the problem. No one I know who has served as AERA President really feels that they deserved the honor; the number of more worthy scholars passed over not only exceeds the number of recipients of the honor by several times, but as a group they probably outshine the few who were honored. Consequently, the need to prove one's worthiness to oneself and one's colleagues is nearly overwhelming, and the most public occasion on which to do it is the Presidential address, where one is assured of an audience of 1,500 or so of the world's top educational researchers. Not a few of my predecessors and contemporaries have cracked under this pressure and succumbed to the temptation to spin out grandiose fantasies about how educational research can become infallible or omnipotent, or about how government at national and world levels must be rebuilt to conform to the dictates of educational researchers. And so I approached the middle of the 1970s knowing that by April 1976 I was expected to release some bombast on the world that proved my worthiness for the AERA Presidency, and knowing that most such speeches were embarrassments spun out of feelings of intimidation and unworthiness. (A man named Richard Krech, I believe, won my undying respect when I was still in graduate school; having been distinguished by the American Psychological Association in the 1960s with one of its highest research awards, Krech, a professor at Berkeley, informed the Association that he was honored, but that he had nothing particularly new to report to the organization at the obligatory annual convention address, but if in the future he did have anything worth saying, they would hear it first.)

The third set of circumstances that joined my wish to annihilate Eysenck and prove that psychotherapy really works and my need to make a big splash

with my Presidential Address was that my training under the likes of Julian Stanley, Chester Harris, Henry Kaiser and George E. P. Box at Wisconsin in statistics and experimental design had left me with a set of doubts and questions about how we were advancing the empirical agenda in educational research. In particular, I had learned to be very skeptical of statistical significance testing; I had learned that all research was imperfect in one respect or another (or, in other words, there are no "perfectly valid" studies nor any line that demarcates "valid" from "invalid" studies); and third, I was beginning to question a taken-for-granted assumption of our work that we progress toward truth by doing what everyone commonly refers to as "studies." (I know that these are complex issues that need to be thoroughly examined to be accurately communicated, and I shall try to return to them.) I recall two publications from graduate school days that impressed me considerably. One was a curve relating serial position of a list of items to be memorized to probability of correct recall that Benton Underwood (1957) had synthesized from a dozen or more published memory experiments. The other was a *Psychological Bulletin* article by Sandy Astin on the effects of glutamic acid on mental performance (whose results presaged a meta-analysis of the Feingold diet research 30 years later in that poorly controlled experiments showed benefits and well controlled experiments did not).

Permit me to say just a word or two about each of these studies because they very much influenced my thinking about how we should "review" research. Underwood had combined the findings of 16 experiments on serial learning to demonstrate a consistent geometrically decreasing curve describing the declining probability of correct recall as a function of number of previously memorized items, thus giving strong weight to an interference explanation of recall errors. What was interesting about Underwood's curve was that it was an amalgamation of studies that had different lengths of lists and different items to be recalled (nonsense syllables, baseball teams, colors and the like). Astin's *Psychological Bulletin* review had attracted my attention in another respect. Glutamic acid – it will now scarcely be remembered – was a discovery of the 1950s that putatively increased the ability of tissue to absorb oxygen. Reasoning with the primitive constructs of the time, researchers hypothesized that more oxygen to the brain would produce more

intelligent behavior. (It is not known what amount of oxygen was reaching the brains of the scientists proposing this hypothesis.) A series of experiments in the 1950s and 1960s tested glutamic acid against "control groups" and by 1961, Astin was able to array these findings in a cross-tabulation that showed that the chances of finding a significant effect for glutamic acid were related (according to a chi-square test) to the presence or absence of various controls in the experiment; placebos and blinding of assessors, for example, were associated with no significant effect of the acid. As irrelevant as the chi-square test now seems, at the time I saw it done, it was revelatory to see "studies" being treated as data points in a statistical analysis. (In 1967, I attempted a similar approach while reviewing the experimental evidence on the Doman-Delacato pattern therapy. Glass and Robbins, 1967) At about the same time I was reading Underwood and Astin, I certainly must have read Ben Bloom's *Stability of Human Characteristics* (1963), but its aggregated graphs of correlation coefficients made no impression on me, because it was many years after work to be described below that I noticed a similarity between his approach and meta-analysis. Perhaps the connections were not made because Bloom dealt with variables such as age, weight, height, IQ and the like where the problems of dissimilarity of variables did not force one to worry about the kinds of problem that lie at the heart of meta-analysis.

If precedence is of any concern, Bob Rosenthal deserves as much credit as anyone for furthering what we now conveniently call "meta-analysis." In 1976, he published *Experimenter Effects in Behavioral Research*, which contained calculations of many "effect sizes" (i.e., standardized mean differences) that were then compared across domains or conditions. If Bob had just gone a little further in quantifying study characteristics and subjecting the whole business to regression analyses and what-not, and then thinking up a snappy name, it would be his name that came up every time the subject is research integration. But Bob had an even more positive influence on the development of meta-analysis than one would infer from his numerous methodological writings on the subject. When I was making my initial forays onto the battlefield of psychotherapy outcome research – about which more soon – Bob wrote me a very nice and encouraging letter in which he indicated that the approach we were taking made perfect sense. Of course, it ought to have

made sense to him, considering that it was not that different from what he had done in *Experimenter Effects*. He probably doesn't realize how important that validation from a stranger was. (And while on the topic of snappy names, although people have suggested or promoted several polysyllabic alternatives – quantitative synthesis, statistical research integration – the name meta-analysis, suggested by Michael Scriven's meta-evaluation (meaning the evaluation of evaluations), appears to have caught on. To press on further into it, the "meta" comes from the Greek preposition meaning "behind" or "in back of." Its application as in "metaphysics" derives from the fact that in the publication of Aristotle's writings during the Middle Ages, the section dealing with the transcendental was bound immediately behind the section dealing with physics; lacking any title provided by its author, this final section became known as Aristotle's "metaphysics." So, in fact, metaphysics is not some grander form of physics, some all encompassing, overarching general theory of everthing; it is merely what Aristotle put after the stuff he wrote on physics. The point of this aside is to attempt to leach out of the term "meta-analysis" some of the grandiosity that others see in it. It is not the grand theory of research; it is simply a way of speaking of the statistical analysis of statistical analyses.)

So positioned in these circumstances, in the summer of 1974, I set about to do battle with Dr. Eysenck and prove that psychotherapy – my psychotherapy – was an effective treatment. (Incidentally, though it may be of only the merest passing interest, my preferences for psychotherapy are Freudian, a predilection that causes Ron Nelson and other of my ASU colleagues great distress, I'm sure.) I joined the battle with Eysenck's 1965 review of the psychotherapy outcome literature.

Eysenck began his famous reviews by eliminating from consideration all theses, dissertations, project reports or other contemptible items not published in peer-reviewed journals. This arbitrary exclusion of literally hundreds of evaluations of therapy outcomes was indefensible. It's one thing to believe that peer review guarantees truth; it is quite another to believe that all truth appears in peer reviewed journals. (The most important paper on the multiple comparisons problem in ANOVA was distributed as an unpublished ditto manuscript from the Princeton University Mathematics Department by John

Tukey; it never was published in a peer reviewed journal.)

Next, Eysenck eliminated any experiment that did not include an untreated control group. This makes no sense whatever, since head-to-head comparisons of two different types of psychotherapy contribute a great deal to our knowledge of psychotherapy effects. If a horse runs 20 mph faster than a man and 35 mph faster than a pig, I can conclude with confidence that the man will outrun the pig by 15 mph.

Having winnowed a huge literature down to 11 studies (!) by whim and prejudice, Eysenck proceeded to describe their findings solely in terms of whether or not statistical significance was attained at the .05 level. No matter that the results may have barely missed the .05 level or soared beyond it. All that Eysenck considered worth noting about an experiment was whether the differences reached significance at the .05 level. If it reached significance at only the .07 level, Eysenck classified it as showing "no effect for psychotherapy."

Finally, Eysenck did something truly staggering in its illogic. If a study showed significant differences favoring therapy over control on what he regarded as a "subjective" measure of outcome (e.g., the Rorschach or the Thematic Apperception Test), he discounted the findings entirely. So be it; he may be a tough case, but that's his right. But then, when encountering a study that showed differences on an "objective" outcome measure (e.g., GPA) but no differences on a subjective measure (like the TAT), Eysenck discounted the entire study because the outcome differences were "inconsistent."

Looking back on it, I can almost credit Eysenck with the invention of meta-analysis by anti-thesis. By doing everything in the opposite way that he did, one would have been led straight to meta-analysis. Adopt an *a posteriori* attitude toward including studies in a synthesis, replace statistical significance by measures of strength of relationship or effect, and view the entire task of integration as a problem in data analysis where "studies" are quantified and the resulting data-base subjected to statistical analysis, and meta-analysis assumes its first formulation. (Thank you, Professor Eysenck.)

Working with my colleague Mary Lee Smith, I set about to collect all the psychotherapy outcome studies that could be found and subjected them to this new form of analysis. By May of 1975, the results were ready to try out on

a friendly group of colleagues. The May 12th Group had been meeting yearly since about 1968 to talk about problems in the area of program evaluation. The 1975 meeting was held in Tampa at Dick Jaeger's place. I worked up a brief handout and nervously gave my friends an account of the preliminary results of the psychotherapy meta-analysis. Lee Cronbach was there; so was Bob Stake, David Wiley, Les McLean and other trusted colleagues who could be relied on to demolish any foolishness they might see. To my immense relief they found the approach plausible or at least not obviously stupid. (I drew frequently in the future on that reassurance when others, whom I respected less, pronounced the entire business stupid.)

The first meta-analysis of the psychotherapy outcome research found that the typical therapy trial raised the treatment group to a level about two-thirds of a standard deviation on average above untreated controls; the average person receiving therapy finished the experiment in a position that exceeded the 75th percentile in the control group on whatever outcome measure happened to be taken. This finding summarized dozens of experiments encompassing a few thousand persons as subjects and must have been cold comfort to Professor Eysenck.

An expansion and reworking of the psychotherapy experiments resulted in the paper that was delivered as the much feared AERA Presidential address in April 1976. Its reception was gratifying. Two months later a long version was presented at a meeting of psychotherapy researchers in San Diego. Their reactions foreshadowed the eventual reception of the work among psychologists. Some said that the work was revolutionary and proved what they had known all along; others said it was wrongheaded and meaningless. The widest publication of the work came in 1977, in a now, may I say, famous article by Smith and Glass in the *American Psychologist*. Eysenck responded to the article by calling it "mega-silliness," a moderately clever play on meta-analysis that nonetheless swayed few. Psychologists tended to fixate on the fact that the analysis gave no warrant to any claims that one type or style of psychotherapy was any more effective than any other: whether called "behavioral" or "Rogerian" or "rational" or "psychodynamic," all the therapies seemed to work and to work to about the same degree of effectiveness. Behavior therapists, who had claimed victory in the psychotherapy horse

race because they were "scientific" and others weren't, found this conclusion unacceptable and took it as reason enough to declare meta-analysis invalid. Non-behavioral therapists – the Rogerians, Adlerians and Freudians, to name a few – hailed the meta-analysis as one of the great achievements of psychological research: a "classic," a "watershed." My cynicism about research and much of psychology dates from approximately this period.

Criticisms of Meta-analysis:

The first appearances of meta-analysis in the 1970s were not met universally with encomiums and expressions of gratitude. There was no shortage of critics who found the whole idea wrong-headed, senseless, misbegotten, etc.

The Apples-and-Oranges Problem:

Of course the most often repeated criticism of meta-analysis was that it was meaningless because it "mixed apples and oranges." I was not unprepared for this criticism; indeed, I had long before prepared my own defense: "Of course it mixes apples and oranges; in the study of fruit nothing else is sensible; comparing apples and oranges is the only endeavor worthy of true scientists; comparing apples to apples is trivial." But I misjudged the degree to which this criticism would take hold of people's opinions and shut down their minds. At times I even began to entertain my own doubts that it made sense to integrate any two studies unless they were studies of "the same thing." But, the same persons who were arguing that no two studies should be compared unless they were studies of the "same thing," were blithely comparing persons (i.e., experimental "subjects") within their studies all the time. This seemed inconsistent. Plus, I had a glimmer of the self-contradictory nature of the statement "No two things can be compared unless they are the same." If they are the same, there is no reason to compare them; indeed, if "they" are the same, then there are not two things, there is only one thing and comparison is not an issue. And yet I had a gnawing insecurity that the critics might be right. One study is an apple, and a second study is an orange; and comparing them is as stupid as comparing apples and oranges, except that sometimes I do hesitate while considering whether I'm hungry for an apple or an orange.

At about this time – late 1970s – I was browsing through a new book that I had bought out of a vague sense that it might be worth my time because

it was written by a Harvard philosopher, carried a title like *Philosophical Explanations* and was written by an author – Robert Nozick – who had written one of the few pieces on the philosophy of the social sciences that ever impressed me as being worth rereading. To my amazement, Nozick spent the first one hundred pages of his book on the problem of "identity," i.e., what does it mean to say that two things are the same? Starting with the puzzle of how two things that are alike in every respect would not be one thing, Nozick unraveled the problem of identity and discovered its fundamental nature underlying a host of philosophical questions ranging from "How do we think?" to "How do I know that I am I?" Here, I thought at last, might be the answer to the "apples and oranges" question. And indeed, it was there.

Nozick considered the classic problem of Theseus's ship. Theseus, King of Thebes, and his men are plying the waters of the Mediterranean. Each day a sailor replaces a wooden plank in the ship. After nearly five years, every plank has been replaced. Are Theseus and his men still sailing in the same ship that was launched five years earlier on the Mediterranean? "Of course," most will answer. But suppose that as each original plank was removed, it was taken ashore and repositioned exactly as it had been on the waters, so that at the end of five years, there exists a ship on shore, every plank of which once stood in exactly the same relationship to every other in what five years earlier had been Theseus's ship. Is this ship on shore – which we could easily launch if we so chose – Theseus's ship? Or is the ship sailing the Mediterranean with all of its new planks the same ship that we originally regarded as Theseus's ship? The answer depends on what we understand the concept of "same" to mean?

Consider an even more troubling example that stems from the problem of the persistence of personal identity. How do I know that I am that person who I was yesterday, or last year, or twenty-five years ago? Why would an old high-school friend say that I am Gene Glass, even though hundreds, no thousands of things about me have changed since high school? Probably no cells are in common between this organism and the organism that responded to the name "Gene Glass" forty years ago; I can assure you that there are few attitudes and thoughts held in common between these two organisms –

or is it one organism? Why then, would an old high-school friend, suitably prompted, say without hesitation, "Yes, this is Gene Glass, the same person I went to high school with." Nozick argued that the only sense in which personal identity survives across time is in the sense of what he called "the closest related continuer." I am still recognized as Gene Glass to those who knew me then because I am that thing most closely related to that person to whom they applied the name "Gene Glass" over forty years ago. Now notice that implied in this concept of the "closest related continuer" are notions of distance and relationship. Nozick was quite clear that these concepts had to be given concrete definition to understand how in particular instances people use the concept of identity. In fact, to Nozick's way of thinking, things are compared by means of weighted functions of constituent factors, and their "distance" from each other is "calculated" in many instances in a Euclidean way.

Consider Theseus's ship again. Is the ship sailing the seas the "same" ship that Theseus launched five years earlier? Or is the ship on the shore made of all the original planks from that first ship the "same" as Theseus's original ship? If I give great weight to the materials and the length of time those materials functioned as a ship (i.e., to displace water and float things), then the vessel on the shore is the closest related continuation of what historically had been called "Theseus's ship." But if, instead, I give great weight to different factors such as the importance of the battles the vessel was involved in (and Theseus's big battles were all within the last three years), then the vessel that now floats on the Mediterranean – not the ship on the shore made up of Theseus's original planks – is Theseus's ship, and the thing on the shore is old spare parts. So here was Nozick saying that the fundamental riddle of how two things could be the same ultimately resolves itself into an empirical question involving observable factors and weighing them in various combinations to determine the closest related continuer. The question of "sameness" is not an a priori question at all; apart from being a logical impossibility, it is an empirical question. For us, no two "studies" are the same. All studies differ and the only interesting questions to ask about them concern how they vary across the factors we conceive of as important. This notion is not fully developed here and I will return to it later.

The "Flat Earth" Criticism:

I may not be the best person to critique meta-analysis, for obvious reasons. However, I will cop to legitimate criticisms of the approach when I see them, and I haven't seen many. But one criticism rings true because I knew at the time that I was being forced into a position with which I wasn't comfortable. Permit me to return to the psychotherapy meta-analysis.

Eysenck was, as I have said, a nettlesome critic of the psychotherapy establishment in the 1960s and 1970s. His exaggerated and inflammatory statements about psychotherapy being worthless (no better than a placebo) were not believed by psychotherapists or researchers, but they were not being effectively rebutted either. Instead of taking him head-on, as my colleagues and I attempted to do, researchers, like Gordon Paul, for example, attempted to argue that the question whether psychotherapy was effective was fundamentally meaningless. Rather, asserted Paul while many others assented, the only legitimate research question was "What type of therapy, with what type of client, produces what kind of effect?" I confess that I found this distracting dodge as frustrating as I found Eysenck's blanket condemnation. Here was a critic – Eysenck – saying that all psychotherapists are either frauds or gullible, self-deluded incompetents, and the establishment's response is to assert that he is not making a meaningful claim. Well, he was making a meaningful claim; and I already knew enough from the meta-analysis of the outcome studies to know that Paul's question was unanswerable due to insufficient data, and that researchers were showing almost no interest in collecting the kind of data that Paul and others argued were the only meaningful data.

It fell to me, I thought, to argue that the general question "Is psychotherapy effective?" is meaningful and that psychotherapy is effective. Such generalizations – across types of therapy, types of client and types of outcome – are meaningful to many people – policy makers, average citizens – if not to psychotherapy researchers or psychotherapists themselves. It was not that I necessarily believed that different therapies did not have different effects for different kinds of people; rather, I felt certain that the available evidence, tons of it, did not establish with any degree of confidence what these differential effects were. It was safe to say that in general psychotherapy works on

many things for most people, but it was impossible to argue that this therapy was better than that therapy for this kind of problem. (I might add that twenty years after the publication of *The Benefits of Psychotherapy*, I still have not seen compelling answers to Paul's questions, nor is there evidence of researchers having any interest in answering them.)

The circumstances of the debate, then, put me in the position of arguing, circa 1980, that there are very few differences among various ways of treating human beings and that, at least, there is scarcely any convincing experimental evidence to back up claims of differential effects. And that policy makers and others hardly need to waste their time asking such questions or looking for the answers. Psychotherapy works; all types of therapy work about equally well; support any of them with your tax dollars or your insurance policies. Class size reductions work – very gradually at first (from 30 to 25 say) but more impressively later (from 15 to 10); they work equally for all grades, all subjects, all types of student. Reduce class sizes, and it doesn't matter where or for whom.

Well, one of my most respected colleagues called me to task for this way of thinking and using social science research. In a beautiful and important paper entitled "Prudent Aspirations for the Social Sciences," Lee Cronbach chastised his profession for promising too much and chastised me for expecting too little. He lumped me with a small group of like-minded souls into what he named the "Flat Earth Society," i.e., a group of people who believe that the terrain that social scientists explore is featureless, flat, with no interesting interactions or topography. All therapies work equally well; all tests predict success to about the same degree; etc.:

... some of our colleagues are beginning to sound like a kind of Flat Earth Society. They tell us that the world is essentially simple: most social phenomena are adequately described by linear relations; one-parameter scaling can discover coherent variables independent of culture and population; and inconsistences among studies of the same kind will vanish if we but amalgamate a sufficient number of studies. ... The Flat Earth folk seek to bury any complex hypothesis with an empirical bulldozer." (Cronbach, 1982, p. 70.)

Cronbach's criticism stung because it was on target. In attempting to refute Eysenck's outlandishness without endorsing the psychotherapy estab-

lishment's obfuscation, I had taken a position of condescending simplicity. A meta-analysis will give you the BIG FACT, I said; don't ask for more sophisticated answers; they aren't there. My own work tended to take this form, and much of what has ensued in the past 25 years has regrettably followed suit. Effect sizes – if it is experiments that are at issue – are calculated, classified in a few ways, perhaps, and all their variability is then averaged across. Little effort is invested in trying to plot the complex, variegated landscape that most likely underlies our crude averages. Consider an example that may help illuminate these matters. Perhaps the most controversial conclusion from the psychotherapy meta-analysis that my colleagues and I published in 1980 was that there was no evidence favoring behavioral psychotherapies over non-behavioral psychotherapies. This finding was vilified by the behavioral therapy camp and praised by the Rogerians and Freudians. Some years later, prodded by Cronbach's criticism, I returned to the database and dug a little deeper. When the nine experiments extant in 1979 – and I would be surprised if there are many more now – in which behavioral and non-behavioral psychotherapies are compared in the same experiment between randomized groups and the effects of treatment are plotted as a function of follow-up time, the two curves intersect. The findings are quite extraordinary and suggestive. Behavioral therapies produce large short-term effects which decay in strength over the first year of follow-up; non-behavioral therapies produce initially smaller effects which increase over time. The two curves appear to be converging on the same long-term effect. I leave it to the reader to imagine why. One answer, I suspect, is not arcane and is quite plausible.

This conclusion is, I believe, truer to Cronbach's conception of reality and how research, even meta-analysis, can lead us to a more sophisticated understanding of our world. Indeed, the world is not flat; it encompasses all manner of interesting hills and valleys, and in general, averages do not do it justice.

Extensions of Meta-analysis:

In the twenty-five years between the first appearance of the word "meta-analysis" in print and today, there have been several attempts to modify the approach, or advance alternatives to it, or extend the method to reach auxiliary issues. If I may be so cruel, few of these efforts have added much.

One of the hardest things to abide in following the developments in meta-analysis methods in the past couple of decades was the frequent observation that what I had contributed to the problem of research synthesis was the idea of dividing mean differences by standard deviations. "Effect sizes," as they are called, had been around for decades before I opened my first statistics text. Having to read that "Glass has proposed integrating studies by dividing mean differences by standard deviations and averaging them" was a bitter pill to swallow. Some of the earliest work that I and my colleagues did involved using a variety of outcome measures to be analyzed and synthesized: correlations, regression coefficients, proportions, odds ratios. Well, so be it; better to be mentioned in any favorable light than not to be remembered at all.

After all, this was not as hard to take as newly minted confections such as "best evidence research synthesis," a come-lately contribution that added nothing whatsoever to what myself and many others had been saying repeatedly on the question of whether meta-analyses should use all studies or only "good" studies. I remain staunchly committed to the idea that meta-analyses must deal with all studies, good bad and indifferent, and that their results are only properly understood in the context of each other, not after having been censored by some a priori set of prejudices. An effect size of 1.50 for 20 studies employing randomized groups has a whole different meaning when 50 studies using matching show an average effect of 1.40 than if 50 matched groups studies show an effect of $-.50$, for example.

Statistical Inference in Meta-analysis:

The appropriate role for inferential statistics in meta-analysis is not merely unclear, it has been seen quite differently by different methodologists in the 25 years since meta-analysis appeared. In 1981, in the first extended discussion of the topic (Glass, McGaw, & Smith, 1981), I raised doubts about the applicability of inferential statistics in meta-analysis. Inference at the level of persons within seemed quite unnecessary, since even a modest size synthesis will involve a few hundred persons (nested within studies) and lead to nearly automatic rejection of null hypotheses. Moreover, the chances are remote that the persons or subjects within studies were drawn from defined populations with anything even remotely resembling probabilistic techniques. Hence, probabilistic calculations advanced as if subjects had been randomly

selected would be dubious. At the level of "studies," the question of the appropriateness of inferential statistics can be posed again, and the answer again seems to be negative. There are two instances in which common inferential methods are clearly appropriate, not just in meta-analysis but in any research: 1) when a well defined population has been randomly sampled, and 2) when subjects have been randomly assigned to conditions in a controlled experiment. In the latter case, Fisher showed how the permutation test can be used to make inferences to the universe of all possible permutations. But this case is of little interest to meta-analysts who never assign units to treatments. Moreover, the typical meta-analysis virtually never meets the condition of probabilistic sampling of a population (though in one instance (Smith, Glass & Miller, 1980), the available population of psychoactive drug treatment experiments was so large that a random sample of experiments was in fact drawn for the meta-analysis). Inferential statistics has little role to play in meta-analysis.

It is common to acknowledge, in meta-analysis and elsewhere, that many data sets fail to meet probabilistic sampling conditions, and then to argue that one ought to treat the data in hand "as if" it were a random sample of some hypothetical population. One must be wary here of the slide from "hypothesis about a population" into "a hypothetical population." They are quite different things, the former being standard and unobjectionable, the latter being a figment with which we hardly know how to deal. Under this stipulation that one is making inferences not to some defined or known population but a hypothetical one, inferential techniques are applied and the results inspected. The direction taken mirrors some of the earliest published opinion on this problem in the context of research synthesis, expressed, for example, by Mosteller and his colleagues in 1977: "One might expect that if our MEDLARS approach were perfect and produced all the papers we would have a census rather than a sample of the papers. To adopt this model would be to misunderstand our purpose. We think of a process producing these research studies through time, and we think of our sample – even if it were a census – as a sample in time from the process. Thus, our inference would still be to the general process, even if we did have all appropriate papers from a time period." (Gilbert, McPeek & Mosteller, 1977, p. 127; quoted in Cook

et al., 1992, p. 291) This position is repeated in slightly different language by Larry Hedges in Chapter 3 "Statistical Considerations" of the *Handbook of Research Synthesis* (1994): "The universe is the hypothetical collection of studies that could be conducted in principle and about which we wish to generalize. The study sample is the ensemble of studies that are used in the review and that provide the effect size data used in the research synthesis." (p. 30)

These notions appear to be circular. If the sample is fixed and the population is allowed to be hypothetical, then surely the data analyst will imagine a population that resembles the sample of data. If I show you a handful of red and green M&Ms, you will naturally assume that I have just drawn my hand out of a bowl of mostly red and green M&Ms, not red and green and brown and yellow ones. Hence, all of these "hypothetical populations" will be merely reflections of the samples in hand and there will be no need for inferential statistics. Or put another way, if the population of inference is not defined by considerations separate from the characterization of the sample, then the population is merely a large version of the sample. With what confidence is one able to generalize the character of this sample to a population that looks like a big version of the sample? Well, with a great deal of confidence, obviously. But then, the population is nothing but the sample writ large and we really know nothing more than what the sample tells us in spite of the fact that we have attached misleadingly precise probability numbers to the result.

Hedges and Olkin (1985) have developed inferential techniques that ignore the *pro forma* testing (because of large $N$) of null hypotheses and focus on the estimation of regression functions that estimate effects at different levels of study. They worry about both sources of statistical instability: that arising from persons within studies and that which arises from variation between studies. The techniques they present are based on traditional assumptions of random sampling and independence. It is, of course, unclear to me precisely how the validity of their methods are compromised by failure to achieve probabilistic sampling of persons and studies.

The irony of traditional hypothesis testing approaches applied to meta-analysis is that whereas consideration of sampling error at the level of persons

always leads to a *pro forma* rejection of "null hypotheses" (of zero correlation or zero average effect size), consideration of sampling error at the level of study characteristics (the study, not the person as the unit of analysis) leads to too few rejections (too many Type II errors, one might say). Hedges's homogeneity test of the hypothesis that all studies in a group estimate the same population parameter [is] frequently seen in published meta-analyses these days. Once a hypothesis of homogeneity is accepted by Hedges's test, one is advised to treat all studies within the ensemble as the same. Experienced data analysts know, however, that there is typically a good deal of meaningful covariation between study characteristics and study findings even within ensembles where Hedges's test can not reject the homogeneity hypothesis. The situation is parallel to the experience of psychometricians discovering that they could easily interpret several more common factors than inferential solutions (maximum-likelihood; LISREL) could confirm. The best data exploration and discovery are more complex and convincing than the most exact inferential test. In short, classical statistics seems not able to reproduce the complex cognitive processes that are commonly applied with success by data analysts.

Donald Rubin (1990) addressed some of these issues squarely and articulated a position that I find very appealing: " ... consider the idea that sampling and representativeness of the studies in a meta-analysis are important. I will claim that this is nonsense – we don't have to worry about representing a population but rather about other far more important things." (p. 155) These more important things to Rubin are the estimation of treatment effects under a set of standard or ideal study conditions. This process, as he outlined it, involves the fitting of response surfaces (a form of quantitative model building) between study effects ($Y$) and study conditions ($X$, $W$, $Z$, etc.). I would only add to Rubin's statement that we are interested in not merely the response of the system under ideal study conditions but under many conditions having nothing to do with an ideally designed study, e.g., person characteristics, follow-up times and the like.

By far most meta-analyses are undertaken in pursuit not of scientific theory but technological evaluation. The evaluation question is never whether some hypothesis or model is accepted or rejected but rather how "outputs"

or "benefits" or "effect sizes" vary from one set of circumstances to another; and the meta-analysis rarely works on a collection of data that can sensibly be described as a probability sample from anything.

Meta-analysis in the Next 25 Years:

If our efforts to research and improve education are to prosper, meta-analysis will have to be replaced by more useful and more accurate ways of synthesizing research findings. To catch a glimpse of what this future for research integration might look like, we need to look back at the deficiencies in our research customs that produced meta-analysis in the first place.

First, the high cost in the past of publishing research results led to cryptic reporting styles that discarded most of the useful information that research revealed. To encapsulate complex relationships in statements like "significant at the .05 level" was a travesty – a travesty that continues today out of bad habit and bureaucratic inertia.

Second, we need to stop thinking of ourselves as scientists testing grand theories, and face the fact that we are technicians collecting and collating information, often in quantitative forms. Paul Meehl (1967; 1978) dispelled once and for all the misconception that we in, what he called, the "soft social sciences" are testing theories in any way even remotely resembling how theory focuses and advances research in the hard sciences. Indeed, the mistaken notion that we are theory driven has, in Meehl's opinion, led us into a worthless *pro forma* ritual of testing and rejecting statistical hypotheses that are a priori known to be 99% false before they are tested.

Third, the conception of our work that held that "studies" are the basic, fundamental unit of a research program may be the single most counterproductive influence of all. This idea that we design a "study," and that a study culminates in the test of a hypothesis and that a hypothesis comes from a theory – this idea has done more to retard progress in educational research than any other single notion. Ask an educational researcher what he or she is up to, and they will reply that they are "doing a study," or "designing a study," or "writing up a study" for publication. Ask a physicist what's up and you'll never hear the word "study." (In fact, if one goes to http://xxx.lanl.gov where physicists archive their work, one will seldom see the word "study." Rather, physicists – the data gathering experimental ones – report data, all

of it, that they have collected under conditions that they carefully described. They contrive interesting conditions that can be precisely described and then they report the resulting observations.)

Meta-analysis was created out of the need to extract useful information from the cryptic records of inferential data analyses in the abbreviated reports of research in journals and other printed sources. "What does this $t$-test really say about the efficacy of ritalin in comparison to caffeine?" Meta-analysis needs to be replaced by archives of raw data that permit the construction of complex data landscapes that depict the relationships among independent, dependent and mediating variables. We wish to be able to answer the question, "What is the response of males ages 5-8 to ritalin at these dosage levels on attention, acting out and academic achievement after one, three, six and twelve months of treatment?"

We can move toward this vision of useful synthesized archives of research now if we simply re-orient our ideas about what we are doing when we do research. We are not testing grand theories, rather we are charting dosage-response curves for technological interventions under a variety of circumstances. We are not informing colleagues that our straw-person null hypothesis has been rejected at the .01 level, rather we are sharing data collected and reported according to some commonly accepted protocols. We aren't publishing "studies," rather we are contributing to data archives.

Five years ago, this vision of how research should be reported and shared seemed hopelessly quixotic. Now it seems easily attainable. The difference is the I-word: the Internet. In 1993, spurred by the ludicrously high costs and glacial turn-around times of traditional scholarly journals, I created an internet-based peer-reviewed journal on education policy analysis. This journal, named *Education Policy Analysis Archives*, is now in its seventh year of publication, has published 150 articles, is accessed daily without cost by nearly 1,000 persons (the other three paper journals in this field have average total subscription bases of fewer than 1,000 persons), and has an average "lag" from submission to publication of about three weeks. Moreover, we have just this year started accepting articles in both English and Spanish. And all of this has been accomplished without funds other than the time I put into it as part of my normal job: no secretaries, no graduate assistants,

nothing but a day or two a week of my time.

Two years ago, we adopted the policy that any one publishing a quantitative study in the journal would have to agree to archive all the raw data at the journal web site so that the data could be downloaded by any reader. Our authors have done so with enthusiasm. I think that you can see how this capability puts an entirely new face on the problem of how we integrate research findings: no more inaccurate conversions of inferential test statistics into something worth knowing like an effect size or a correlation coefficient or an odds ratio; no more speculating about distribution shapes; no more frustration at not knowing what violence has been committed when linear coefficients mask curvilinear relationships. Now we simply download each others' data, and the synthesis prize goes to the person who best assembles the pieces of the jigsaw puzzle into a coherent picture of how the variables relate to each other.

References:

Cook, T. D. (1992). *Meta-analysis for explanation – a casebook.* New York: Russell Sage Foundation; 1992.

Cooper, H. M. (1989). *Integrating research: a guide for literature reviews.* 2nd ed. Newbury Park, CA: SAGE Publications.

Cooper, H. M., & Hedges, L. V. (Eds.) (1994). *The handbook of research synthesis.* New York: Russell Sage Foundation.

Cronbach, L. J. (1982). Prudent aspirations for social inquiry. Chapter 5 (pp. 61–81) in Kruskal, W. H. (Ed.), *The social sciences: Their nature and uses.* Chicago: The University of Chicago Press.

Eysenck, H. J. (1965). The effects of psychotherapy. *International Journal of Psychiatry*, *1*, 97–187.

Glass, G. V (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*, 3–8.

Glass, G. V (1978). Integrating findings: The meta-analysis of research. *Review of Research in Education*, *5*, 351–379.

Glass, G. V, McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* Beverly Hills, CA: SAGE Publications.

Glass, G. V et al. (1982). *School class size: Research and policy.* Beverly Hills, CA: SAGE Publications.

Glass, G. V, & Robbins, M. P. (1967). A critique of experiments on the role of neurological organization in reading performance. *Reading Research Quarterly*, *3*, 5–51.

Hedges, L. V., Laine, R. D., & Greenwald, R. (1994). Does Money Matter? A Meta-Analysis of Studies of the Effects of Differential School Inputs on Student Outcomes. *Educational Researcher*, *23(3)*, 5–14.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* New York: Academic Press.

Hunt, M. (1997). *How science takes stock: The story of meta-analysis.* New York: Russell Sage Foundation.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park (CA): SAGE Publications.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies.* Beverly Hills, CA: SAGE Publications.

Light, R. J., Singer, J. D., & Willett, J. B. (1990). *By design: Planning research on higher education.* Cambridge, MA: Harvard University Press.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*, 103–15.

Meehl. P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*, 806–34.

Rosenthal, R. (1976). *Experimenter effects in behavioral research.* New York: John Wiley.

Rosenthal, R. (1991). *Meta-analytic procedures for social research.* Rev. ed. Newbury Park, CA: SAGE Publications.

Rubin, D. (1990). A new perspective. Chapter 14 (pp. 155–166) in Wachter, K. W. & Straf, M. L. (Eds.). *The future of meta-analysis.* New York: Russell Sage Foundation.

Smith, M. L., & Glass, G. V (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, *32*, 752–60.

Smith, M. L., Glass, G. V, & Miller, T. I. (1980). *The benefits of psychotherapy.* Baltimore: Johns Hopkins University Press.

Wachter, K. W., & Straf, M. L. (Eds.). (1990). *The future of meta-analysis*. New York: Russell Sage Foundation.

Wolf, F. M. (1986). *Meta-analysis: quantitative methods for research synthesis*. Beverly Hills, CA: SAGE Publications.

### 0.0.20 The Treatment, Malcolm Gladwell (*New Yorker*), May 17, 2010

May 17, 2010

Malcolm Gladwell (*New Yorker*)

Why is it so difficult to develop drugs for cancer?

In the world of cancer research, there is something called a Kaplan-Meier curve, which tracks the health of patients in the trial of an experimental drug. In its simplest version, it consists of two lines. The first follows the patients in the "control arm," the second the patients in the "treatment arm." In most cases, those two lines are virtually identical. That is the sad fact of cancer research: nine times out of ten, there is no difference in survival between those who were given the new drug and those who were not. But every now and again – after millions of dollars have been spent, and tens of thousands of pages of data collected, and patients followed, and toxicological issues examined, and safety issues resolved, and manufacturing processes fine-tuned – the patients in the treatment arm will live longer than the patients in the control arm, and the two lines on the Kaplan-Meier will start to diverge.

Seven years ago, for example, a team from Genentech presented the results of a colorectal-cancer drug trial at the annual meeting of the American Society of Clinical Oncology – a conference attended by virtually every major cancer researcher in the world. The lead Genentech researcher took the audience through one slide after another – click, click, click – laying out the design and scope of the study, until he came to the crucial moment: the Kaplan-Meier. At that point, what he said became irrelevant. The members of the audience saw daylight between the two lines, for a patient population in which that almost never happened, and they leaped to their feet and gave him an ovation. Every drug researcher in the world dreams of standing in front of thousands of people at ASCO and clicking on a Kaplan-Meier like that. "It is why we are in this business," Safi Bahcall says. Once he thought that this dream would come true for him. It was in the late summer of 2006, and is among the greatest moments of his life.

Bahcall is the C.E.O. of Synta Pharmaceuticals, a small biotechnology company. It occupies a one-story brick nineteen-seventies building outside

Boston, just off Route 128, where many of the region's high-tech companies have congregated, and that summer Synta had two compounds in development. One was a cancer drug called elesclomol. The other was an immune modulator called apilimod. Experimental drugs must pass through three phases of testing before they can be considered for government approval. Phase 1 is a small trial to determine at what dose the drug can be taken safely. Phase 2 is a larger trial to figure out if it has therapeutic potential, and Phase 3 is a definitive trial to see if it actually works, usually in comparison with standard treatments. Elesclomol had progressed to Phase 2 for soft-tissue sarcomas and for lung cancer, and had come up short in both cases. A Phase 2 trial for metastatic melanoma – a deadly form of skin cancer – was also under way. But that was a long shot: nothing ever worked well for melanoma. In the previous thirty-five years, there had been something like seventy large-scale Phase 2 trials for metastatic-melanoma drugs, and if you plotted all the results on a single Kaplan-Meier there wouldn't be much more than a razor's edge of difference between any two of the lines.

That left apilimod. In animal studies and early clinical trials for autoimmune disorders, it seemed promising. But when Synta went to Phase 2 with a trial for psoriasis, the results were underwhelming. "It was ugly," Bahcall says. "We had lung cancer fail, sarcoma next, and then psoriasis. We had one more trial left, which was for Crohn's disease. I remember my biostats guy coming into my office, saying, 'I've got some good news and some bad news. The good news is that apilimod is safe. We have the data. No toxicity. The bad news is that it's not effective.' It was heartbreaking."

Bahcall is a boyish man in his early forties, with a round face and dark, curly hair. He was sitting at the dining-room table in his sparsely furnished apartment in Manhattan, overlooking the Hudson River. Behind him, a bicycle was leaning against a bare wall, giving the room a post-college feel. Both his parents were astrophysicists, and he, too, was trained as a physicist, before leaving academia for the business world. He grew up in the realm of the abstract and the theoretical – with theorems and calculations and precise measurements. But drug development was different, and when he spoke about the failure of apilimod there was a slight catch in his voice.

Bahcall started to talk about one of the first patients ever treated with

elesclomol: a twenty-four-year-old African-American man. He'd had Kaposi's sarcoma; tumors covered his lower torso. He'd been at Beth Israel Deaconess Medical Center, in Boston, and Bahcall had flown up to see him. On a Monday in January of 2003, Bahcall sat by his bed and they talked. The patient was just out of college. He had an I.V. in his arm. You went to the hospital and you sat next to some kid whose only wish was not to die, and it was impossible not to get emotionally involved. In physics, failure was disappointing. In drug development, failure was heartbreaking. Elesclomol wasn't much help against Kaposi's sarcoma. And now apilimod didn't work for Crohn's. "I mean, we'd done charity work for the Crohn's & Colitis Foundation," Bahcall went on. "I have relatives and friends with Crohn's disease, personal experience with Crohn's disease. We had Crohn's patients come in and talk in meetings and tell their stories. We'd raised money for five years from investors. I felt terrible. Here we were with our lead drug and it had failed. It was the end of the line."

That summer of 2006, in one painful meeting after another, Synta began to downsize. "It was a Wednesday," Bahcall said. "We were around a table, and we were talking about pruning the budget and how we're going to contain costs, one in a series of tough discussions, and I noticed my chief medical officer, Eric Jacobson, at the end of the table, kind of looking a little unusually perky for one of those kinds of discussions." After the meeting, Bahcall pulled Jacobson over: "Is something up?" Jacobson nodded. Half an hour before the meeting, he'd received some news. It was about the melanoma trial for elesclomol, the study everyone had given up on. "The consultant said she had never seen data this good," Jacobson told him.

Bahcall called back the management team for a special meeting. He gave the floor to Jacobson. Eric was, like, 'Well, you know we've got this melanoma trial,' " Bahcall began, "and it took a moment to jog people's memories, because we'd all been so focussed on Crohn's disease and the psoriasis trials. And Eric said, 'Well, we got the results. The drug worked! It was a positive trial!' ' One person slammed the table, stood up, and hollered. Others peppered Eric with questions. "Eric said, 'Well, the group analyzing the data is trying to disprove it, and they can't disprove it.' And he said, 'The consultant handed me the data on Wednesday morning, and she said it was

boinking good.' And everyone said, 'What?' Because Eric is the sweetest guy, who never swears. A bad word cannot cross his lips. Everyone started yelling, 'What? What? What did she say, Eric? Eric! Eric! Say it! Say it!"

Bahcall contacted Synta's board of directors. Two days later, he sent out a company-wide e-mail saying that there would be a meeting that afternoon. At four o'clock, all hundred and thirty employees trooped into the building's lobby. Jacobson stood up. "So the lights go down," Bahcall continued. "Clinical guys, when they present data, tend to do it in a very bottoms-up way: this is the disease population, this is the treatment, and this is the drug, and this is what was randomized, and this is the demographic, and this is the patient pool, and this is who had toenail fungus, and this is who was Jewish. They go on and on and on, and all anyone wants is, Show us the fucking Kaplan-Meier! Finally he said, 'All right, now we can get to the efficacy.' It gets really silent in the room. He clicks the slide. The two lines separate out beautifully – and a gasp goes out, across a hundred and thirty people. Eric starts to continue, and one person goes like this" – Bahcall started clapping slowly – "and then a couple of people joined in, and then soon the whole room is just going like this – clap, clap, clap. There were tears. We all realized that our lives had changed, the lives of patients had changed, the way of treating the disease had changed. In that moment, everyone realized that this little company of a hundred and thirty people had a chance to win. We had a drug that worked, in a disease where nothing worked. That was the single most moving five minutes of all my years at Synta."

In the winter of 1955, a young doctor named Emil Freireich arrived at the National Cancer Institute, in Bethesda, Maryland. He had been drafted into the Army, and had been sent to fulfill his military obligation in the public-health service. He went to see Gordon Zubrod, then the clinical director for the N.C.I. and later one of the major figures in cancer research. "I said, 'I'm a hematologist,'" Freireich recalls. "He said, 'I've got a good idea for you. Cure leukemia.' It was a military assignment." From that assignment came the first great breakthrough in the war against cancer.

Freireich's focus was on the commonest form of childhood leukemia – acute lymphoblastic leukemia (ALL). The diagnosis was a death sentence. "The children would come in bleeding," Freireich says. "They'd have infections.

They would be in pain. Median survival was about eight weeks, and everyone was dead within the year." At the time, three drugs were known to be useful against ALL. One was methotrexate, which, the pediatric pathologist Sidney Farber had shown seven years earlier, could push the disease into remission. Corticosteroids and 6-mercaptopurine (6-MP) had since proved useful. But even though methotrexate and 6-MP could kill a lot of cancer cells, they couldn't kill them all, and those which survived would regroup and adjust and multiply and return with a vengeance. "These remissions were all temporary – two or three months," Freireich, who now directs the adult-leukemia research program at the M. D. Anderson Cancer Center, in Houston, says. "The authorities in hematology didn't even want to use them in children. They felt it just prolonged the agony, made them suffer, and gave them side effects. That was the landscape."

In those years, the medical world had made great strides against tuberculosis, and treating t.b. ran into the same problem as treating cancer: if doctors went after it with one drug, the bacteria eventually developed resistance. Their solution was to use multiple drugs simultaneously that worked in very different ways. Freireich wondered about applying that model to leukemia. Methotrexate worked by disrupting folic-acid uptake, which was crucial in the division of cells; 6-MP shut down the synthesis of purine, which was also critical in cell division. Putting the two together would be like hitting the cancer with a left hook and a right hook. Working with a group that eventually included Tom Frei, of the N.C.I., and James Holland, of the Roswell Park Cancer Institute, in Buffalo, Freireich started treating ALL patients with methotrexate and 6-MP in combination, each at two-thirds its regular dose to keep side effects in check. The remissions grew more frequent. Freireich then added the steroid prednisone, which worked by a mechanism different from that of either 6-MP or methotrexate; he could give it at full dose and not worry about the side effects getting out of control. Now he had a left hook, a right hook, and an uppercut.

"So things are looking good," Freireich went on. "But still everyone dies. The remissions are short. And then out of the blue came the gift from Heaven" – another drug, derived from periwinkle, that had been discovered by Irving Johnson, a researcher at Eli Lilly. "In order to get two milligrams

of drug, it took something like two train-car loads of periwinkle," Freireich said. "It was expensive. But Johnson was persistent." Lilly offered the new drug to Freireich. "Johnson had done work in mice, and he showed me the results. I said, 'Gee whiz, I've got ten kids on the ward dying. I'll give it to them tomorrow.' So I went to Zubrod. He said, 'I don't think it's a good' But I said, 'These kids are dying. What's the difference?' He said, 'O.K., I'll let you do a few children.' The response rate was fifty-five per cent. The kids jumped out of bed." The drug was called vincristine, and, by itself, it was no wonder drug. Like the others, it worked only for a while. But the good news was that it had a unique mechanism of action – it interfered with cell division by binding to what is called the spindle protein – and its side effects were different from those of the other drugs. "So I sat down at my desk one day and I thought, Gee, if I can give 6-MP and meth at two-thirds dose and prednisone at full dose and vincristine has different limiting toxicities, I bet I can give a full dose of that, too. So I devised a trial where we would give all four in combination." The trial was called VAMP. It was a left hook, a right hook, an uppercut, and a jab, and the hope was that if you hit leukemia with that barrage it would never get up off the canvas.

The first patient treated under the experimental regimen was a young girl. Freireich started her off with a dose that turned out to be too high, and she almost died. She was put on antibiotics and a respirator. Freireich saw her eight times a day, sitting at her bedside. She pulled through the chemo-induced crisis, only to die later of an infection. But Freireich was learning. He tinkered with his protocol and started again, with patient No. 2. Her name was Janice. She was fifteen, and her recovery was nothing short of miraculous. So was the recovery of the next patient and the next and the next, until nearly every child was in remission, without need of antibiotics or transfusions. In 1965, Frei and Freireich published one of the most famous articles in the history of oncology, "Progress and Perspective in the Chemotherapy of Acute Leukemia," in *Advances in Chemotherapy.* Almost three decades later, a perfectly healthy Janice graced the cover of the journal *Cancer Research.*

What happened with ALL was a formative experience for an entire generation of cancer fighters. VAMP proved that medicine didn't need a magic

bullet – a superdrug that could stop all cancer in its tracks. A drug that worked a little bit could be combined with another that worked a little bit and another that worked a little bit, and, as long as all three worked in different ways and had different side effects, the combination could turn out to be spectacular. To be valuable, a cancer drug didn't have to be especially effective on its own; it just had to be novel in the way it acted. And, from the beginning, this was what caused so much excitement about elesclomol.

Safi Bahcall's partner in the founding of Synta was a cell biologist at Harvard Medical School named Lan Bo Chen. Chen, who is in his mid-sixties, was born in Taiwan. He is a mischievous man, with short-cropped straight black hair and various quirks – including a willingness to say whatever is on his mind, a skepticism about all things Japanese (the Japanese occupied Taiwan during the war, after all), and a keen interest in the marital prospects of his unattached co-workers. Bahcall, who is Jewish, describes him affectionately as "the best and worst parts of a Jewish father and the best and worst parts of a Jewish mother rolled into one." (Sample e-mail from Chen: "Safi is in Israel. Hope he finds wife.")

Drug hunters like Chen fall into one of two broad schools. The first school, that of "rational design," believes in starting with the disease and working backward – designing a customized solution based on the characteristics of the problem. Herceptin, one of the most important of the new generation of breast-cancer drugs, is a good example. It was based on genetic detective work showing that about a quarter of all breast cancers were caused by the overproduction of a protein called HER2. HER2 kept causing cells to divide and divide, and scientists set about designing a drug to turn HER2 off. The result is a drug that improved survival in twenty-five per cent of patients with advanced breast cancer. (When Herceptin's Kaplan-Meier was shown at ASCO, there was stunned silence.) But working backward to a solution requires a precise understanding of the problem, and cancer remains so mysterious and complex that in most cases scientists don't have that precise understanding. Or they think they do, and then, after they turn off one mechanism, they discover that the tumor has other deadly tricks in reserve.

The other approach is to start with a drug candidate and then hunt for diseases that it might attack. This strategy, known as "mass screening," doesn't

involve a theory. Instead, it involves a random search for matches between treatments and diseases. This was the school to which Chen belonged. In fact, he felt that the main problem with mass screening was that it wasn't mass enough. There were countless companies outside the drug business – from industrial research labs to photography giants like Kodak and Fujifilm – that had millions of chemicals sitting in their vaults. Yet most of these chemicals had never been tested to see if they had potential as drugs. Chen couldn't understand why. If the goal of drug discovery was novelty, shouldn't the hunt for new drugs go as far and wide as possible?

"In the early eighties, I looked into how Merck and Pfizer went about drug discovery," Chen recalls. "How many compounds are they using? Are they doing the best they can? And I come up with an incredible number. It turns out that mankind had, at this point, made tens of millions of compounds. But Pfizer was screening only six hundred thousand compounds, and Merck even fewer, about five hundred thousand. How could they screen for drugs and use only five hundred thousand, when mankind has already made so many more?"

An early financial backer of Chen's was Michael Milken, the junk-bond king of the nineteen-eighties who, after being treated for prostate cancer, became a major cancer philanthropist. "I told Milken my story," Chen said, "and very quickly he said, 'I'm going to give you four million dollars. Do whatever you want.' Right away, Milken thought of Russia. Someone had told him that the Russians had had, for a long time, thousands of chemists in one city making compounds, and none of those compounds had been disclosed." Chen's first purchase was a batch of twenty-two thousand chemicals, gathered from all over Russia and Ukraine. They cost about ten dollars each, and came in tiny glass vials. With his money from Milken, Chen then bought a six-hundred-thousand-dollar state-of-the-art drug-screening machine. It was a big, automated Rube Goldberg contraption that could test ninety-six compounds at a time and do a hundred batches a day. A robotic arm would deposit a few drops of each chemical onto a plate, followed by a clump of cancer cells and a touch of blue dye. The mixture was left to sit for a week, and then rexamined. If the cells were still alive, they would show as blue. If the chemical killed the cancer cells, the fluid would be clear.

Chen's laboratory began by testing his compounds against prostate-cancer cells, since that was the disease Milken had. Later, he screened dozens of other cancer cells as well. In the first go-around, his batch of chemicals killed everything in sight. But plenty of compounds, including pesticides and other sorts of industrial poisons, will kill cancer cells. The trouble is that they'll kill healthy cells as well. Chen was looking for something that was selective – that was more likely to kill malignant cells than normal cells. He was also interested in sensitivity – in a chemical's ability to kill at low concentrations. Chen reduced the amount of each chemical on the plate a thousand-fold, and tried again. Now just one chemical worked. He tried the same chemical on healthy cells. It left them alone. Chen lowered the dose another thousand-fold. It still worked. The compound came from the National Taras Shevchenko University of Kiev. It was an odd little chemical, the laboratory equivalent of a jazz musician's riff. "It was pure chemist's joy," Chen said. "Homemade, random, and clearly made for no particular purpose. It was the only one that worked on everything we tried."

Mass screening wasn't as elegant or as efficient as rational drug design. But it provided a chance of stumbling across something by accident – something so novel and unexpected that no scientist would have dreamed it up. It provided for serendipity, and the history of drug discovery is full of stories of serendipity. Alexander Fleming was looking for something to fight bacteria, but didn't think the answer would be provided by the mold that grew on a petri dish he accidentally left out on his bench. That's where penicillin came from. Pfizer was looking for a new heart treatment and realized that a drug candidate's unexpected side effect was more useful than its main effect. That's where Viagra came from. "The end of surprise would be the end of science," the historian Robert Friedel wrote in the 2001 essay "Serendipity Is No Accident." "To this extent, the scientist must constantly seek and hope for surprises." When Chen gathered chemical compounds from the farthest corners of the earth and tested them against one cancer-cell line after another, he was engineering surprise.

What he found was exactly what he'd hoped for when he started his hunt: something he could never have imagined on his own. When cancer cells came into contact with the chemical, they seemed to go into crisis mode: they

acted as if they had been attacked with a blowtorch. The Ukrainian chemical, elesclomol, worked by gathering up copper from the bloodstream and bringing it into cells' mitochondria, sparking an electrochemical reaction. His focus was on the toxic, oxygen-based compounds in the cell called ROS, reactive oxygen species. Normal cells keep ROS in check. Many kinds of cancer cells, though, generate so much ROS that the cell's ability to keep functioning is stretched to the breaking point, and elesclomol cranked ROS up even further, to the point that the cancer cells went up in flames. Researchers had long known that heating up a cancer cell was a good way of killing it, and there had been plenty of interest over the years in achieving that effect with ROS. But the idea of using copper to set off an electrochemical reaction was so weird – and so unlike the way cancer drugs normally worked – that it's not an approach anyone would have tried by design. That was the serendipity. It took a bit of "chemist's joy," constructed for no particular reason by some bench scientists in Kiev, to show the way. Elesclomol was wondrously novel. "I fell in love," Chen said. "I can't explain it. I just did."

When Freireich went to Zubrod with his idea for VAMP, Zubrod could easily have said no. Drug protocols are typically tested in advance for safety in animal models. This one wasn't. Freireich freely admits that the whole idea of putting together poisonous drugs in such dosages was "insane," and, of course, the first patient in the trial had nearly been killed by the toxic regimen. If she had died from it, the whole trial could have been derailed.

The ALL success story provided a hopeful road map for a generation of cancer fighters. But it also came with a warning: those who pursued the unexpected had to live with unexpected consequences. This was not the elegance of rational drug design, where scientists perfect their strategy in the laboratory before moving into the clinic. Working from the treatment to the disease was an exercise in uncertainty and trial and error.

If you're trying to put together a combination of three or four drugs out of an available pool of dozens, how do you choose which to start with? The number of permutations is vast. And, once you've settled on a combination, how do you administer it? A child gets sick. You treat her. She goes into remission, and then she relapses. VAMP established that the best way to induce remission was to treat the child aggressively when she first showed

up with leukemia. But do you treat during the remission as well, or only when the child relapses? And, if you treat during remission, do you treat as aggressively as you did during remission induction, or at a lower level? Do you use the same drugs in induction as you do in remission and as you do in relapse? How do you give the drugs, sequentially or in combination? At what dose? And how frequently – every day, or do you want to give the child's body a few days to recover between bouts of chemo?

Oncologists compared daily 6-MP plus daily methotrexate with daily 6-MP plus methotrexate every four days. They compared methotrexate followed by 6-MP, 6-MP followed by methotrexate, and both together. They compared prednisone followed by full doses of 6-MP, methotrexate, and a new drug, cyclophosphamide (CTX), with prednisone followed by half doses of 6-MP, methotrexate, and CTX. It was endless: vincristine plus prednisone and then methotrexate every four days or vincristine plus prednisone and then methotrexate daily? They tried new drugs, and different combinations. They tweaked and refined, and gradually pushed the cure rate from forty per cent to eighty-five per cent. At St. Jude Children's Research Hospital, in Memphis – which became a major center of ALL research – no fewer than sixteen clinical trials, enrolling 3,011 children, have been conducted in the past forty-eight years.

And this was just childhood leukemia. Beginning in the nineteen-seventies, Lawrence Einhorn, at Indiana University, pushed cure rates for testicular cancer above eighty per cent with a regimen called BEP: three to four rounds of bleomycin, etoposide, and cisplatin. In the nineteen-seventies, Vincent T. DeVita, at the N.C.I., came up with MOPP for advanced Hodgkin's disease: mustargen, oncovin, procarbazine, and prednisone. DeVita went on to develop a combination therapy for breast cancer called CMF – cyclophosphamide, methotrexate, and 5-fluorouracil. Each combination was a variation on the combination that came before it, tailored to its target through a series of iterations. The often asked question "When will we find a cure for cancer?" implies that there is some kind of master code behind the disease waiting to be cracked. But perhaps there isn't a master code. Perhaps there is only what can be uncovered, one step at a time, through trial and error.

When elesclomol emerged from the laboratory, then, all that was known

about it was that it did something novel to cancer cells in the laboratory. Nobody had any idea what its best target was. So Synta gave elesclomol to an oncologist at Beth Israel in Boston, who began randomly testing it out on his patients in combination with paclitaxel, a standard chemotherapy drug. The addition of elesclomol seemed to shrink the tumor of someone with melanoma. A patient whose advanced ovarian cancer had failed multiple rounds of previous treatment had some response. There was dramatic activity against Kaposi's sarcoma. They could have gone on with Phase 1s indefinitely, of course. Chen wanted to combine elesclomol with radiation therapy, and another group at Synta would later lobby hard to study elesclomol's effects on acute myeloid leukemia (AML), the commonest form of adult leukemia. But they had to draw the line somewhere. Phase 2 would be lung cancer, soft-tissue sarcomas, and melanoma.

Now Synta had its targets. But with this round of testing came an even more difficult question. What's the best way to conduct a test of a drug you barely understand? To complicate matters further, melanoma, the disease that seemed to be the best of the three options, is among the most complicated of all cancers. Sometimes it confines itself to the surface of the skin. Sometimes it invades every organ in the body. Some kinds of melanoma have a mutation involving a gene called BRAF; others don't. Some late-stage melanoma tumors pump out high levels of an enzyme called LDH. Sometimes they pump out only low levels of LDH, and patients with low-LDH tumors lived so much longer that it was as if they had a different disease. Two patients could appear to have identical diagnoses, and then one would be dead in six months and the other would be fine. Tumors sometimes mysteriously disappeared. How did you conduct a drug trial with a disease like this?

It was entirely possible that elesclomol would work in low-LDH patients and not in high-LDH patients, or in high-LDH patients and not in low-LDH ones. It might work well against the melanoma that confined itself to the skin and not against the kind that invaded the liver and other secondary organs; it might work in the early stages of metastasis and not in the later stages. Then, there was the prior-treatment question. Because of how quickly tumors become resistant to drugs, new treatments sometimes work better on "naive" patients – those who haven't been treated with other forms of chemotherapy.

So elesclomol might work on chemo-naive patients and not on prior-chemo patients. And, in any of these situations, elesclomol might work better or worse depending on which other drug or drugs it was combined with. There was no end to the possible combinations of patient populations and drugs that Synta could have explored.

At the same time, Synta had to make sure that whatever trial it ran was as big as possible. With a disease as variable as melanoma, there was always the risk in a small study that what you thought was a positive result was really a matter of spontaneous remissions, and that a negative result was just the bad luck of having patients with an unusually recalcitrant form of the disease. John Kirkwood, a melanoma specialist at the University of Pittsburgh, had done the math: in order to guard against some lucky or unlucky artifact, the treatment arm of a Phase 2 trial should have at least seventy patients.

Synta was faced with a dilemma. Given melanoma's variability, the company would ideally have done half a dozen or more versions of its Phase 2 trial: low-LDH, high-LDH, early-stage, late-stage, prior-chemo, chemo-nave, multi-drug, single-drug. There was no way, though, that they could afford to do that many trials with seventy patients in each treatment arm. The American biotech industry is made up of lots of companies like Synta, because small start-ups are believed to be more innovative and adventurous than big pharmaceutical houses. But not even big firms can do multiple Phase 2 trials on a single disease – not when trials cost more than a hundred thousand dollars per patient and not when, in the pursuit of serendipity, they are simultaneously testing that same experimental drug on two or three other kinds of cancer. So Synta compromised. The company settled on one melanoma trial: fifty-three patients were given elesclomol plus paclitaxel, and twenty-eight, in the control group, were given paclitaxel alone, representing every sort of LDH level, stage of disease, and prior-treatment status. That's a long way from half a dozen trials of seventy each.

Synta then went to Phase 3: six hundred and fifty-one chemo-naive patients, drawn from a hundred and fifty hospitals, in fifteen countries. The trial was dubbed SYMMETRY. It was funded by the pharmaceutical giant GlaxoSmithKline. Glaxo agreed to underwrite the cost of the next round of clinical trials and – should the drug be approved by the Food and Drug

Administration – to split the revenues with Synta.

But was this the perfect trial? Not really. In the Phase 2 trial, elesclomol had been mixed with an organic solvent called Cremophore and then spun around in a sonicator, which is like a mini washing machine. Elesclomol, which is rock-hard in its crystalline form, needed to be completely dissolved if it was going to work as a drug. For SYMMETRY, though, sonicators couldn't be used. "Many countries said that it would be difficult, and some hospitals even said, 'We don't allow sonication in the preparation room,'" Chen explained. "We got all kinds of unbelievable feedback. In the end, we came up with something that, after mixing, you use your hand to shake it." Would hand shaking be a problem? No one knew.

Then a Synta chemist, Mitsunori Ono, figured out how to make a water-soluble version of elesclomol. When the head of Synta's chemistry team presented the results, he "sang a Japanese drinking song," Chen said, permitting himself a small smile at the eccentricities of the Japanese. "He was very happy." It was a great accomplishment. The water-soluble version could be given in higher doses. Should they stop SYMMETRY and start again with elesclomol 2.0? They couldn't. A new trial would cost many millions of dollars more, and set the whole effort back two or three years. So they went ahead with a drug that didn't dissolve easily, against a difficult target, with an assortment of patients who may or may not have been ideal – and crossed their fingers.

SYMMETRY began in late 2007. It was a double-blind, randomized trial. No one had any idea who was getting elesclomol and who wasn't, and no one would have any idea how well the patients on elesclomol were doing until the trial data were unblinded. Day-to-day management of the study was shared with a third-party contractor. The trial itself was supervised by an outside group, known as a data-monitoring committee. "We send them all the data in some database format, and they plug that into their software package and then they type in the code and press 'Enter,'" Bahcall said. "And then this line" – he pointed at the Kaplan-Meier in front of him – "will, hopefully, separate into two lines. They will find out in thirty seconds. It's, literally, those guys press a button and for the next five years, ten years, the life of the drug, that's really the only bit of evidence that matters." It was January,

2009, and the last of the six hundred and fifty-one patients were scheduled to be enrolled in the trial in the next few weeks. According to protocol, when the results began to come in, the data-monitoring committee would call Jacobson, and Jacobson would call Bahcall. "ASCO starts May 29th," Bahcall said. "If we get our data by early May, we could present at ASCO this year."

In the course of the SYMMETRY trial, Bahcall's dining-room-table talks grew more reflective. He drew Kaplan-Meiers on the back of napkins. He talked about the twists and turns that other biotech companies had encountered on the road to the marketplace. He told wry stories about Lan Bo Chen, the Jewish mother and Jewish father rolled into one – and, over and over, he brought up the name of Judah Folkman. Folkman died in 2008, and he was a legend. He was the father of angiogenesis – a wholly new way of attacking cancer tumors. Avastin, the drug that everyone cheered at ASCO seven years ago, was the result of Folkman's work.

Folkman's great breakthrough had come while he was working with mouse melanoma cells at the National Naval Medical Center: when the tumors couldn't set up a network of blood vessels to feed themselves, they would stop growing. Folkman realized that the body must have its own system for promoting and halting blood-vessel formation, and that if he could find a substance that prevented vessels from being formed he would have a potentially powerful cancer drug. One of the researchers in Folkman's laboratory, Michael O'Reilly, found what seemed to be a potent inhibitor: angiostatin. O'Reilly then assembled a group of mice with an aggressive lung cancer, and treated half with a saline solution and half with angiostatin. In the book "Dr. Folkman's War" (2001), Robert Cooke describes the climactic moment when the results of the experiment came in:

With a horde of excited researchers jam-packed into a small laboratory room, Folkman euthanized all fifteen mice, then began handing them one by one to O'Reilly to dissect. O'Reilly took the first mouse, made an incision in its chest, and removed the lung. The organ was overwhelmed by cancer. Folkman checked a notebook to see which group the mouse had been in. It was one of those that had gotten only saline. O'Reilly cut into the next mouse and removed its lung. It was perfect. What treatment had it gotten?

The notebook revealed it was angiostatin.

It wasn't Folkman's triumph that Bahcall kept coming back to, however. It was his struggle. Folkman's great insight at the Naval Medical Center occurred in 1960. O'Reilly's breakthrough experiment occurred in 1994. In the intervening years, Folkman's work was dismissed and attacked, and confronted with every obstacle.

At times, Bahcall tried to convince himself that elesclomol's path might be different. Synta had those exciting Phase 2 results, and the endorsement of the Glaxo deal. "For the results not to be real, you'd have to believe that it was just a statistical fluke that the patients who got drugs are getting better," Bahcall said, in one of those dining-room-table moments. "You'd have to believe that the fact that there were more responses in the treatment group was also a statistical fluke, along with the fact that we've seen these signs of activity in Phase 1, and the fact that the underlying biology strongly says that we have an extremely active anti-cancer agent."

But then he would remember Folkman. Angiostatin and a companion agent also identified by Folkman's laboratory, endostatin, were licensed by a biotech company called EntreMed. And EntreMed never made a dime off either drug. The two drugs failed to show any clinical effects in both Phase 1 and Phase 2. Avastin was a completely different anti-angiogenesis agent, discovered and developed by another team entirely, and brought to market a decade after O'Reilly's experiment. What's more, Avastin's colorectal-cancer trial – the one that received a standing ovation at ASCO – was the drug's second go-around. A previous Phase 3 trial, for breast cancer, had been a crushing failure. Even Folkman's beautifully elaborated theory about angiogenesis may not fully explain the way Avastin works. In addition to cutting off the flow of blood vessels to the tumor, Avastin seems also to work by repairing some of the blood vessels feeding the tumor, so that the drugs administered in combination with Avastin can get to the tumor more efficiently.

Bahcall followed the fortunes of other biotech companies the way a teenage boy follows baseball statistics, and he knew that nothing ever went smoothly. He could list, one by one, all the breakthrough drugs that had failed their first Phase 3 or had failed multiple Phase 2s, or that turned out

not to work the way they were supposed to work. In the world of serendipity and of trial and error, failure was a condition of discovery, because, when something was new and worked in ways that no one quite understood, every bit of knowledge had to be learned, one experiment at a time. You ended up with VAMP, which worked, but only after you compared daily 6-MP and daily methotrexate with daily 6-MP and methotrexate every four days, and so on, through a great many iterations, none of which worked very well at all. You had results that looked "boinking good," but only after a trial with a hundred compromises.

Chen had the same combination of realism and idealism that Bahcall did. He was the in-house skeptic at Synta. He was the one who worried the most about the hand shaking of the drugs in the SYMMETRY trial. He had never been comfortable with the big push behind melanoma. "Everyone at Dana-Farber" – the cancer hospital at Harvard – "told me, 'Don't touch melanoma,'" Chen said. "'It is so hard. Maybe you save it as the last, after you have already treated and tried everything else.'" The scientists at Synta were getting better and better at understanding just what it was that elesclomol did when it confronted a cancer cell. But he knew that there was always a gap between what could be learned in the laboratory and what happened in the clinic. "We just don't know what happens in vivo," he said. "That's why drug development is still so hard and so expensive, because the human body is such a black box. We are totally shooting in the dark." He shrugged. "You have to have good science, sure. But once you shoot the drug in humans you go home and pray."

Chen was sitting in the room at Synta where Eric Jacobson had revealed the "boinking good" news about elesclomol's Phase 2 melanoma study. Down the hall was a huge walk-in freezer, filled with thousands of chemicals from the Russian haul. In another room was the Rube Goldberg drug-screening machine, bought with Milken's money. Chen began to talk about elesclomol's earliest days, when he was still scavenging through the libraries of chemical companies for leads and Bahcall was still an ex-physicist looking to start a biotech company. "I could not convince anyone that elesclomol had potential," Chen went on. "Everyone around me tried to stop it, including my research partner, who is a Nobel laureate. He just hated it." At one

point, Chen was working with Fujifilm. The people there hated elesclomol. He worked for a while for the Japanese chemical company Shionogi. The Japanese hated it. "But you know who I found who believed in it?" Chen's eyes lit up: "Safi!"

Last year, on February 25th, Bahcall and Chen were at a Synta board meeting in midtown Manhattan. It was five-thirty in the afternoon. As the meeting was breaking up, Bahcall got a call on his cell phone. "I have to take this," he said to Chen. He ducked into a nearby conference room, and Chen waited for him, with the company's chairman, Keith Gollust. Fifteen minutes passed, then twenty. "I tell Keith it must be the data-monitoring committee," Chen recalls. "He says, 'No way. Too soon. How could the D.M.C. have any news just yet?' I said, 'It has to be.' So he stays with me and we wait. Another twenty minutes. Finally Safi comes out, and I looked at him and I knew. He didn't have to say anything. It was the color of his face."

The call had been from Eric Jacobson. He had just come back from Florida, where he had met with the D.M.C. on the SYMMETRY trial. The results of the trial had been unblinded. Jacobson had spent the last several days going over the data, trying to answer every question and double-check every conclusion. "I have some really bad news," he told Bahcall. The trial would have to be halted: more people were dying in the treatment arm than in the control arm. "It took me about a half hour to come out of primary shock," Bahcall said. "I didn't go home. I just grabbed my bag, got into a cab, went straight to LaGuardia, took the next flight to Logan, drove straight to the office. The chief medical officer, the clinical guys, statistical guys, operational team were all there, and we essentially spent the rest of the night, until about one or two in the morning, reviewing the data." It looked as if patients with high-LDH tumors were the problem: elesclomol seemed to fail them completely. It was heartbreaking. Glaxo, Bahcall knew, was certain to pull out of the deal. There would have to be many layoffs.

The next day, Bahcall called a meeting of the management team. They met in the Synta conference room. "Eric has some news," Bahcall said. Jacobson stood up and began. But before he got very far he had to stop, because he was overcome with emotion, and soon everyone else in the room

was, too.

On December 7, 2009, Synta released the following statement:

Synta Pharmaceuticals Corp. (NASDAQ: SNTA), a biopharmaceutical company focused on discovering, developing, and commercializing small molecule drugs to treat severe medical conditions, today announced the results of a study evaluating the activity of elesclomol against acute myeloid leukemia (AML) cell lines and primary leukemic blast cells from AML patients, presented at the Annual Meeting of the American Society of Hematology (ASH) in New Orleans. "The experiments conducted at the University of Toronto showed elesclomol was highly active against AML cell lines and primary blast cells from AML patients at concentrations substantially lower than those already achieved in cancer patients in clinical trials," said Vojo Vukovic, M.D., Ph.D., Senior Vice President and Chief Medical Officer, Synta. "Of particular interest were the *ex vivo* studies of primary AML blast cells from patients recently treated at Toronto, where all 10 samples of leukemic cells responded to exposure to elesclomol. These results provide a strong rationale for further exploring the potential of elesclomol in AML, a disease with high medical need and limited options for patients."

"I will bet anything I have, with anybody, that this will be a drug one day," Chen said. It was January. The early AML results had just come in. Glaxo was a memory. "Now, maybe we are crazy, we are romantic. But this kind of characteristic you have to have if you want to be a drug hunter. You have to be optimistic, you have to have supreme confidence, because the odds are so incredibly against you. I am a scientist. I just hope that I would be so romantic that I become deluded enough to keep hoping."

### 0.0.21 The Ghost's Vocabulary: How the Computer Listens for Shakespeare's "Voiceprint", Edward Dolnick (*The Atlantic*), October, 1991

October, 1991

Edward Dolnick (*The Atlantic*)

In 1842 literature and science met with a thud. Alfred Tennyson had just published his poem "The Vision of Sin." Among the appreciative letters he received was one from Charles Babbage, the mathematician and inventor who is known today as the father of the computer. Babbage wrote to suggest a correction to Tennyson's "otherwise beautiful" poem – in particular to the lines "Every moment dies a man, Every moment one is born."

"It must be manifest," Babbage pointed out, "that, were this true, the population of the world would be at a standstill." Since the population was in fact growing slightly, Babbage continued, "I would suggest that in the next edition of your poem you have it read: 'Every moment dies a man, Every moment 1-1/16 is born.'" Even this was not strictly correct, Babbage conceded, "but I believe 1-1/16 will be sufficiently accurate for poetry."

Today computers are standard tools for amateur and professional literary investigators alike. Shakespeare is both the most celebrated object of this effort and the most common. At Claremont McKenna College, in California, for example, two highly regarded faculty members have devoted years of their lives to a computer-based attempt to find out whether Shakespeare, rather than Francis Bacon or the Earl of Oxford or any of a myriad of others, wrote the plays and poems we associate with his name.

As Babbage's venture into criticism foreshadowed, the marriage of computers and literature has been an uneasy one. At the mention of computers or statistics, many Shakespeareans and others in the literary establishment wrinkle their noses in distaste. To approach the glories of literature in this plodding way is misguided, they say, and misses the point in the same way as does the oft-cited remark that the human body is worth just a few dollars – the market value of the various chemicals of which it is composed. "This is just madness," says Ricardo Quinones, the chairman of the literature department at Claremont McKenna. "Why don't they simply read the plays?"

Rather than read, these literary sleuths prefer to count. Their strategy is straightforward. Most are in search of a statistical fingerprint, a reliable and objective mark of identity unique to a given author. Every writer will sooner or later reveal himself, they contend, by quirks of style that may be too subtle for the eye to note but are well within the computer's power to identify.

For a University of Chicago statistician named Ronald Thisted, the call to enter this quasi-literary enterprise came on a Sunday morning in December of 1985. Thisted had settled down with *New York Times Book Review* and an article by Gary Taylor, a Shakespeare scholar, caught his eye. Taylor claimed that he had found a new poem by Shakespeare at Oxford's Bodleian Library. Among the many reasons Taylor advanced for believing in the authenticity of the poem, called "Shall I Die?," Thisted focused on one. "One of his arguments," Thisted says, "was that several words in the poem don't appear previously in Shakespeare. And that was evidence that Shakespeare wrote it. One's first reaction is, that's dumb. If Shakespeare didn't use these words, why would that be evidence that he wrote the poem?" But Taylor's article went on to explain that in practically everything he wrote, Shakespeare used words he hadn't used elsewhere. Thisted conceded the point in his own mind, but raised another objection. "If ALL the words in there were ones that Shakespeare had never used," he thought, "if it were in Sanskrit or something, you'd say, 'No way Shakespeare could have written this.' So there had to be about the right number of new words." That question – how many new words an authentic Shakespeare text should contain – was similar to one that Thisted himself had taken on a decade before. Together with the Stanford statistician Bradley Efron, then his graduate adviser, Thisted had published a paper that proposed a precise answer to the question "How many words did Shakespeare know but never use?" The question sounds ludicrous, like "How many New Year's resolutions have I not yet made?" Nonetheless, Efron and Thisted managed to answer it. They found the crucial insight in a generation-old story, perhaps apocryphal, about an encounter between a mathematician and a butterfly collector.

R. A. Fisher, the statistical guru of his day, had been consulted by a butterfly hunter newly back from Malaysia. The naturalist had caught members

of some species once or twice, other species several times, and some species time and time again. Was it worth the expense, the butterfly collector asked, to go back to Malaysia for another season's trapping? Fisher recast the question as a mathematical problem. The collector knew how many species he had seen exactly once, exactly twice, and so on. Now, how many species were out there that he had yet to see? If the collector had many butterflies from each species he had seen, Fisher reasoned, then quite likely he had sampled all the species that were out there. Another hunting trip would be superfluous. But if he had only one or two representatives of most species, then there might be many species yet to find. It would be worth returning to Malaysia. Fisher devised a mathematical way to make that rough idea precise (and reportedly suggested another collecting trip). Efron and Thisted's question was essentially the same.

Where the naturalist had tramped through the rain forest in search of exotic butterflies, the mathematicians could scan Shakespeare in search of unusual words. By counting how many words he used exactly once, exactly twice, and so on, they would attempt to calculate how many words he knew but had yet to use.

Neither Efron nor Thisted had imagined that their statistical sleight of hand could ever be put to a live test. No new work of Shakespeare's had been unearthed for decades. Now Taylor had given them their chance. A new Shakespeare poem, like a new butterfly-collecting trip to the jungle, should yield a certain number of new words, a certain number that Shakespeare had used once before, and so on. If Shakespeare did write "Shall I Die?," which has 429 words, according to the mathematicians' calculations it should have about seven words he never used elsewhere; it has nine. To Efron and Thisted's surprise, the number of words in the poem which Shakespeare had used once before also came close to matching their predictions, as did the number of twice-used words, all the way through to words he had used ninety-nine times before. The poem, which sounds nothing like Shakespeare, fit Shakespeare like a glove.

This is work that can suck up lives. One Defoe scholar, trying to pick out true Defoe from a slew of anonymous and pseudonymous works, has pursued his quarry for twenty years, with no end in sight. A team trying to

determine if the *Book of Mormon* was composed by ancient authors or by the nineteenth-century American Joseph Smith took 10,000 hours to produce a single essay. (The largely Mormon team of researchers concluded that Smith had not written the *Book of Mormon*. Confirmed samples of Smith's prose, the researchers argued, showed patterns of word usage different from those in the *Book of Mormon*.) Paper after paper begins with a trumpet fanfare and ends with a plaintive bleat. One writer, for instance, decided to determine whether Jonathan Swift or one of his contemporaries had written a particular article, by pigeonholing his words according to what part of speech they were. "The only positive conclusion from over a year of effort and the coding of over 40,000 words," she lamented, "is that a great deal of further study will be needed." (Swift himself had satirized, in Gulliver's Travels, a professor who had "employed all his Thoughts from his Youth" in making "the strictest Computation of the general Proportion there is in Books between the Numbers of Particles, Nouns, and Verbs, and other Parts of Speech.")

Despite the shortage of triumphs the field is growing, because more and more of the work can be assigned to electronic drudges. Scholars once had to count words by hand. Later they had the option of typing entire books into a computer, so that the machine could do the counting. Today computers are everywhere, and whole libraries of machine-readable texts are available. Software to do deluxe slicing and dicing is easy to obtain.

As a result, everything imaginable is being counted somewhere. Someone at this moment is tallying up commas or meticulously computing adjective-to-adverb ratios. But sophisticated tools don't automatically produce good work. A future Academy of Statistics and Style might take as its motto the warning that the Nobel laureate P. B. Medawar issued to his fellow scientists: "An experiment not worth doing is not worth doing well."

Among those least likely to be fazed by such pronouncements is a professor of political science at Claremont McKenna College named Ward Elliott. Elliott is an authority on voting rights, a cheerful eccentric, and, like his father before him, inclined to view the Earl of Oxford as the true author of Shakespeare's works. Four years ago Elliott recruited Robert Valenza, an expert programmer also on the Claremont McKenna faculty, and the two set

to work on the authorship question.

This time the model would be not butterfly hunting but radar. Valenza had spent considerable time devising mathematical procedures to find the patterns obscured by noisy and jumbled electronic signals. Adapted to Shakespeare, the idea was to go beyond counting various words, as many others had done, and see whether consistent patterns could be found in the way certain key words were used together. Two writers might use the words "blue" and Green equally often throughout a text, for example, but the writers could be distinguished if one always used them on the same page while the other never used them together.

This pattern-finding mathematics is widely used in physics and engineering, in deciphering television and radar signals, for example. Given a long list of words – not simply the "blue" and Green of the example, but dozens more – the computer can quickly tell how Shakespeare typically balanced those words. "You might have a pattern," Valenza says, "with a lot of 'love,' very little 'hate,' and a good deal of 'woe.'" A different writer might use the same words, and even use them at the same rates as Shakespeare, but the configurations might be different. The result is that a given list of words produces a kind of voiceprint for each author.

Valenza and Elliott examined "common but not too common" words that Shakespeare used. To examine rare words, Valenza had reasoned, would be like trying to identify a voice from a whisper, and to examine common words would be to let the writer shout into your ear. The final, fifty-two-word list – with such miscellaneous entries as "about," "death," "desire," "secret," and "set" – was assembled by trial and error. It consisted of words with two key properties. In various works of Shakespeare's those words are used in patterns that yield the same voiceprint each time. And when other writers are tested, the same words yield voiceprints that are different from Shakespeare's.

The machinery in place, Valenza and Elliott began by testing Shakespeare's poetry against that of thirty other writers. Exciting results came quickly: The disputed "Shall I Die?" poem seemed not to be Shakespeare's after all. Three of the leading claimants to Shakespeare's work – Francis Bacon, Christopher Marlowe, and Sir Edward Dyer – were decisively ruled out. To Elliott's good-humored consternation, the test dealt just as harshly with

the claims put forward on behalf of the Earl of Oxford. Worse was to follow. For even as this first round of tests ruled out the best-known Shakespeare candidates, it left a few surprising contenders. One possibility for the "real" Shakespeare: Queen Elizabeth I. "That did it for our chance of appearing in *Science*," Elliott laments, "But it vastly increased our chance of getting into the *National Enquirer*." (To his dismay, Elliott did find himself in *Science*, not as the co-author of a weighty research paper but as the subject of a skeptical news brief with the headline "Did Queen Write Shakespeare's Sonnets?")

Valenza and Elliott have since conducted more extensive tests that have ruled out Queen Elizabeth. But the mishap highlights a risk that is shared by all the number-crunching methods. "If the glass slipper doesn't fit, it's pretty good evidence that you're not Cinderella," Elliott points out. "But if it does fit, that doesn't prove that you are."

The risk of being fooled is least for someone who combines a deep knowledge of literature with some statistical insight. Donald Foster, a professor of English at Vassar College, fits that bill. Foster's scholarship is highly regarded. Soon after "Shall I Die?" was presented to the world, for example, he wrote a long debunking essay that persuaded many readers that the poem was not Shakespeare's. In a more recent essay he consigned whole libraries of research to the scrap heap. Hundreds or thousands of articles have been written to explain the epigraph to Shakespeare's Sonnets, which begins, "To the onlie begetter of these insuing sonnets, Master W.H." Who was W.H.? Foster's solution to the mystery, which won him the Modern Language Association's Parker Prize, is that W.H. was ... a typo. The publisher, who wrote the epigraph as a bit of flowery praise to honor Shakespeare, had intended to print "W.SH."

Those essays had nothing to do with statistics, but Foster has done some statistical sleuthing of his own, and he is well aware of the hazards. One scholar compared Shakespeare's plays with someone else's poems, for example, and concluded that Shakespeare used the present tense more than other writers do. Another compared Shakespeare with later writers and concluded that he used many four-letter words, whereas other writers used shorter words – forgetting that archaic words like "thou" and "hath" drive Shakespeare's

average up. "There are strong and compelling reasons for avoiding this kind of research," Foster says, "because it's so difficult to anticipate all the pitfalls." But Foster himself has often given way to temptation. Like many Shakespeareans, he steers clear of the "authorship question," but he has looked into a pertinent mystery.

Shakespeare acted in his plays. But with two exceptions, we don't know what roles he took. Foster believes he has found a statistical way to gather that long-vanished knowledge. "It occurred to me," he says, "that Shakespeare may have been influenced in his writing by the parts he had memorized for performances and was reciting on a more or less daily basis." Last year Foster figured out a way to test that hunch. "The results," he says, "have been absolutely stunning."

"We started by using a concordance to type in all the words that Shakespeare used ten times or fewer," Foster says. These aren't exotic words, necessarily, just ones that don't crop up often in Shakespeare. Scholars have known for some time that these "rare" words tend to be clustered chronologically. Foster found that if two plays shared a considerable number of rare words, in the later play those words were scattered randomly among all the characters. In the earlier play, the shared words were not scattered. "In one role," Foster says, "there would be two to six times the expected number of rare words." There stood Shakespeare: the words that Shakespeare the writer had at the tip of his pen were the ones he had been reciting as Shakespeare the actor.

If Foster is right, Shakespeare played Theseus in *A Midsummer Night's Dream* and "Chorus" in *Henry V* and *Romeo and Juliet*. In play after play the first character to come on stage and speak is the one that Foster's test identifies as Shakespeare: John Gower in *Pericles*, Bedford in *Henry VI, Part I*, Suffolk in *Henry VI, Part II*, and Warwick in *Henry VI, Part III*. And Foster's test picks out as Shakespeare's the two roles that we have seventeenth-century evidence he played: the ghost in *Hamlet* and Adam in *As You Like It*.

The theory can be tested in other ways. It never assigns to Shakespeare a role we know another actor took. The roles it does label as Shakespeare's all seem plausible – male characters rather than women or children. The test

never runs in the wrong direction, with the unusual words scattered randomly in an early play and clustered in one role in a later play. On those occasions when Foster's test indicates that Shakespeare played TWO roles in a given play – Gaunt and a gardener in *Richard II*, for example – the characters are never onstage together. Foster's theory passes another test. When Foster looks at the rare words that *Hamlet* shares with *Macbeth*, written a few years later, those words point to the ghost in *Hamlet* as Shakespeare's role. And if Foster looks at rare words that Hamlet shares with a different play also written a few years later – *King Lear*, for example – those shared words also pick out the ghost as Shakespeare's role.

Additional evidence has been uncovered. After *Hamlet*, the ghost's vocabulary exerted a strong influence on Shakespeare's writing and then tapered off. But Shakespeare's plays went in and out of production. When *Hamlet* was revived several years after its first staging, and Shakespeare was again playing the ghost, he began again to recycle the ghost's vocabulary.

It is a strange image, a computer fingering a ghost. But it is a sign of things to come. Eventually the prejudice against computers in literary studies will give way. "The walls are sure to crumble," Ward Elliott says, "just as they did in baseball and popular music ... Some high-tech Jackie Robinson will score a lot of runs, and thereafter all the teams in the league will pursue the newest touch as ardently and piously as they now shrink from it."

### 0.0.22 Influence of Funding Source on Outcome, Validity, and Reliability of Pharmaceutical Research, Report 10 of the Council on Scientific Affairs of the American Medical Association

Report 10 of the Council on Scientific Affairs of the American Medical Association — Note: This report represents the medical/scientific literature on the subject as of June 2004.

Background of report:

Resolution 514 (A-03), introduced by the American Psychiatric Association and the American Academy of Child and Adolescent Psychiatry and referred to the American Medical Association (AMA) Board of Trustees, asked that the Council on Scientific Affairs (1) study the impact of funding sources on the outcome, validity, and reliability of pharmaceutical research; and (2) develop guidelines to assist physician-researchers in evaluating and preserving the scientific integrity, validity, and reliability of pharmaceutical research, regardless of funding source.

Considerable research has been conducted on the issues raised in Resolution 514, and systematic reviews on the subject have been recently published. This report summarizes the findings of these reviews, updates new information, provides some perspectives on the topic, and offers some recommendations on how the AMA can continue to assist in improving the scientific integrity, validity, and reliability of pharmaceutical research. JAMA and other major-impact medical journals also have recently taken steps to address the problem of publication bias and to properly identify the conflicts of interest that inevitably emerge with the sponsorship of a market-driven enterprise such as prescription drug approval.

Methods: Literature searches were conducted in the MEDLINE and LexisNexis databases for English-language review articles published between 1985 and 2003 using the search terms clinical trials; drug industry; financing, organized; publishing; and research or research support, in combination with economics or standards. This search identified 12 systematic reviews on the relationship between pharmaceutical industry sponsorship and research outcome, quality, or publication bias. Three of these reviews evaluated pre-

viously published information on the relationship between industry funding and outcome or quality; more than 2000 original studies were included in the original studies covered by these reviews. The findings of these 12 systematic reviews form the basis for discussion in this report. Recently published articles not covered in these systematic reviews and other original studies and editorials relevant to related issues also were analyzed. In addition, the draft report was offered to the AMA's Council on Ethical and Judicial Affairs for its review and contribution to the discussion in the "Potential Remedies" section (see below).

Introduction:

The results (and analysis) of clinical research that are published in peer-reviewed journals inform most treatment decisions, and influence public and private health care policy. A longstanding concern exists about the potential for publication bias in pharmaceutical research. Publication bias is the selective publication of studies based on the direction (positive), magnitude, and statistical significance of the treatment effect. Publication bias is often attributed to decisions made by author/investigators and journal editors, but in fact can intrude during the entire process of planning and conducting the clinical trial and publishing the results, leading to outcome bias.

Studies with positive findings are more likely to be published than studies with negative or null results and an association exists between pharmaceutical industry sponsorship of clinical research and publication of results favoring the sponsor's products. Additionally, the publication of negative results may be delayed compared with the time to publication of studies with positive results. This relative time lag is not limited to industry-sponsored trials.

This pattern of publication distorts the medical literature, thereby affecting the validity and findings of systematic reviews and meta-analyses, the decisions of funding agencies, and ultimately the optimal practice of medicine. However, without pharmaceutical industry sponsorship, important therapeutic advances would not have occurred. Modern drug therapy has changed the clinical landscape and represents a cost-effective intervention for disease prevention and treatment. Productive collaborations between the pharmaceutical industry and academic medicine or other research organizations have flourished over the last 30 years enabling a steady stream of new and inno-

vative treatments to improve patient care. In 2002, total grant spending for clinical trials involving human subjects was approximately $5.6 billion, with more than 70% provided by the biopharmaceutical industry. If device manufacturers are included, the total fraction of grant support rises to 80%, with the remainder ($1.1 billion) supplied primarily by the National Institutes of Health (NIH). In 2002, approximately 50,000 clinical investigators received funding for at least one clinical trial conducted in the United States.

Publication bias involving industry-sponsored trials may be exacerbated because many drug trials are conducted to gain regulatory approval, not to test a novel scientific hypothesis or to examine relative therapeutic efficacy versus another treatment. Much of the clinical trial information is unpublished at the time of marketing. Physicians would like to know how one drug compares with other available therapeutic options, and the health care system wants to take cost utility into account, but such information is usually not available initially and may never become available. These deficiencies raise broader issues related to drug approval and marketing.

Sources of publication bias:

Investigators and authors. One direct source of publication bias is the failure of investigators to submit completed research for publication. Quantitative studies with significant results are more likely to be submitted, and studies with positive results are submitted more rapidly. Given available time and resources, some investigators are unwilling to submit studies with null or unimportant results because they believe the manuscript will be rejected anyway.

Also, investigators (including academic faculty who receive industry funding) may be subject to prepublication review or restricted access to data. Cross-sectional surveys indicate that industry-sponsored faculty are more likely to report restrictions on the dissemination of their research findings and to be denied access to the entire data. Delayed publication may occur because of clinical trial agreements that require information to remain confidential for an extended period to enable patent protection, protect a corporate "lead" in the field, or resolve disputes over intellectual property. Another behavior of academic faculty that is associated with pharmaceutical sponsorship or the presence of commercial agreements is refusal to share

information with colleagues.

Because pharmaceutical companies now exert more direct control and ownership of the clinical trial process, published reports also may contain declared authors who have not participated in the design or interpretation of the study, with only limited access to the original data.

Journal editors and reviewers. As mentioned above, authors may fail to submit negative studies because of fear of rejection by journal editors. The ideal articles for journals are those with findings that will affect clinical practice. Indeed, instructions to authors may contain phrases such as the journal "gives priority to reports of original research that are likely to change clinical practice or thinking about a disease." Under this rubric, even confirmatory trials (positive or negative) would achieve lower priority. In terms of external advice received by journal editors, selected referees may demonstrate bias against papers with results that are contrary to their own perspectives. This type of confirmatory bias, in which the referees' judgment is colored by their own preconceptions and experience, contributes to poor interrater reliability. In a study of articles submitted to JAMA from 1996–1999 that reported results of controlled trials involving an intervention and comparison group, the odds ratio for publishing studies with positive results was 1.30. This difference in publication rates between those with positive and negative results was not statistically significant, suggesting publication bias was not evident in editorial decision-making.

Clinical trial agreements. A large percentage of companies providing support for clinical research obtain patents and market products as a result of this relationship. One survey revealed that 53 percent of signed clinical trial agreements in a sample of university-industry research centers allowed publication to be delayed, 35 percent allowed the sponsor to delete information from the publication, and 30% allowed both. Commercial interests and intellectual property are often the main reasons for lack of publication of clinical studies funded by the pharmaceutical industry. In the last decade, some high-profile cases of direct interference have been noted.

Nevertheless, industry's dependence on academia has lessened in the last decade with a shift to contract research organizations (CROs), site management organizations, and commercial drug trial networks. There are several

reasons for this development, including changes in the pharmaceutical industry itself (i.e., employment of competent researchers to design, run, and interpret trials) and the fact that CROs have been able to use community physicians as a reliable source of patient enrollees, can run trials less expensively and more efficiently than academic medical centers, and generally impose a less cumbersome trial agreement process.

Outcome bias. When control lies with the commercial rather than academic or public sector, bias can also envelope the process through the trial design. Outcome bias can result from the use of unreliable methods or instruments, as well as inadequate sample size or comparison groups. Favorable results are more likely if: (1) the drug is tested in a healthier population (i.e., younger, fewer comorbidities, milder disease) that is not representative of the actual patient population that will take the drug; (2) the drug is compared with an insufficient dose of an alternate product; or (3) multiple surrogate endpoints (which may not correlate with more important clinical endpoints) are studied but only results favoring the product are published. Industry-funded studies are also much more likely to use placebos or inactive controls, a practice that increases the likelihood of achieving positive study results.

Funding source and outcome:

For reasons cited above, growing concern exists about the influence that industry sponsorship exerts on clinical trial design and outcome, academic freedom, transparency in the research process, and ultimately, the public good. With regard to the extent of financial relationships, at least $1.5 billion flows from industry to academia annually, so a significant percentage (at least 25 percent) of academic investigators receive industry funding for their biomedical research, and at least one-third have personal financial ties with industry sponsors. Correspondingly, many universities may hold equity in the sponsor's business (as well as accept funding).

Because of these financial relationships, questions have been raised about whether such financial relationships could create institutional conflicts of interest involving research integrity and/or the safety and welfare of human subjects. Because of the increasing trend for faculty researchers to be involved in financial relationships with their research sponsors, Boyd and Bero concluded that "guidelines for what constitutes a conflict and how the conflict

should be managed are needed if researchers are to have consistent standards of behavior among institutions." In keeping with this recommendation, the Association of American Medical Colleges (AAMC) formed a task force on conflicts of interest in clinical research. In 2002, the task force released a report that provided guidance on addressing the financial interests of faculty, as well as other individuals involved in the clinical research enterprise.

Funding source and publication outcome or status. Results of three recent systematic reviews confirm that industry-sponsored research tends to yield pro-industry conclusions. Even among comparison trials, the sponsor's drug is almost always deemed equivalent or superior to the comparator.

The favorable relationship between funding source and outcome extends to pharmacoeconomic studies and sponsored meta-analyses. Authors who have financial relationships with pharmaceutical companies may be more likely to produce proindustry conclusions. This fact and other evidence support the notion that conclusions are generally more positive in trials funded by the pharmaceutical industry, in part due to biased interpretation of trial results.

Although lower quality studies lend themselves inherently to biased results, this does not appear to account for the relationship between industry funding and the bias towards positive published outcomes. With one major exception, most authors have concluded that industry-funded studies published in peer-reviewed journals are of equivalent or higher quality than non-industry funded clinical trials. This view does not necessarily apply to industry-sponsored symposia journal supplements. Research funded by drug companies is more likely to be published in the proceedings of symposia than non-industry sponsored research. Some, but not all studies have concluded that randomized controlled trials published in these supplements were generally of lower quality, although the relationship between funding and positive outcome persists.

Despite these consistent findings over the last 15 years, results of a recent pilot study of randomized controlled trials published in 5 leading medical journals (including JAMA) failed to document any association between funding source, trial outcome, and study quality. This could reflect the beneficial effects of more stringent editorial policies that have been adopted by the editors of these journals.

Potential remedies:

Investigators and authors. In addition to ethical and legal guidelines that are intended to help protect human subjects, individual investigators should be mindful of guidelines developed to prevent or mitigate potential conflicts of interest that could lead to publication bias. In particular, 2 ethical opinions included in the AMA's Code of Medical Ethics, E-8.031, "Conflicts of Interest: Biomedical Research," and E-8.0315, "Managing Conflicts of Interest in the Conduct of Clinical Trials," (AMA Policy Database), recommend the disclosure of financial compensation from or other material ties to companies whose products they are investigating to various stakeholders, including journal editors. It is also recommended that physicians should not enter into research contracts that permit the sponsoring company to unduly delay or otherwise obstruct the presentation or publication of results.

Uniform institutional practices. A 2001 report issued by the U.S. General Accounting Office noted that equity ownership or investment in a research sponsor may "color [an institution's] review, approval, or monitoring of research ... or its allocation of equipment, facilities, and staff for research." In response to these concerns, the AAMC task force authored a second report that laid out a conceptual framework for assessing institutional conflicts of interest in human subjects research. As a fundamental principle "institutions should ensure that, in practice, the functions and administrative responsibilities related to human subjects research are separate from those related to investment management and technology licensing." The report also contained a series of recommendations regarding: (1) authority/responsibility of institutional officials; (2) circumstances or other financial relationships that should trigger close scrutiny; (3) an appropriate internal committee structure and review process; (4) institutional behavior within the confines of multicenter or single, primary site trials; (5) use of external institutional review boards (IRBs); (6) conditions where recusal may be warranted; (7) policies applying to IRB members; and (8) disclosure requirements.

Journals and journal editors. The publishing community has taken steps to increase the quality of published reports and to reduce publication bias associated with clinical trials and the conflicts of interest that may arise out of industry funding of investigators. The Consolidated Standards for Reporting

of Trials Group (CONSORT) developed a statement intended as an evidence-based approach to improve the quality of reports emanating from randomized controlled trials (RCTs). Originally published in 1996 and revised in 2001, the CONSORT statement comprises a checklist and flow diagram for the reporting of RCTs, primarily those constructed as parallel-group studies. The checklist includes items, based on evidence, that need to be addressed in the report to avoid biased estimates of treatment effect and to properly evaluate the findings. The flow diagram assists readers in analyzing the progress of trial participants from the time they are randomized until the end of their involvement with the trial. Adoption of the CONSORT statement by journals (including JAMA) is associated with improvement in the quality of published reports. Companion efforts to increase the quality of meta-analyses of RCTs (QUOROM), observational studies (MOOSE), and assessments of diagnostic tests (STARD) have been drafted or are under way.

Journal editors also have taken steps to reduce publication bias by revising the "Uniform Requirements for Manuscripts Submitted to Biomedical Journals: Writing and Editing for Biomedical Publication." This document was developed by the International Committee of Medical Journal Editors (ICMJE) and is widely used as the basis for editorial policy. The revision established more rigorous criteria for the acceptance of research sponsored by the pharmaceutical industry, particularly regarding declaration of potential conflicts of interest related to individual authors' commitments or project support. Among the salient points are:

researchers should not enter into agreements that interfere with their access to the data or their ability to analyze the data independently, to prepare manuscripts, and to publish them.

authors should describe the role of the study sponsor(s), if any, in study design; in the collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the report for publication.

These tenets hold that the sponsor must impose no impediment, direct or indirect, on the publication of the study's full results, including data perceived to be detrimental to marketing of the drug. Journals adhering to this approach will not review or publish articles based on studies that are conducted under conditions that allow the sponsor to have sole control of

the data or withhold publication. Accordingly, journals should encourage a culture of transparency in research and reporting by publishing study protocols and publishing all data from drug studies, including negative results, in concert with a comprehensive trials registry (see below) and development of accessible electronic databases.

Voluntary industry guidances. In recognition of the need for proactive involvement of the pharmaceutical industry in addressing publication bias, the Pharmaceutical and Research Manufacturers of America adopted voluntary principles on its members' relationships with those involved in the clinical research process. While these principles "commit to timely communication of meaningful results of controlled trials ... that are approved for marketing regardless of outcome," they do not commit "to mak[ing] the designs of clinical trial protocols available publicly at inception, as in a clinical trials registry."

Staff from a small cadre of pharmaceutical companies have also developed guidelines for good publication practices, which they have offered for voluntary use by sponsors when they seek to publish results from their clinical trials. The Good Publication Practice for Pharmaceutical Companies is intended for use in conjunction with the ICMJE-derived "Uniform Requirements" and the CONSORT statement. The guidelines cover publication standards, including a commitment to publish results from all clinical trials; premature, duplicate, or redundant publication; contractual relationships between sponsors and investigators; study tags or identification; and authorship, including the role of professional medical writers. The guidelines and a list of pharmaceutical companies and CROs that have endorsed their use can be found at www.gpp-guidelines.org.

Other recommendations to reduce publication bias. A number of other steps have been recommended to address the problem of publication bias, to improve the quality and reliability of published drug studies, and to assist physicians in accessing, summarizing, and applying information on new drug treatments.

1. Register all clinical trials at inception. Over the last 30 years there have been a number of events and initiatives related to the goal of trial registration. Industry, government, and certain collaborative registers have been

formed, but no comprehensive system for tracking, organizing, and disseminating information about ongoing clinical trials currently exists. This could be implemented by having the Department of Health and Human Services, the parent agency that encompasses the NIH and the Food and Drug Administration, take responsibility for ensuring trial registration in the United States. A recent example of such collaborative action is the GemCRIS system for registration of gene-transfer studies and facilitated reporting of adverse events.

Institutional review boards could make registration a condition of approval. Industry compliance would also be enhanced if both the researcher and the individual patient insist that the trial be registered before enrollment, with explicit language to that effect in the informed consent document. Finally, legislation to fund, require, and enforce public registration of all trials involving human subjects (regardless of funding source) could be sought. While clinical trial registers address some problems, their effects on patient care will be limited unless they are backed by sound publication policies.

2. Reduce bias in study design by conducting studies of comparative effectiveness and requiring data comparing new drugs with available alternatives for effectiveness and cost as part of the regulatory approval process. Alternatively, to provide physicians with the kind of information needed for optimal treatment decisions, establish a center for the assessment of pharmaceutical effectiveness funded by subscription fees on third-party payers, contributions by payers to address specific research questions, user fees, or taxes on pharmaceutical products.

3. Reduce bias in study conduct by leaving the planning and monitoring of the research design completely to the funded investigators.

Summary and comment:

When an investigator has a financial interest in or funding from a company with activities related to his or her research, the research is more likely to favor the sponsor's product, less likely to be published, and more likely to have delayed publication. Investigators and academic institutions can help ensure the integrity of clinical research by negotiating ethically acceptable contracts that allow full access to data and control of publication rights. Publication bias involving drug studies has been reduced by journal editors

who have adopted the revised CONSORT statement, and via revision of the "Uniform Requirements" by the ICMJE. Adoption of these guidelines helps readers make informed judgments about clinical trials and potential biases involving authors, data analyses, and publication/interpretation of results. Thus, guidelines are already established that "assist physician-researchers in evaluating and preserving the scientific integrity, validity and reliability of pharmaceutical research, regardless of funding source" as requested by Resolution 514 (A-03).

Additionally, some progress has been made in registering clinical trials at inception. Development of a comprehensive trials registry will require a cooperative venture among all participants. Development of a registry will not guarantee publication of results, but would assist authors conducting systematic reviews and meta-analyses in identifying relevant studies for inclusion. Electronic databases that can serve as a repository of published results are needed to accommodate all trial results. Absent legislative action, IRBs and patient advocates form a potentially powerful coalition in requiring clinical trials involving human subjects to be registered as a condition for approval.

Recommendations (adopted AMA policy):

The following statements, recommended by the Council on Scientific Affairs, were adopted by the AMA House of Delegates as AMA policy and directives at the 2004 AMA Annual Meeting:

1. All medical journal editors and authors should adhere to the revised CONSORT Statement and Uniform Requirements for Manuscripts Submitted to Biomedical Journals.

2. The AMA recommends that (a) the Department of Health and Human Services establish a comprehensive registry for all clinical trials conducted in the United States; (b) every clinical trial should have a unique identifier; and (c) all results from registered clinical trials should be made publicly available through either publication or an electronic data-repository.

3. The AMA urges that Institutional Review Boards consider registration of clinical trials to an existing registry as condition of approval.

### 0.0.23 Sponsorship, Authorship, and Accountability: International Committee of Medical Journal Editors (August, 2007)

As editors of general medical journals, we recognize that the publication of clinical-research findings in respected peer-reviewed journals is the ultimate basis for most treatment decisions. Public discourse about this published evidence of efficacy and safety rests on the assumption that clinical-trials data have been gathered and are presented in an objective and dispassionate manner. This discourse is vital to the scientific practice of medicine because it shapes treatment decisions made by physicians and drives public and private health care policy. We are concerned that the current intellectual environment in which some clinical research is conceived, study participants are recruited, and the data analyzed and reported (or not reported) may threaten this precious objectivity.

Clinical trials are powerful tools; like all powerful tools, they must be used with care. They allow investigators to test biologic hypotheses in living patients, and they have the potential to change the standards of care. The secondary economic impact of such changes can be substantial. Well-done trials, published in high-profile journals, may be used to market drugs and medical devices, potentially resulting in substantial financial gain for the sponsor. But powerful tools must be used carefully. Patients participate in clinical trials largely for altruistic reasons – that is, to advance the standard of care. In the light of that truth, the use of clinical trials primarily for marketing, in our view, makes a mockery of clinical investigation and is a misuse of a powerful tool.

Until recently, academic, independent clinical investigators were key players in design, patient recruitment, and data interpretation in clinical trials. The intellectual and working home of these investigators, the academic medical center, has been at the hub of this enterprise, and many institutions have developed complex infrastructures devoted to the design and conduct of clinical trials. The academic enterprise has been a critical part of the process that led to the introduction of many new treatments into medical practice and contributed to the quality, intellectual rigor, and impact of such clinical trials. But, as economic pressures mount, this may be a thing of the past.

Many clinical trials are performed to facilitate regulatory approval of a device or drug rather than to test a specific novel scientific hypothesis. As trials have become more sophisticated and the margin of untreated disease harder to reach, there has been a great increase in the size of the trials and consequently the costs of developing new drugs. It is estimated that the average cost of bringing a new drug to market in the United States is about $500 million. The pharmaceutical industry has recognized the need to control costs and has discovered that private nonacademic research groups – that is, contract research organizations (CROs) – can do the job for less money and with fewer hassles than academic investigators. Over the past few years, CROs have received the lion's share of clinical-trial revenues. For example, in 2000 in the United States, CROs received 60% of the research grants from pharmaceutical companies, as compared with only 40% for academic trialists.

As CROs and academic medical centers compete head to head for the opportunity to enroll patients in clinical trials, corporate sponsors have been able to dictate the terms of participation in the trial, terms that are not always in the best interests of academic investigators, the study participants, or the advancement of science generally. Investigators may have little or no input into trial design, no access to the raw data, and limited participation in data interpretation. These terms are draconian for self-respecting scientists, but many have accepted them because they know that if they do not, the sponsor will find someone else who will. And, unfortunately, even when an investigator has had substantial input into trial design and data interpretation, the results of the finished trial may be buried rather than published if they are unfavorable to the sponsor's product. Such issues are not theoretical. There have been a number of recent public examples of such problems, and we suspect that many more go unreported.

As editors, we strongly oppose contractual agreements that deny investigators the right to examine the data independently or to submit a manuscript for publication without first obtaining the consent of the sponsor. Such arrangements not only erode the fabric of intellectual inquiry that has fostered so much high-quality clinical research but also make medical journals party to potential misrepresentation, since the published manuscript may not reveal the extent to which the authors were powerless to control the conduct

of a study that bears their names. Because of our concern, we have recently revised and strengthened the section on publication ethics in the "Uniform Requirements for Manuscripts Submitted to Biomedical Journals: Writing and Editing for Biomedical Publication," a document developed by the International Committee of Medical Journal Editors (ICMJE) and widely used by individual journals as the basis for editorial policy. The revised section follows this editorial. (The entire "Uniform Requirements" document is undergoing revision; the revised version should be available at the beginning of 2002.) As part of the reporting requirements, we will routinely require authors to disclose details of their own and the sponsor's role in the study. Many of us will ask the responsible author to sign a statement indicating that he or she accepts full responsibility for the conduct of the trial, had access to the data, and controlled the decision to publish.

We believe that a sponsor should have the right to review a manuscript for a defined period (for example, 30 to 60 days) before publication to allow for the filing of additional patent protection, if required. When the sponsor employs some of the authors, these authors' contributions and perspective should be reflected in the final paper, as are those of the other authors, but the sponsor must impose no impediment, direct or indirect, on the publication of the study's full results, including data perceived to be detrimental to the product. Although we most commonly associate this behavior with pharmaceutical sponsors, research sponsored by government or other agencies may also fall victim to this form of censorship, especially if the results of such studies appear to contradict current policy.

Authorship means both accountability and independence. A submitted manuscript is the intellectual property of its authors, not the study sponsor. We will not review or publish articles based on studies that are conducted under conditions that allow the sponsor to have sole control of the data or to withhold publication. We encourage investigators to use the revised ICMJE requirements on publication ethics to guide the negotiation of research contracts. Those contracts should give the researchers a substantial say in trial design, access to the raw data, responsibility for data analysis and interpretation, and the right to publish – the hallmarks of scholarly independence and, ultimately, academic freedom. By enforcing adherence to these revised

requirements, we can as editors assure our readers that the authors of an article have had a meaningful and truly independent role in the study that bears their names. The authors can then stand behind the published results, and so can we.

The section on publication ethics from the "Uniform Requirements for Manuscripts Submitted to Biomedical Journals: Writing and Editing for Biomedical Publication" follows below. The full revised "Uniform Requirements" will be published later.

Conflict of Interest:

Public trust in the peer review process and the credibility of published articles depend in part on how well conflict of interest is handled during writing, peer review, and editorial decision making. Conflict of interest exists when an author (or the author's institution), reviewer, or editor has financial or personal relationships with other persons or organizations that inappropriately influence (bias) his or her actions. The potential of such relationships to create bias varies from negligible to extremely great; the existence of such relationships does not necessarily represent true conflict of interest, therefore. (Relationships that do not bias judgment are sometimes known as dual commitments, competing interests, or competing loyalties.) The potential for conflict of interest can exist whether or not an individual believes that the relationship affects his or her scientific judgment. Financial relationships (such as employment, consultancies, stock ownership, honoraria, paid expert testimony) are the most easily identifiable conflicts of interest and the most likely to undermine the credibility of the journal, the authors, and science itself. Conflicts can occur for other reasons, however, such as personal and family relationships, academic competition, and intellectual passion

All participants in the peer review and publication process must disclose all relationships that could be viewed as presenting a potential conflict of interest. Disclosure of these relationships is particularly important in connection with editorials and review articles, because bias can be more difficult to detect in those publications than in reports of original research. Editors may use information disclosed in conflict of interest and financial interest statements as a basis for editorial decisions. Editors should publish this information if they believe it will be important to readers in judging the manuscript.

Potential Conflicts of Interest Related to Individual Authors' Commitments:

When authors submit a manuscript, whether an article or a letter, they are responsible for disclosing all financial and personal relationships between themselves and others that might bias their work. To prevent ambiguity, authors must state explicitly whether potential conflicts do or do not exist. Authors should do so in the manuscript on a conflict of interest notification page that follows the title page, providing additional detail, if necessary, in the accompanying cover letter. Investigators should disclose potential conflicts to study participants, and should state in the manuscript whether they have done so.

Editors also need to decide when to publish information disclosed by authors about potential conflicts. If doubt exists, it is best to err on the side of publication.

Potential Conflicts of Interest Related to Project Support:

Increasingly, biomedical studies receive funding from commercial firms, private foundations, and government. The conditions of this funding have the potential to bias and otherwise discredit the research.

Scientists have an ethical obligation to submit creditable research results for publication. As the persons directly responsible for their work, researchers therefore should not enter into agreements that interfere with their access to the data or their ability to analyze the data independently, to prepare manuscripts, and to publish them. Authors should describe the role of the study sponsor(s), if any, in study design; in the collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the report for publication. If the supporting source had no such involvement, the authors should so state. Biases potentially introduced when sponsors are directly involved in research are analogous to methodological biases of other sorts; some journals therefore choose to include information about the sponsor's involvement in the methods section of the published paper.

If a study is funded by an agency with a proprietary or financial interest in the outcome, editors may ask authors to sign a statement such as, "I had full access to all of the data in this study and I take complete responsibility for the

integrity of the data and the accuracy of the data analysis." Editors should be encouraged to review copies of the protocol and/or contracts associated with project-specific studies before accepting such studies for publication. Editors may choose not to consider an article if a sponsor has asserted control over the authors' right to publish.

Conflicts of Interest Related to Commitments of Editors, Journal Staff, or Reviewers:

Editors should avoid selecting external peer reviewers with obvious potential conflicts of interest, for example, those who work in the same department or institution as any of the authors. Authors often provide editors with the names of persons they feel should not be asked to review a manuscript because of potential conflicts of interest, usually professional. When possible, authors should be asked to explain or justify their concerns; that information is important to editors in deciding whether to honor such requests.

Reviewers must disclose to editors any conflicts of interest that could bias their opinions of the manuscript, and they should disqualify themselves from reviewing specific manuscripts if they believe such disqualification would be appropriate. As in the case of authors, silence on the part of reviewers concerning potential conflicts may mean either that such conflicts exist that they have failed to disclose, or that conflicts do not exist. Reviewers must therefore also be asked to state explicitly whether conflicts do or do not exist. Reviewers must not use knowledge of the work, before its publication, to further their own interests.

Editors who make final decisions about manuscripts must have no personal, professional, or financial involvement in any of the issues they might judge. Other members of the editorial staff, if they participate in editorial decisions, must provide editors with a current description of their financial interests (as they might relate to editorial judgments) and disqualify themselves from any decisions where they have a conflict of interest. Editorial staff must not use the information gained through working with manuscripts for private gain.

Editors should avoid submitting to their own journal reports of original research to which they have contributed as authors. If they do so, they should recuse themselves from the editorial process, and delegate editorial decisions

on those manuscripts to other members of the editorial staff. Editors should publish regular disclosure statements about potential conflicts of interests related to the commitments of journal staff.

### 0.0.24 Whose Body is it, Anyway?, Atul Gawande (*New Yorker*), October 4, 1999

Excerpted by Meenal and Bashir Mamdani for the *Indian Journal of Medical Ethics*: from Atul Gawande, *Whose Body is it Anyway?* (*New Yorker*, October 4, 1999).

What doctors should do when patients make bad decisions.

Joseph Lazaroff's cancer had spread throughout his body. Eight months earlier, he had seen his doctor about a backache. A scan revealed tumours in Lazaroff's liver, bowel, and up and down his spine. A biopsy showed an untreatable cancer. Lazaroff went on around-the-clock morphine to control his pain. ... his legs had become weak and he became incontinent. A scan showed a metastasis compressing his thoracic spinal cord. Radiation had no effect. Spinal surgery offered a last-ditch chance of restoring some strength to his legs and sphincters. The risks, however, were severe and his chance of surviving the procedure and getting back home was slim. The alternative was to do nothing. He'd go home with hospice care, which would keep him comfortable and help him maintain a measure of control over his life. It was his best chance of dying peacefully surrounded by his loved ones. The decision was Lazaroff's.

Only a decade ago, doctors made the decisions; patients did what they were told. People were put on machines, given drugs, and subjected to operations they would not have chosen. And they missed out on treatments that they might have preferred. Then in 1984 a book, *The Silent World of Doctor and Patient*, by a Yale doctor and ethicist named Jay Katz, dealt a devastating critique of traditional medical decision-making. Katz argued that medical decisions could and should be made by the patients involved. By the early '90s, we were taught to see patients as autonomous decision-makers.

In practice, patients make bad decisions too. But when you see your patients making a grave mistake, should you simply do what the patients' want? The current medical orthodoxy says yes. After all, whose body is it, anyway?

Lazaroff wanted surgery. The oncologist was dubious about the choice, but she called in a neurosurgeon who warned them about the risks. But

Lazaroff wasn't to be dissuaded. Outside the room, David, his son, told me that his mother had spent a long time in intensive care on a ventilator before dying of emphysema, and since then his father had often said that he did not want anything like that to happen to him. But now he was adamant about doing 'everything.' Lazaroff had his surgery the next day. The operation was a technical success. Lazaroff's lungs wouldn't recover however, and we struggled to get him off the ventilator. It became apparent that our efforts were futile. It was exactly the way Lazaroff hadn't wanted to die - strapped down and sedated, tubes in every natural orifice and in several new ones, and on a ventilator.

Lazaroff chose badly because his choice ran against his deepest interests as he conceived them. It was clear that he wanted to live. He would take any risk - even death - to live. But life was not what we had to offer. We could offer only a chance of preserving minimal lower-body function at cost of severe violence to him and extreme odds of a miserable death. But he did not hear us. Couldn't it have been a mistake, then, even to have told him about the surgical option? We are exquisitely attuned to the requirements of patient autonomy. But there are still times when a doctor has to steer patients to do what's right for themselves.

This is a controversial suggestion. People are rightly suspicious of those claiming to know better than they do what's best for them. But a good physician cannot simply stand aside when patients make bad or self-defeating decisions.

Suppose you are a doctor seeing a female patient in her 40s. She had a mammogram before seeing you, and now you review the radiologist's report, which reads, "There is a faint group of punctate clustered calcifications. Biopsy may be considered to exclude the possibility of malignancy." You suggest a biopsy. Three times in the past five years, her annual mammogram has revealed an area of suspicious calcifications. Three times a surgeon has taken her to the operating room and removed the tissue in question. And three times under the pathologist's microscope, it has proved to be benign. "I'm not getting another goddam biopsy," she says, and she stands up to get dressed. Do you let her go? It's not an unreasonable thing to do. She's an adult, after all. Still, these calcifications are not equivocal findings. They

often do indicate cancer. Now people have to be permitted to make their own mistakes. But when the stakes are high, and the bad choice may be irreversible, doctors are reluctant to sit back. You could tell her she's making a big mistake. And in all likelihood you'll lose her. The aim isn't to show her how wrong she is. The aim is to win her over. Notice what good doctors do. They sit her down. And when you sit close by, on the same level as your patients, you're no longer the rushed, bossy doctor with no time for them; patients feel less imposed upon and more inclined to think you are both on the same side of the issue. Oddly enough, nine times out of ten this approach works. People feel as if they've been heard, and have had an opportunity to vent. At this point, they finally begin to ask questions, voice doubts, even work through the logic themselves. And they come around.

But it is misleading to view all this simply as the art of doctorly manipulation: when you see patients cede authority to the doctor, something else may be going on. The new orthodoxy about patient autonomy has a hard time acknowledging an awkward truth: patients frequently don't want the freedom that we have given them. That is, they are glad to have their autonomy respected, but the exercise of that autonomy means being able to relinquish it. It turns out that patients commonly prefer to have others make their medical decisions. One study found that although sixty-four percent of the general public thought they'd want to select their own treatment if they develop cancer, only 12 percent of newly diagnosed cancer patients actually did want to do so. Carl Schneider, a professor of law and medicine at the University of Michigan, recently published a book called *The Practice of Autonomy* in which he sorted through a welter of studies on medical decision-making. He found that ill patients were often in a poor position to make good choices. Schneider found that physicians, being less emotionally engaged, are able to reason through the uncertainties without the distortions of fear and attachment. They work in a scientific culture that disciplines the way they make decisions. They have the benefit of 'group rationality' – norms based on scholarly literature and refined practice and the relevant experience.

Just as there is an art to being a doctor, there is an art to being a patient. You must choose wisely when to submit and when to assert yourself. Even when patients decide not to decide, they should still question their physicians

and insist on explanations. The doctor should not make all these decisions and neither should the patient. Something must be worked out between them. Where many ethicists go wrong is in promoting patient autonomy as a kind of ultimate value in medicine rather than recognizing it as one value among others. Schneider found that what patients want most from doctors isn't autonomy per se; it's competence and kindness. Now, kindness will often be involved in respecting patients' autonomy, assuring that they have control over vital decisions. But it may also mean taking on burdensome decisions when patients don't want them, or guiding patients in the right direction when they do. Many ethicists find this disturbing, and medicine will continue to struggle with how patients and doctors ought to make decisions. But, as the field grows ever more complex and technological, the real task isn't to banish paternalism; the real task is to preserve kindness.

Mr. Howe was in his late 30s, in the hospital following an operation for a badly infected gallbladder. Three days after his surgery, he spiked a high fever and become short of breath. I found him sweating profusely, he had an oxygen mask on, his heart was racing and his blood pressure was much too low. I drew blood for tests and cultures, and went into the hall and paged S., one of the chief residents, for help. S. came right over and went over to him, put a hand on his shoulder, and asked how he was doing. She explained the situation: the sepsis, the likely pneumonia, and the probability that he would get worse before he got better. The antibiotics would fix the problem, but not instantly, she said, and he was tiring out quickly. To get him through it, she would need to place him on a breathing machine. "No," he gasped and sat straight up. "Don't ... put me ... on a ... machine." It would not be for long, she said. Maybe a couple of days. We'd give him sedatives so he'd be as comfortable as possible the whole time. And – she wanted to be sure he understood – without the ventilator he would die. He shook his head. "No ... machine!" He was, we believed, making a bad decision. With antibiotics and some high-tech support, we had every reason to believe he'd recover fully. Could we be certain we were right? No, but if we were right, could we really just let him die? S, looked over at Howe's wife, who was stricken with fear and, in an effort to enlist her in the cause, asked what she thought her husband should do. She burst into tears and left the room.

Soon Howe did tire out and he gradually fell into unconsciousness. That was when S. went into action. She slipped a breathing tube into his trachea. We wheeled Howe to the intensive care unit. Over the next twenty-four hours, his lungs improved markedly. He woke up and opened his eyes, the breathing tube sticking out of his mouth. "I'm going to take this tube out of your mouth now, OK?" I said. Then I pulled it out, and he coughed violently a few times. "You had a pneumonia," I told him, "but you're doing fine now." He swallowed hard, wincing from the soreness. Then he looked at me, and, in a horse but steady voice, he said, "Thank you."

### 0.0.25 Drug Companies & Doctors: A Story of Corruption, Marcia Angell (*New York Review of Books*), January 15, 2009

January 15, 2009

Marcia Angell (*New York Review of Books*)

*Side Effects: A Prosecutor, a Whistleblower, and a Bestselling Antidepressant on Trial*
by Alison Bass
Algonquin Books of Chapel Hill, 260 pp., $24.95

*Our Daily Meds: How the Pharmaceutical Companies Transformed Themselves into Slick Marketing Machines and Hooked the Nation on Prescription Drugs*
by Melody Petersen
Sarah Crichton/Farrar, Straus and Giroux, 432 pp., $26.00

*Shyness: How Normal Behavior Became a Sickness*
by Christopher Lane
Yale University Press, 263 pp., $27.50; $18.00 (paper)

Recently Senator Charles Grassley, ranking Republican on the Senate Finance Committee, has been looking into financial ties between the pharmaceutical industry and the academic physicians who largely determine the market value of prescription drugs. He hasn't had to look very hard.

Take the case of Dr. Joseph L. Biederman, professor of psychiatry at Harvard Medical School and chief of pediatric psychopharmacology at Harvard's Massachusetts General Hospital. Thanks largely to him, children as young as two years old are now being diagnosed with bipolar disorder and treated with a cocktail of powerful drugs, many of which were not approved by the Food and Drug Administration (FDA) for that purpose and none of which were approved for children below ten years of age.

Legally, physicians may use drugs that have already been approved for a particular purpose for any other purpose they choose, but such use should be based on good published scientific evidence. That seems not to be the case here. Biederman's own studies of the drugs he advocates to treat childhood bipolar disorder were, as *New York Times* summarized the opinions

of its expert sources, "so small and loosely designed that they were largely inconclusive [1]."

In June, Senator Grassley revealed that drug companies, including those that make drugs he advocates for childhood bipolar disorder, had paid Biederman $1.6 million in consulting and speaking fees between 2000 and 2007. Two of his colleagues received similar amounts. After the revelation, the president of the Massachusetts General Hospital and the chairman of its physician organization sent a letter to the hospital's physicians expressing not shock over the enormity of the conflicts of interest, but sympathy for the beneficiaries: "We know this is an incredibly painful time for these doctors and their families, and our hearts go out to them."

Or consider Dr. Alan F. Schatzberg, chair of Stanford's psychiatry department and president-elect of the American Psychiatric Association. Senator Grassley found that Schatzberg controlled more than $6 million worth of stock in Corcept Therapeutics, a company he cofounded that is testing mifepristone – the abortion drug otherwise known as RU-486 – as a treatment for psychotic depression. At the same time, Schatzberg was the principal investigator on a National Institute of Mental Health grant that included research on mifepristone for this use and he was coauthor of three papers on the subject. In a statement released in late June, Stanford professed to see nothing amiss in this arrangement, although a month later, the university's counsel announced that it was temporarily replacing Schatzberg as principal investigator "to eliminate any misunderstanding."

Perhaps the most egregious case exposed so far by Senator Grassley is that of Dr. Charles B. Nemeroff, chair of Emory University's department of psychiatry and, along with Schatzberg, coeditor of the influential *Textbook of Psychopharmacology* [2]. Nemeroff was the principal investigator on a five-year $3.95 million National Institute of Mental Health grant – of which $1.35 million went to Emory for overhead – to study several drugs made by GlaxoSmithKline. To comply with university and government regulations, he was required to disclose to Emory income from GlaxoSmithKline, and Emory was required to report amounts over $10,000 per year to the National Institutes of Health, along with assurances that the conflict of interest would be managed or eliminated.

But according to Senator Grassley, who compared Emory's records with those from the company, Nemeroff failed to disclose approximately $500,000 he received from GlaxoSmithKline for giving dozens of talks promoting the company's drugs. In June 2004, a year into the grant, Emory conducted its own investigation of Nemeroff's activities, and found multiple violations of its policies. Nemeroff responded by assuring Emory in a memorandum, "In view of the NIMH/Emory/GSK grant, I shall limit my consulting to GSK to under $10,000/year and I have informed GSK of this policy." Yet that same year, he received $171,031 from the company, while he reported to Emory just $9,999 – a dollar shy of the $10,000 threshold for reporting to the National Institutes of Health.

Emory benefited from Nemeroff's grants and other activities, and that raises the question of whether its lax oversight was influenced by its own conflicts of interest. As reported by Gardiner Harris in *New York Times* [3], Nemeroff himself had pointed out his value to Emory in a 2000 letter to the dean of the medical school, in which he justified his membership on a dozen corporate advisory boards by saying:

Surely you remember that Smith-Kline Beecham Pharmaceuticals donated an endowed chair to the department and there is some reasonable likelihood that Janssen Pharmaceuticals will do so as well. In addition, Wyeth-Ayerst Pharmaceuticals has funded a Research Career Development Award program in the department, and I have asked both AstraZeneca Pharmaceuticals and Bristol-Meyers [sic] Squibb to do the same. Part of the rationale for their funding our faculty in such a manner would be my service on these boards.

Because these psychiatrists were singled out by Senator Grassley, they received a great deal of attention in the press, but similar conflicts of interest pervade medicine. (The senator is now turning his attention to cardiologists.) Indeed, most doctors take money or gifts from drug companies in one way or another. Many are paid consultants, speakers at company-sponsored meetings, ghost-authors of papers written by drug companies or their agents [4], and ostensible "researchers" whose contribution often consists merely of putting their patients on a drug and transmitting some token information to the company. Still more doctors are recipients of free meals and other out-and-out gifts. In addition, drug companies subsidize most meetings of pro-

fessional organizations and most of the continuing medical education needed by doctors to maintain their state licenses.

No one knows the total amount provided by drug companies to physicians, but I estimate from the annual reports of the top nine US drug companies that it comes to tens of billions of dollars a year. By such means, the pharmaceutical industry has gained enormous control over how doctors evaluate and use its own products. Its extensive ties to physicians, particularly senior faculty at prestigious medical schools, affect the results of research, the way medicine is practiced, and even the definition of what constitutes a disease.

Consider the clinical trials by which drugs are tested in human subjects [5]. Before a new drug can enter the market, its manufacturer must sponsor clinical trials to show the Food and Drug Administration that the drug is safe and effective, usually as compared with a placebo or dummy pill. The results of all the trials (there may be many) are submitted to the FDA, and if one or two trials are positive – that is, they show effectiveness without serious risk – the drug is usually approved, even if all the other trials are negative. Drugs are approved only for a specified use – for example, to treat lung cancer – and it is illegal for companies to promote them for any other use.

But physicians may prescribe approved drugs "off label" – i.e., without regard to the specified use – and perhaps as many as half of all prescriptions are written for off-label purposes. After drugs are on the market, companies continue to sponsor clinical trials, sometimes to get FDA approval for additional uses, sometimes to demonstrate an advantage over competitors, and often just as an excuse to get physicians to prescribe such drugs for patients. (Such trials are aptly called "seeding" studies.)

Since drug companies don't have direct access to human subjects, they need to outsource their clinical trials to medical schools, where researchers use patients from teaching hospitals and clinics, or to private research companies (CROs), which organize office-based physicians to enroll their patients. Although CROs are usually faster, sponsors often prefer using medical schools, in part because the research is taken more seriously, but mainly because it gives them access to highly influential faculty physicians – referred to by the industry as "thought-leaders" or "key opinion leaders" (KOLs). These are the people who write textbooks and medical journal papers, issue practice

guidelines (treatment recommendations), sit on FDA and other governmental advisory panels, head professional societies, and speak at the innumerable meetings and dinners that take place every year to teach clinicians about prescription drugs. Having KOLs like Dr. Biederman on the payroll is worth every penny spent.

A few decades ago, medical schools did not have extensive financial dealings with industry, and faculty investigators who carried out industry-sponsored research generally did not have other ties to their sponsors. But schools now have their own manifold deals with industry and are hardly in a moral position to object to their faculty behaving in the same way. A recent survey found that about two thirds of academic medical centers hold equity interest in companies that sponsor research within the same institution [6]. A study of medical school department chairs found that two thirds received departmental income from drug companies and three fifths received personal income [7]. In the 1980s medical schools began to issue guidelines governing faculty conflicts of interest but they are highly variable, generally quite permissive, and loosely enforced.

Because drug companies insist as a condition of providing funding that they be intimately involved in all aspects of the research they sponsor, they can easily introduce bias in order to make their drugs look better and safer than they are. Before the 1980s, they generally gave faculty investigators total responsibility for the conduct of the work, but now company employees or their agents often design the studies, perform the analysis, write the papers, and decide whether and in what form to publish the results. Sometimes the medical faculty who serve as investigators are little more than hired hands, supplying patients and collecting data according to instructions from the company.

In view of this control and the conflicts of interest that permeate the enterprise, it is not surprising that industry-sponsored trials published in medical journals consistently favor sponsors' drugs – largely because negative results are not published, positive results are repeatedly published in slightly different forms, and a positive spin is put on even negative results. A review of seventy-four clinical trials of antidepressants, for example, found that thirty-seven of thirty-eight positive studies were published [8]. But of the thirty-six

negative studies, thirty-three were either not published or published in a form that conveyed a positive outcome. It is not unusual for a published paper to shift the focus from the drug's intended effect to a secondary effect that seems more favorable.

The suppression of unfavorable research is the subject of Alison Bass's engrossing book, *Side Effects: A Prosecutor, a Whistleblower, and a Bestselling Antidepressant on Trial*. This is the story of how the British drug giant GlaxoSmithKline buried evidence that its top-selling antidepressant, Paxil, was ineffective and possibly harmful to children and adolescents. Bass, formerly a reporter for the *Boston Globe*, describes the involvement of three people – a skeptical academic psychiatrist, a morally outraged assistant administrator in Brown University's department of psychiatry (whose chairman received in 1998 over $500,000 in consulting fees from drug companies, including GlaxoSmithKline), and an indefatigable New York assistant attorney general. They took on GlaxoSmithKline and part of the psychiatry establishment and eventually prevailed against the odds.

The book follows the individual struggles of these three people over many years, culminating with GlaxoSmithKline finally agreeing in 2004 to settle charges of consumer fraud for $2.5 million (a tiny fraction of the more than $2.7 billion in yearly Paxil sales about that time). It also promised to release summaries of all clinical trials completed after December 27, 2000. Of much greater significance was the attention called to the deliberate, systematic practice of suppressing unfavorable research results, which would never have been revealed without the legal discovery process. Previously undisclosed, one of GlaxoSmithKline's internal documents said, "It would be commercially unacceptable to include a statement that efficacy had not been demonstrated, as this would undermine the profile of paroxetine [Paxil] [9]."

Many drugs that are assumed to be effective are probably little better than placebos, but there is no way to know because negative results are hidden. One clue was provided six years ago by four researchers who, using the Freedom of Information Act, obtained FDA reviews of every placebo-controlled clinical trial submitted for initial approval of the six most widely used antidepressant drugs approved between 1987 and 1999 – Prozac, Paxil, Zoloft, Celexa, Serzone, and Effexor [10]. They found that on average, placebos were

80 percent as effective as the drugs. The difference between drug and placebo was so small that it was unlikely to be of any clinical significance. The results were much the same for all six drugs: all were equally ineffective. But because favorable results were published and unfavorable results buried (in this case, within the FDA), the public and the medical profession believed these drugs were potent antidepressants.

Clinical trials are also biased through designs for research that are chosen to yield favorable results for sponsors. For example, the sponsor's drug may be compared with another drug administered at a dose so low that the sponsor's drug looks more powerful. Or a drug that is likely to be used by older people will be tested in young people, so that side effects are less likely to emerge. A common form of bias stems from the standard practice of comparing a new drug with a placebo, when the relevant question is how it compares with an existing drug. In short, it is often possible to make clinical trials come out pretty much any way you want, which is why it's so important that investigators be truly disinterested in the outcome of their work.

Conflicts of interest affect more than research. They also directly shape the way medicine is practiced, through their influence on practice guidelines issued by professional and governmental bodies, and through their effects on FDA decisions. A few examples: in a survey of two hundred expert panels that issued practice guidelines, one third of the panel members acknowledged that they had some financial interest in the drugs they considered [11]. In 2004, after the National Cholesterol Education Program called for sharply lowering the desired levels of "bad" cholesterol, it was revealed that eight of nine members of the panel writing the recommendations had financial ties to the makers of cholesterol-lowering drugs [12]. Of the 170 contributors to the most recent edition of the American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders (DSM), ninety-five had financial ties to drug companies, including all of the contributors to the sections on mood disorders and schizophrenia [13]. Perhaps most important, many members of the standing committees of experts that advise the FDA on drug approvals also have financial ties to the pharmaceutical industry [14].

In recent years, drug companies have perfected a new and highly effective method to expand their markets. Instead of promoting drugs to treat

diseases, they have begun to promote diseases to fit their drugs. The strategy is to convince as many people as possible (along with their doctors, of course) that they have medical conditions that require long-term drug treatment. Sometimes called "disease-mongering," this is a focus of two new books: Melody Petersen's *Our Daily Meds: How the Pharmaceutical Companies Transformed Themselves into Slick Marketing Machines and Hooked the Nation on Prescription Drugs* and Christopher Lane's *Shyness: How Normal Behavior Became a Sickness.*

To promote new or exaggerated conditions, companies give them serious-sounding names along with abbreviations. Thus, heartburn is now "gastro-esophageal reflux disease" or GERD; impotence is "erectile dysfunction" or ED; premenstrual tension is "premenstrual dysphoric disorder" or PMMD; and shyness is "social anxiety disorder" (no abbreviation yet). Note that these are ill-defined chronic conditions that affect essentially normal people, so the market is huge and easily expanded. For example, a senior marketing executive advised sales representatives on how to expand the use of Neurontin: "Neurontin for pain, Neurontin for monotherapy, Neurontin for bipolar, Neurontin for everything [15]." It seems that the strategy of the drug marketers – and it has been remarkably successful – is to convince Americans that there are only two kinds of people: those with medical conditions that require drug treatment and those who don't know it yet. While the strategy originated in the industry, it could not be implemented without the complicity of the medical profession.

Melody Petersen, who was a reporter for *New York Times*, has written a broad, convincing indictment of the pharmaceutical industry [16]. She lays out in detail the many ways, both legal and illegal, that drug companies can create "blockbusters" (drugs with yearly sales of over a billion dollars) and the essential role that KOLs play. Her main example is Neurontin, which was initially approved only for a very narrow use – to treat epilepsy when other drugs failed to control seizures. By paying academic experts to put their names on articles extolling Neurontin for other uses – bipolar disease, post-traumatic stress disorder, insomnia, restless legs syndrome, hot flashes, migraines, tension headaches, and more – and by funding conferences at which these uses were promoted, the manufacturer was able to parlay the drug into

a blockbuster, with sales of $2.7 billion in 2003. The following year, in a case covered extensively by Petersen for the *Times*, Pfizer pleaded guilty to illegal marketing and agreed to pay $430 million to resolve the criminal and civil charges against it. A lot of money, but for Pfizer, it was just the cost of doing business, and well worth it because Neurontin continued to be used like an all-purpose tonic, generating billions of dollars in annual sales.

Christopher Lane's book has a narrower focus – the rapid increase in the number of psychiatric diagnoses in the American population and in the use of psychoactive drugs (drugs that affect mental states) to treat them. Since there are no objective tests for mental illness and the boundaries between normal and abnormal are often uncertain, psychiatry is a particularly fertile field for creating new diagnoses or broadening old ones [17]. Diagnostic criteria are pretty much the exclusive province of the current edition of the *Diagnostic and Statistical Manual of Mental Disorders*, which is the product of a panel of psychiatrists, most of whom, as I mentioned earlier, had financial ties to the pharmaceutical industry. Lane, a research professor of literature at Northwestern University, traces the evolution of the DSM from its modest beginnings in 1952 as a small, spiral-bound handbook (DSM-I) to its current 943-page incarnation (the revised version of DSM-IV) as the undisputed "bible" of psychiatry – the standard reference for courts, prisons, schools, insurance companies, emergency rooms, doctors' offices, and medical facilities of all kinds.

Given its importance, you might think that the DSM represents the authoritative distillation of a large body of scientific evidence. But Lane, using unpublished records from the archives of the American Psychiatric Association and interviews with the principals, shows that it is instead the product of a complex of academic politics, personal ambition, ideology, and, perhaps most important, the influence of the pharmaceutical industry. What the DSM lacks is evidence. Lane quotes one contributor to the DSM-III task force:

There was very little systematic research, and much of the research that existed was really a hodgepodge – scattered, inconsistent, and ambiguous. I think the majority of us recognized that the amount of good, solid science upon which we were making our decisions was pretty modest. Lane uses

shyness as his case study of disease-mongering in psychiatry. Shyness as a psychiatric illness made its debut as "social phobia" in DSM-III in 1980, but was said to be rare. By 1994, when DSM-IV was published, it had become "social anxiety disorder," now said to be extremely common. According to Lane, GlaxoSmithKline, hoping to boost sales for its antidepressant, Paxil, decided to promote social anxiety disorder as "a severe medical condition." In 1999, the company received FDA approval to market the drug for social anxiety disorder. It launched an extensive media campaign to do it, including posters in bus shelters across the country showing forlorn individuals and the words "Imagine being allergic to people ... ," and sales soared. Barry Brand, Paxil's product director, was quoted as saying, "Every marketer's dream is to find an unidentified or unknown market and develop it. That's what we were able to do with social anxiety disorder."

Some of the biggest blockbusters are psychoactive drugs. The theory that psychiatric conditions stem from a biochemical imbalance is used as a justification for their widespread use, even though the theory has yet to be proved. Children are particularly vulnerable targets. What parents dare say "No" when a physician says their difficult child is sick and recommends drug treatment? We are now in the midst of an apparent epidemic of bipolar disease in children (which seems to be replacing attention-deficit hyperactivity disorder as the most publicized condition in childhood), with a forty-fold increase in the diagnosis between 1994 and 2003 [18]. These children are often treated with multiple drugs off-label, many of which, whatever their other properties, are sedating, and nearly all of which have potentially serious side effects.

The problems I've discussed are not limited to psychiatry, although they reach their most florid form there. Similar conflicts of interest and biases exist in virtually every field of medicine, particularly those that rely heavily on drugs or devices. It is simply no longer possible to believe much of the clinical research that is published, or to rely on the judgment of trusted physicians or authoritative medical guidelines. I take no pleasure in this conclusion, which I reached slowly and reluctantly over my two decades as an editor of *New England Journal of Medicine*.

One result of the pervasive bias is that physicians learn to practice a very drug-intensive style of medicine. Even when changes in lifestyle would be

more effective, doctors and their patients often believe that for every ailment and discontent there is a drug. Physicians are also led to believe that the newest, most expensive brand-name drugs are superior to older drugs or generics, even though there is seldom any evidence to that effect because sponsors do not usually compare their drugs with older drugs at equivalent doses. In addition, physicians, swayed by prestigious medical school faculty, learn to prescribe drugs for off-label uses without good evidence of effectiveness.

It is easy to fault drug companies for this situation, and they certainly deserve a great deal of blame. Most of the big drug companies have settled charges of fraud, off-label marketing, and other offenses. TAP Pharmaceuticals, for example, in 2001 pleaded guilty and agreed to pay $875 million to settle criminal and civil charges brought under the federal False Claims Act over its fraudulent marketing of Lupron, a drug used for treatment of prostate cancer. In addition to GlaxoSmithKline, Pfizer, and TAP, other companies that have settled charges of fraud include Merck, Eli Lilly, and Abbott. The costs, while enormous in some cases, are still dwarfed by the profits generated by these illegal activities, and are therefore not much of a deterrent. Still, apologists might argue that the pharmaceutical industry is merely trying to do its primary job – further the interests of its investors – and sometimes it goes a little too far.

Physicians, medical schools, and professional organizations have no such excuse, since their only fiduciary responsibility is to patients. The mission of medical schools and teaching hospitals – and what justifies their tax-exempt status – is to educate the next generation of physicians, carry out scientifically important research, and care for the sickest members of society. It is not to enter into lucrative commercial alliances with the pharmaceutical industry. As reprehensible as many industry practices are, I believe the behavior of much of the medical profession is even more culpable [19]. Drug companies are not charities; they expect something in return for the money they spend, and they evidently get it or they wouldn't keep paying.

So many reforms would be necessary to restore integrity to clinical research and medical practice that they cannot be summarized briefly. Many would involve congressional legislation and changes in the FDA, including its drug

approval process. But there is clearly also a need for the medical profession to wean itself from industry money almost entirely. Although industry-academic collaboration can make important scientific contributions, it is usually in carrying out basic research, not clinical trials, and even here, it is arguable whether it necessitates the personal enrichment of investigators. Members of medical school faculties who conduct clinical trials should not accept any payments from drug companies except research support, and that support should have no strings attached, including control by drug companies over the design, interpretation, and publication of research results.

Medical schools and teaching hospitals should rigorously enforce that rule, and should not enter into deals with companies whose products members of their faculty are studying. Finally, there is seldom a legitimate reason for physicians to accept gifts from drug companies, even small ones, and they should pay for their own meetings and continuing education.

After much unfavorable publicity, medical schools and professional organizations are beginning to talk about controlling conflicts of interest, but so far the response has been tepid. They consistently refer to "potential" conflicts of interest, as though that were different from the real thing, and about disclosing and "managing" them, not about prohibiting them. In short, there seems to be a desire to eliminate the smell of corruption, while keeping the money. Breaking the dependence of the medical profession on the pharmaceutical industry will take more than appointing committees and other gestures. It will take a sharp break from an extremely lucrative pattern of behavior. But if the medical profession does not put an end to this corruption voluntarily, it will lose the confidence of the public, and the government (not just Senator Grassley) will step in and impose regulation. No one in medicine wants that.

Notes:

[1] Gardiner Harris and Benedict Carey, "Researchers Fail to Reveal Full Drug Pay," *New York Times*, June 8, 2008.

[2] Most of the information in these paragraphs, including Nemeroff's quote in the summer of 2004, is drawn from a long letter written by Senator Grassley to James W. Wagner, President of Emory University, on October 2, 2008.

[3] See Gardiner Harris, "Leading Psychiatrist Didn't Report Drug Makers' Pay," *New York Times*, October 4, 2008.

[4] Senator Grassley is current investigating Wyeth for paying a medical writing firm to ghost-write articles favorable to its hormone-replacement drug Prempro.

[5 Some of this material is drawn from my article "Industry-Sponsored Clinical Research: A Broken System," *The Journal of the American Medical Association*, September 3, 2008.

[6] Justin E. Bekelman et al., "Scope and Impact of Financial Conflicts of Interest in Biomedical Research: A Systematic Review, *The Journal of the American Medical Association*, January 22, 2003.

[7 ] Eric G. Campbell et al., "Institutional Academic-Industry Relationships," *The Journal of the American Medical Association*, October 17, 2007.

[8] Erick H. Turner et al., "Selective Publication of Antidepressant Trials and Its Influence on Apparent Efficacy," *New England Journal of Medicine*, January 17, 2008.

[9] See Wayne Kondro and Barb Sibbald, "Drug Company Experts Advised Staff to Withhold Data About SSRI Use in Children, *Canadian Medical Association Journal*, March 2, 2004.

[10] Irving Kirsch et al., "The Emperor's New Drugs: An Analysis of Antidepressant Medication Data Submitted to the US Food and Drug Administration," *Prevention & Treatment*, July 15, 2002.

[11] Rosie Taylor and Jim Giles, "Cash Interests Taint Drug Advice," *Nature*, October 20, 2005.

[12] David Tuller, "Seeking a Fuller Picture of Statins," *New York Times*, July 20, 2004.

[13] Lisa Cosgrove et al., "Financial Ties Between DSM-IV Panel Members and the Pharmaceutical Industry," *Psychotherapy and Psychosomatics*, Vol. 75, No. 3 (2006).

[14] On August 4, 2008, the FDA announced that $50,000 is now the "maximum personal financial interest an advisor may have in all companies that may be affected by a particular meeting." Waivers may be granted for amounts less than that.

[15] See Petersen, *Our Daily Meds*, p. 224.

[16] Petersen's book is a part of a second wave of books exposing the deceptive practices of the pharmaceutical industry. The first included Katharine

Greider's *The Big Fix: How the Pharmaceutical Industry Rips Off American Consumers* (PublicAffairs, 2003), Merrill Goozner's The *$800 Million Pill: The Truth Behind the Cost of New Drugs* (University of California Press, 2004), Jerome Avorn's *Powerful Medicines: The Benefits, Risks, and Costs of Prescription Drugs* (Knopf, 2004), John Abramson's *Overdo$ed America: The Broken Promise of American Medicine* (HarperCollins, 2004), and my own *The Truth About the Drug Companies: How They Deceive Us and What to Do About It* (Random House, 2004).

[17] See the review by Frederick Crews of Lane's book and two others, *New York Review*, December 6, 2007.

[18] See Gardiner Harris and Benedict Carey, "Researchers Fail to Reveal Full Drug Pay," *New York Times*, June 8, 2008.

[19] This point is made powerfully in Jerome P. Kassirer's disturbing book, *On the Take: How Medicine's Complicity With Big Business Can Endanger Your Health* (Oxford University Press, 2005).

**A Note to Readers: By The Editors of the *New York Review of Books***

February 12, 2009

In response to Drug Companies & Doctors: A Story of Corruption (January 15, 2009)

New York Review has received a letter from a legal representative of Dr. Alan F. Schatzberg objecting to the use of the word "corruption" in the headline and text of a review by Marcia Angell, M.D., of several books that are critical of the financial ties between the pharmaceutical industry and medical researchers [*NYR*, January 15]. The letter contends that, in its use of the word "corruption," the review may be read as implying that Dr. Schatzberg engaged in "bribery" or "other similar dishonest dealings."

We believe Dr. Schatzberg's legal representative is mistaken. The word "corruption" did not appear in the vicinity of the one brief paragraph that mentioned Dr. Schatzberg in the midst of a long article. Moreover, *New York Review* and Dr. Angell did not state or imply that he engaged in any unlawful conduct. The article specifically said that Stanford University had stated that it saw, as Dr. Angell put it, "nothing amiss" in Dr. Schatzberg's

own record. We feel sure readers would not have drawn from Dr. Angell's article the conclusion that Dr. Schatzberg had engaged in unlawful dealings such as those mentioned by his legal representative; and we would deeply regret it if any readers had such a mistaken impression.

More fundamentally, readers would not, in fact, interpret the word "corruption," as it appeared in the title and final paragraph of the review, in the manner that the letter suggests. In context, the word "corruption" summarized the opinion (set forth in some of the books mentioned in the review) that, as a general matter, the growing financial dependence of the medical profession on the pharmaceutical industry is profoundly detrimental to sound public, medical, and scientific policy.

We invited Dr. Schatzberg to submit a letter to the editor explaining in his own words his disagreement with the review, but he has not chosen to do so. *New York Review* and Dr. Angell regret that a disagreement has arisen with Dr. Schatzberg on this matter.

We have received many letters about Dr. Angell's review and will publish several of them – including one from Stanford University's General Counsel in support of Dr. Schatzberg – in the next issue.

**'Drug Companies & Doctors': An Exchange**

February 26, 2009

By Nada L. Stotland, Debra L. Zumwalt, Reply by Marcia Angell

In response to Drug Companies & Doctors: A Story of Corruption (January 15, 2009)

To the Editors:

In her article ["Drug Companies & Doctors: A Story of Corruption," *NYR*, January 15], Dr. Marcia Angell incorrectly implies that Dr. Alan Schatzberg and Stanford University did not address a conflict of interest issue until US Senator Charles Grassley made inquiries into the matter. In fact, extensive documentation sent to Senator Grassley and available online to the public since June 2008:

(ucomm.stanford.edu/news/conflict_of_interest_schatzerg_grassley.html)

shows that both Stanford University and Dr. Schatzberg did comply with all rules regarding disclosure and management of Dr. Schatzberg's conflict. The National Institute of Mental Health (NIMH) was informed about the potential conflict before any federal funds were expended on Dr. Schatzberg's grant. The requirements established by the National Institutes of Health were met, and appropriate notice was given to the NIMH. Stanford's reviews and management of the potential conflict were extensive and well documented. The integrity and safety of the research were assured by Data Safety Management Boards at both Stanford and the NIMH.

Dr. Angell also took misleading liberties with a letter I wrote to Dr. Jane Steinberg of the NIMH by quoting selective phrases out of context. Her review says that "the university's [general] counsel announced that it was temporarily replacing Schatzberg as principal investigator 'to eliminate any misunderstanding.'" Dr. Angell's selective quote leaves a negative impression not justified by the fuller explanation in my letter. Dr. Schatzberg had already requested that the grant for which he was principal investigator be restructured to eliminate any appearance of conflict. I noted that Stanford University, with the concurrence of Dr. Schatzberg, would temporarily appoint another faculty member as principal investigator while working with the NIMH to restructure the grant, to eliminate any misunderstanding of Dr. Schatzberg's role in mifepristone research for those not familiar with the grant and federal oversight process. This step was taken to put to rest any confusion about the integrity of the research and to allow important mental health research to continue.

It may be that Dr. Angell does not agree with the federal requirements for management of conflicts of interest; however, she should not allege misconduct or corruption where there is none.

Debra L. Zumwalt
Vice President and General Counsel Stanford University
Stanford, California

---

To the Editors:

We appreciate the concerns expressed by Marcia Angell in her review of several books addressing the activities of pharmaceutical companies, their

relationships with physicians, and the significant impact of those relationships on pharmacologic research and the prescribing of medications. The exchange of money or other items of value between the medical community and medically related industries should be transparent and limited to payment for legitimate services. Some of the statements in the review, however, are misleading and prejudicial. For example, because research on children is fraught with ethical challenges, very few medications of any kind have been tested on children and thereby approved for pediatric use by the Food and Drug Administration. Psychiatric medications are not at all unique in that respect.

Many of the most commonly treated medical conditions – hypertension, hypercholesterolemia, migraine – exist on a spectrum of normal – abnormal and are defined as diseases when they cause significant distress or are associated with bad outcomes in the long term. The distinction between a normal cell and a cancerous or potentially cancerous cell is not always clear. Similarly, there are people who are "shy" and people who are unable to eat a meal with friends, make a presentation required in the course of their work, or stand up in church to announce that money is being collected to buy food for the hungry.

As Angell notes, the provision of money and other valuables by pharmaceutical companies to medical schools, medical societies, and individual physicians has been widely accepted for many years. Some of the effects, such as the development of effective new treatments, have been positive. It is good that society and the profession are finally paying attention to the consequences that are negative. But standards – of diagnosis, research, and behavior – change over time. It is unfair to suggest physicians are "corrupt" for activities that were virtually universal when they occurred. The diagnostic manual of psychiatric disorders to which Angell refers was produced fifteen years ago.

The American Psychiatric Association is currently engaged in an inclusive, wide-ranging, transparent, and science-based process of diagnostic classification: the fifth edition of that manual. That process began in 1999, will involve over 160 experts from all over the world, all of whom have agreed to disclose and limit their income from the pharmaceutical industry, and is

slated to end with publication in 2012. The public is invited to read reports on that process and to send comments and suggestions via a dedicated Web site, www.dsm5.org. We are aware of no parallel process in any other branch of medicine.

There has been an explosive increase in the development and marketing of new medications over the last ten or fifteen years, and physicians are and have always been eager to adopt new interventions that hold promise for relieving the suffering of our patients.

Our concern is that the majority of people with painful, disabling, and too often fatal psychiatric conditions are not diagnosed and not treated. According to the World Health Organization, clinical depression is the primary cause of disability worldwide in terms of years lost due to disability. Suicide was the third leading cause of death in the United States in 2004 for children, adolescents, and young adults, ages 10–24. Much of that suffering, disability, and death can be avoided with expert diagnosis and treatment. Psychiatric diseases have been recognized and categorized since the time of the ancient Greeks. Sadly, prejudice against psychiatric diseases, those who suffer from them, and those who treat them, has lasted centuries as well. Much of what stands between those who are ill and the treatment that could help them is stigma. We hope that open-minded examination and appropriate limitation of the sometimes cloudy relationships between physicians and the pharmaceutical industry will clear the air without contributing to that stigma.

Nada L. Stotland, M.D., M.P.H.
President
American Psychiatric Association
Chicago, Illinois

---

Marcia Angell replies:

The above letters, as well as the letter from Dr. Schatzberg's legal representative that was the subject of the editors' "Note to Readers" in the February 12 issue of *New York Review*, mischaracterize my position. My article was about the conflicts of interest that permeate medicine, not failures to disclose them. And nowhere did I state or imply that they were unlawful, as Schatzberg's lawyer charges. My point was that pervasive conflicts of in-

terest corrupt the medical profession, not in a criminal sense, but in the sense of undermining the impartiality that is essential both to medical research and clinical practice. That use of the word "corruption" is supported by every dictionary I have checked (e.g., "a change for the worse of an institution, custom, etc.; a departure from a state of original purity" in the *New Shorter Oxford English Dictionary* of 1993). It is used in that way by many writers, including the author of one of the books I reviewed. A recent example occurs in "The Neurontin Legacy – Marketing Through Misinformation and Manipulation," an article in the January 8 issue of *New England Journal of Medicine*, which states: " ... evidence indicates that drug promotion can corrupt the science, teaching, and practice of medicine."

My short paragraph concerning Dr. Schatzberg focused entirely on the fact that he was principal investigator in a research project funded by a National Institute of Mental Health grant that included testing a drug (mifepristone) to treat psychotic depression that was developed for this use by a company (Corcept Therapeutics) he co-founded and partly owned. He also served as a member of the company's board of directors and chairman of its scientific advisory board. All of this surely constitutes a conflict of interest, and it was longstanding. He initiated a patent application on mifepristone to treat psychotic depression in 1997 (Stanford, where he is on the faculty, owns the patent); he cofounded Corcept Therapeutics in 1998, and, in 1999, extended the National Institute of Mental Health grant for the study of psychotic depression to include testing mifepristone. Stanford also had equity interest in the company. Later the university divested itself of the stock, but still receives royalties for licensing the patent to Corcept Therapeutics. Contrary to the letter from Debra L. Zumwalt, I did not imply that Stanford and Dr. Schatzberg did not "comply with all rules regarding disclosure and management of Dr. Schatzberg's conflict." However, I believe those rules are entirely too lax, and in my article, I was just as critical of the institutions that make the rules as of individuals who follow them.

Stanford believes it adequately managed Dr. Schatzberg's conflict of interest, but I disagree. As an example of its "management" of the conflict of interest, Stanford said in a statement released on June 25, 2008, that although Dr. Schatzberg was principal investigator on the National Institute of Mental

Health grant, he "has not had responsibility for any aspect of the conduct of the grant's research related to mifepristone." Yet he remained principal investigator until July 31, 2008, and was coauthor of published papers on mifepristone in 2001, 2002, and 2006, all favorable to the drug.

Dr. Stotland is incorrect in stating that very few medications of any kind have been tested in children. Many have been tested and found not to warrant FDA approval; others have been tested in poorly designed trials for marketing purposes, not to gain FDA approval. Although it is illegal to promote drugs for use in children if the FDA has not approved them for that use, the law is frequently circumvented by disguising marketing as education or research. Eli Lilly recently agreed to pay $1.4 billion to settle civil and criminal charges of marketing the anti-psychotic drug Zyprexa for uses not approved by the FDA (known as "off-label" uses). Zyprexa, which has serious side effects, is one of the drugs frequently used off-label to treat children diagnosed with bipolar disorder. I don't deny the serious effects of psychiatric conditions, but it is still necessary to show in adequate clinical trials that the drugs used to treat them do more good than harm.

Dr. Stotland says that many medical conditions exist on a spectrum from normal to abnormal, but those she mentions present different problems in diagnosis. Unlike migraines or shyness, hypertension or high cholesterol can be defined by an objective measurement – a blood pressure or cholesterol level. One can dispute the threshold chosen as abnormal, but the measurement is easily verifiable. The fact that psychiatric conditions are not objectively verifiable underscores the necessity for both diagnosis and treatment to be as impartial as possible. That is why conflicts of interest are more serious in this field than in most others, and why it is so important that the authors of the *Diagnostic and Statistical Manual of Mental Disorders*, which is now the standard reference for defining and diagnosing psychiatric conditions, be free of them. Yet, as was the case with the authors of the current edition of the manual (DSM-IV), most members of the task force now working on DSM-V have financial ties to industry. Dr. Stotland seems to think that it is enough to disclose those ties and to limit the amount of money members may receive, but I am far less sanguine.

Most conflicts of interest in medicine can and should be eliminated, not

"managed." One of Dr. Stotland's predecessors as president of the American Psychiatric Association, Dr. Steven S. Sharfstein, warned his colleagues in a column published in *Psychiatric News* on August 19, 2005: "Drug company representatives bearing gifts are frequent visitors to psychiatrists' offices and consulting rooms. We should have the wisdom and distance to call these gifts what they are – kickbacks and bribes." While industry sponsorship of research may be acceptable, it should be for work of scientific, not just commercial, importance. Academic researchers should have no other financial ties to companies whose products they are evaluating and should bear full responsibility for the way the research is conducted and reported.

Judges do not hear cases in which they have a financial interest. Reporters do not write stories about companies in which they have a financial interest. By the same token, doctors should not have a financial interest in treatments they are evaluating or accept gifts from the makers of drugs they prescribe. In a January 5 editorial supporting legislation to require disclosure of industry payments to doctors, *New York Times* ended with this sentence: "Better yet, the medical profession needs to wean itself almost entirely from its pervasive dependence on industry money." I agree, and I suspect a growing part of the public does too.

### 0.0.26 Science and Society: The Interdependence of Science and Law, Stephen Breyer (*Science*), April 24, 1998

April 24, 1998

Stephen Breyer (*Science*)

The practice of science depends on sound law – law that at a minimum supports science by offering the scientist breathing space, within which he or she may search freely for the truth on which all knowledge depends. It is equally true that the law itself increasingly requires access to sound science. This need arises because society is becoming more dependent for its well-being on scientifically complex technology, so, to an increasing degree, this technology underlies legal issues of importance to all of us. We see this conclusion illustrated throughout the legal system.

Consider, for example, the U.S. Supreme Court's docket. Two cases the Court heard last year concerned the right to die. The specific legal question was whether the federal Constitution, which prohibits government from depriving "any person" of "liberty" without "due process of law," requires a state to permit a doctor's assistance in the suicide of a terminally ill patient. Is that "right to assisted suicide" part of the liberty that the Constitution protects? Underlying the legal question was a medical question: To what extent can medical technology reduce or eliminate the risk of dying in severe pain? The medical question did not determine the answer to the legal question, but to do our legal job properly we needed to develop an informed – although necessarily approximate – understanding of the state of that relevant scientific art.

Nor is the right-to-die case unique in this respect. A different case in 1992 challenged the constitutionality of a state sexual psychopath statute. The law required a determination of when a person is both dangerous and mentally ill to the point that the public safety may justify indefinite non-criminal confinement, a question that implicates science and medicine as well as law. One case on our docket this year concerns the sharing of responsibility – by juries, trial judges, and appellate judges – for determining such scientific matters as the potential toxicity or carcinogenicity of chemical substances, such as Bendectin or PCBs. A different criminal case involves the reliability

of polygraph lie detector tests. A third case investigates whether scientific advances in proving paternity may influence statutes that confer citizenship on children born out of wedlock.

The U.S. Supreme Court's docket is only illustrative. Scientific issues permeate the law. Criminal courts consider the scientific validity of, say, DNA sampling, or voice prints, or expert predictions of defendants' "future dangerousness," which can lead courts or juries to authorize or to withhold the punishment of death. Courts review the reasonableness of administrative agency conclusions about the safety of a drug, the risks attending nuclear waste disposal, the leakage potential of a toxic waste dump, or the risks to wildlife associated with the building of a dam. Patent law cases can turn almost entirely on an understanding of the underlying technical or scientific subject matter. And, of course, tort law, which assesses civil liability for injury or death, often requires difficult determinations about the degree of risk of death or injury associated with a chemical ingredient of a pesticide or other product.

The importance of scientific accuracy in the decision of such cases reaches well beyond the case itself. A decision wrongly denying compensation in a toxic substance case, for example, can deprive not only the plaintiff of warranted compensation but can discourage other similarly situated individuals from even trying to obtain compensation and can encourage the continued use of a dangerous substance. On the other hand, a decision wrongly granting compensation, although of immediate benefit to the plaintiff, through the strong financial disincentives that accompany a finding of tort liability, can improperly force abandonment of the substance. Thus if the decision is wrong, it will improperly deprive the public of what can be far more important benefits – those surrounding a drug that cures many while subjecting a few to less serious risk, for example. The upshot is that we must search for law that reflects an understanding of the relevant underlying science, not for law that frees companies to cause serious harm or forces them unnecessarily to abandon the thousands of artificial substances on which modern life depends.

That search is not a search for scientific precision. One could not hope to replicate the subtleties and uncertainties that characterize good scientific

work. A judge is not a scientist, and a courtroom is not a scientific laboratory. Consider the remark made by the physicist Wolfgang Pauli. After a colleague asked whether a certain scientific paper was wrong, Pauli replied, "Oh, no. Certainly not. That paper is not good enough to be wrong." That is our objective. It is to avoid legal decisions that reflect that paper's so-called science. Rather, the law must seek decisions that fall within the boundaries of scientifically sound knowledge and approximately reflect the scientific state of the art.

This objective is sometimes difficult to achieve in practice. The most obvious reason is that most judges lack the scientific training that might facilitate the evaluation of scientific claims or the evaluation of expert witnesses who make such claims. They typically are generalists, dealing with cases that can vary widely in subject matter. Their primary objective is usually process-related: that of seeing that a decision is reached fairly and in a timely way. And the decision in a court of law typically (though not always) focuses on a particular event and specific individualized evidence.

Furthermore, science itself may be highly uncertain and controversial with respect to many of the matters that come before the courts. Scientists often express considerable uncertainty about the dangers of a particular substance. And their views may differ about many related questions that courts may have to answer. What, for example, is the relevance to human cancer of studies showing that a substance causes some cancers, perhaps only a few, in test groups of mice or rats? What is the significance of extrapolations from toxicity studies with high doses of a substance to situations where the doses are much much smaller? Can lawyers or judges or anyone else expect scientists always to be certain or always to have uniform views with respect to an extrapolation from a large to a small dose, when the causes of and mechanisms related to cancer are generally not well known? Many difficult legal cases fall within the heartland of this kind of scientific uncertainty.

Finally, a court proceeding, such as a trial, is not simply a search for dispassionate truth. The law must be fair. In our country, it must always seek to protect basic human liberties. One important procedural safeguard, guaranteed by our Constitution's Seventh Amendment, is the right to a trial by jury. Any effort to bring better science into the courtroom must respect

the jury's constitutionally specified role – even if doing so means that, from a scientific perspective, an incorrect result is sometimes produced.

Despite the difficulties, I believe there is an increasingly important need for law to reflect sound science. I remain optimistic about the likelihood that it will do so. It is common to find cooperation between governmental institutions and the scientific community where the need for that cooperation is apparent. Today, as a matter of course, the president works with a science adviser, Congress solicits advice on the potential dangers of food additives from the National Academy of Sciences, and a scientific regulatory agency will often work with outside scientists, as well as their own, to develop a product that reflects good science.

The judiciary, too, has begun to look for ways to improve the quality of the science on which scientifically related judicial determinations will rest. In the U.S. Supreme Court, as a matter of course, we hear not only from the parties to a case but also from outside groups, which file briefs – 30-page *amicus curiae* briefs – that help us to become more informed about the relevant science. In the "right-to-die" case, we received about 60 such documents from organizations of doctors, psychologists, nurses, hospice workers, and handicapped persons, among others. Many discussed pain control technology, thereby helping us to identify areas of technical consensus and disagreement. Such briefs help to educate the judges on potentially relevant technical matters, helping to make us, not experts, but moderately educated laypersons, and that education improves the quality of our decisions.

Moreover, our Court recently made clear that the law imposes on trial judges the duty, with respect to scientific evidence, to become evidentiary gatekeepers. The judge, without interfering with the jury's role as trier of fact, must determine whether purported scientific evidence is "reliable" and will "assist the trier of fact," thereby keeping from juries testimony that, in Pauli's sense, isn't even good enough to be wrong. Trial judges, looking for ways to perform this function better, increasingly have used pretrial conferences to narrow the scientific issues in dispute, pretrial hearings where potential experts are subject to examination by the court, and the appointment of specially trained law clerks or scientific special masters.

Judge Weinstein of New York suggests that courts sometimes "go beyond

the experts proffered by the parties" and "appoint independent experts" as the federal Rules of Evidence allow. Judge Rosen of Michigan recently appointed a University of Michigan Medical School professor to testify as an expert witness for the court, helping to determine the relevant facts in a case challenging a Michigan law prohibiting partial-birth abortions. Judge Stearns of Massachusetts, acting with the consent of the parties in a recent, highly technical, genetic engineering patent case, appointed a Harvard Medical School professor to serve "as a sounding board for the court to think through the scientific significance of the evidence," to "assist the court in determining the validity of any scientific evidence," and to "assist the court in determining the validity of any scientific evidence, hypothesis or theory on which the experts base their [testimony]."

These techniques are neutral, in principle favoring neither plaintiffs nor defendants. When used, they have typically proved successful. Nonetheless, judges have not often invoked their Rules-provided authority to appoint their own experts. They may hesitate simply because the process is unfamiliar or because the use of this kind of technique inevitably raises questions. Will use of an independent expert, in effect, substitute that expert's judgment for that of the court? Will it inappropriately deprive the parties of control over the presentation of the case? Will it improperly intrude on the proper function of the jury? Where is one to find a truly neutral expert? After all, different experts, in total honesty, often can interpret the same data differently. Will the search for the expert create inordinate delay or significantly increase costs? Who will pay the expert? Judge Acker of Alabama writes: "Unless and until there is a national register of experts on various subjects and a method by which they can be fairly compensated, the federal amateurs wearing black robes will have to overlook their new gatekeeping function lest they assume the intolerable burden of becoming experts themselves in every discipline known to the physical and social sciences, and some as yet unknown but sure to blossom."

The AAAS, working with the American Bar Association and Federal Judicial Center, has begun to explore these matters with an eye toward finding practical ways to provide scientific help: a pilot project to test the feasibility of increased use of court-appointed experts in cases that present technical

issues. The project "will provide a slate of candidates to serve as court-appointed experts in cases in which the court has determined that the traditional means of clarifying issues under the adversarial system are unlikely to yield the information that is necessary for a reasoned and principled resolution of the disputed issues." The project might also examine in some detail instances in which courts have successfully used their own outside experts. How were those experts identified? How might this better be done? How did the court, while protecting the interests of the lawyers and the parties they represent, also protect the experts from unreasonable demands, say on their time? How did the court prepare the expert to encounter what may be an unfamiliar and sometimes hostile legal environment?

The project might also ask whether criteria emerge that help to determine when a court-appointed expert will prove useful and whether that expert might better serve in an adviser-type or witness-like capacity. It would undoubtedly also be helpful to recommend methods for efficiently educating (that is, in a few hours) willing scientists in the ways of the courts, just as it would be helpful to develop training that might better equip judges to understand the ways of science and the ethical, as well as the practical and legal, aspects of the matter. The answers to some of these questions will help determine the practicality of promising methods to help bring science and law closer together.

I believe that in this age of science we must build legal foundations that are sound in science as well as in law. Scientists have offered their help. We in the legal community should accept that offer, and we are in the process of doing so. The result, in my view, will further not only the interests of truth but also those of justice. The law will work better to resolve many of the most important human problems of our time.

The author is an Associate Justice of the Supreme Court of the United States. The text is revised from an address given 16 February 1998 at the 150th Annual Meeting of the American Association for the Advancement of Science.

### 0.0.27 Something Rotten At the Core of Science?, David F. Horrobin (*Trends in Pharmacological Sciences*), February, 2001

February, 2001

David F. Horrobin (*Trends in Pharmacological Sciences*)

The US Supreme Court has recently been wrestling with the issues of the acceptability and reliability of scientific evidence. In its judgement in the case of Daubert versus Merrell Dow, the Court attempted to set guidelines for US judges to follow when listening to scientific experts. Whether or not findings had been published in a peer-reviewed journal provided one important criterion. But in a key caveat, the Court emphasized that peer review might sometimes be flawed and therefore this criterion was not unequivocal evidence of validity or otherwise. A recent analysis of peer review adds to this controversy by identifying an alarming lack of correlation between reviewers' recommendations.

Many scientists and lawyers are unhappy about the admission by the top legal authority in the US, the US Supreme Court, that peer review might in some circumstances be flawed. David Goodstein, writing in a *Guide to the Federal Rules of Evidence*, one of whose functions is to interpret the judgement in the case of Daubert versus Merrell Dow Pharmaceuticals, states that 'Peer review is one of the sacred pillars of the scientific edifice'. In public, at least, almost all scientists would agree. Those who disagree are almost always dismissed in pejorative terms such as 'maverick', 'failure' and 'driven by bitterness'.

Peer review is central to the organization of modern science. The peer-review process for submitted manuscripts is a crucial determinant of what sees the light of day in a particular journal. Fortunately, it is less effective in blocking publication completely; there are so many journals that most even modestly competent studies will be published provided that the authors are determined enough. The publication might not be in a prestigious journal but at least it will get into print. However, peer review is also the process that controls access to funding and here the situation becomes much more serious. There might often be only two or three realistic sources of funding for a project, and the networks of reviewers for these sources are often interacting

and interlocking. Failure to pass the peer review process might well mean that a project is never funded.

Science bases its presumed authority in the world on the reliability and objectivity of the evidence that is produced. If the pronouncements of science are to be greeted with public confidence – and there is plenty of evidence to suggest that such confidence is low and eroding – it should be able to demonstrate that peer review, 'one of the sacred pillars of the scientific edifice', is a process that has been validated objectively as a reliable process for putting a stamp of approval on work that has been done. Peer review should also have been validated as a reliable method for making appropriate choices as to what work should be done. Yet when one looks for that evidence it is simply not there.

For 30 years or so, I and others have been pointing out the fallibility of peer review and have been calling for much more openness and objective evaluation of its procedures. For the most part, the scientific establishment, its journals and its grant-giving bodies have resisted such open evaluation. They fail to understand that if a process that is as central to the scientific endeavour as peer review has no validated experimental base, and if it consistently refuses open scrutiny, it is not surprising that the public is increasingly sceptical about the agenda and the conclusions of science.

Largely because of this antagonism to openness and evaluation there is a great lack of good evidence either way concerning the objectivity and validity of peer review. What evidence there is does not give confidence but is open to many criticisms. Now, Rothwell and Martyn have thrown a bombshell. [Rothwell, P. M., et al. (2000). Reproducibility of peer review in clinical neuroscience – is agreement between reviewers greater than would be expected by chance alone? *Brain*, *123*, 1964 – 1969. ] Their conclusions are measured and cautious but there is little doubt that they have provided solid evidence of something truly rotten at the core of science.

Rothwell and Martyn performed a detailed evaluation of the reviews of papers submitted to two neuroscience journals. Each journal normally sent papers out to two reviewers. Reviews of abstracts and oral presentations sent to two neuroscience meetings were also evaluated. One meeting sent its abstracts to 16 reviewers and the other to 14 reviewers, which provides a

good opportunity for statistical evaluation. Rothwell and Martyn analysed the correlations among reviewers' recommendations by analysis of variance. Their report should be read in full. However, the conclusions are alarmingly clear. For one journal, the relationships among the reviewers' opinions were no better than that obtained by chance. For the other journal, the relationship was only fractionally better. For the meeting abstracts, the content of the abstract accounted for only about 10-20% of the variance in opinion of referees, and other factors accounted for 80-90% of the variance.

These appalling figures will not be surprising to critics of peer review, but they give solid substance to what these critics have been saying. The core system by which the scientific community allots prestige (in terms of oral presentations at major meetings and publication in major journals) and funding is a non-validated charade whose processes generate results little better than does chance. Given the fact that most reviewers are likely to be mainstream and broadly supportive of the existing organization of the scientific enterprise, it would not be surprising if the likelihood of support for truly innovative research was considerably less than that provided by chance.

Scientists frequently become very angry about the public's rejection of the conclusions of the scientific process. However, the Rothwell and Martyn findings, coming on top of so much other evidence, suggest that the public might be right in groping its way to a conclusion that there is something rotten in the state of science. Public support can only erode further if science does not put its house in order and begin a real attempt to develop validated processes for the distribution of publication rights, credit for completed work, and funds for new work. Funding is the most important issue that most urgently requires opening up to rigorous research and objective evaluation.

What relevance does this have for pharmacology and pharmaceuticals? Despite enormous amounts of hype and optimistic puffery, pharmaceutical research is actually failing. The annual number of new chemical entities submitted for approval is steadily falling in spite of the enthusiasm for techniques such as combinatorial chemistry, high-throughput screening and pharmacogenomics. The drive to merge pharmaceutical companies is driven by failure, and not by success.

Could the peer-review processes in both academia and industry have de-

stroyed rather than promoted innovation? In my own field of psychopharmacology, could it be that peer review has ensured that in depression and schizophrenia we are still largely pursuing themes that were initiated in the 1950s? Could peer review explain the fact that in both diseases the efficacy of modern drugs is no better than those compounds developed in 1950? Even in terms of side-effects, where the differences between old and new drugs are much hyped, modern research has failed substantially. Is it really a success that 27 of every 100 patients taking the selective 5-HT reuptake inhibitors stop treatment within six weeks compared with the 30 of every 100 who take a 1950s tricyclic antidepressant compound?

The Rothwell-Martyn bombshell is a wake-up call to the cosy establishments who run science. If science is to have any credibility – and also if it is to be successful – the peer-review process must be put on a much sounder and properly validated basis or scrapped altogether.

### 0.0.28 Is Science Different for Lawyers?, David L. Faigman (*Science*), July 19, 2002

July 19, 2002

David L. Faigman (*Science*)

On 7 January of this year, Judge Lewis Pollak, former dean of the Yale Law School and a highly respected federal district court judge, stunned the legal world when he held that fingerprint experts could not testify that a latent print found at the scene of the crime "matched" the defendant's print. Despite being admitted into courts for nearly 100 years, Judge Pollak found that no one had bothered to conduct any meaningful research on the technique. His ruling was based on the landmark opinion in Daubert v. Merrell Dow Pharmaceuticals, Inc., in which the Supreme Court held that trial court judges must ensure, as gatekeepers, that proffered scientific evidence is valid and reliable. On 13 March, however, Judge Pollak reversed his earlier ruling. Upon reflection, he said, he had come to the realization that fingerprint identification was not a matter of "science," it was a "specialization," and thus need not meet the rigors of the scientific method to be admitted in court. This distinction between science and specialization is premised on a basic skepticism of the scientific method and its usefulness to judicial decision-making. Although this skepticism is not universally held by judges, it threatens the fundamental reform wrought by Daubert, because it is shared by many and is thought intellectually respectable by most. In fact, however, this skepticism stems from ignorance, a condition that can only be remedied by a sustained effort to educate these decision-makers about the practices and culture of hypothesis testing in science. Scientists should lead this effort.

Shortly after Daubert was decided, experts in many disciplines that assist the law looked about and realized that they did not have the research to support the opinions they routinely offered in court. These purveyors of science – "lite" – cleverly sought to avoid the new rule. They claimed that the tough new gate-keeping standards did not apply to them because the decision involved only "scientific" evidence. These experts, and the lawyers who employ them, argued that when witnesses are "technical or specialized experts," Daubert does not apply. All sorts of formerly proud scientists joined

the exodus from science, including psychologists, engineers, and medical doctors. To their everlasting shame, many forensic scientists also disclaimed the science mantle. Obviously, however, any failed science might make such an argument. Allowing every failed science to proclaim itself a "specialty" would have gutted the new standard. In a 1999 decision, Kumho Tire Co. v. Carmichael, the Supreme Court closed this loophole. It held that all expert testimony is subject to the requirement that it be demonstrated to have a valid and reliable basis, whether it is accounting or rocket science.

The challenge that arose after Kumho Tire, and one foreseen in Justice Breyer's opinion for the Court, was the question, by what criteria should the multitude of experts who appear daily in court be evaluated? This is no easy task, because this multitude includes physicists, biologists, toxicologists, epidemiologists, psychologists, engineers, medical doctors, historians, accountants, auto mechanics, and the list goes on almost without end. This extraordinarily broad array of expertise is simply not susceptible to any one scheme of evaluation. In Daubert, the Court had suggested four criteria that might be used to assess the reliability of scientific opinion: (i) Is the opinion testable and has it been tested? (ii) Is the error rate associated with the technique or opinion acceptable? (iii) Has the basis for the opinion survived peer review and has it been published? And (iv) Is it generally accepted among scientists in the pertinent field? Justice Breyer in Kumho Tire, however, did not attempt to offer similar criteria by which to judge the "many different kinds of experts, and many different kinds of expertise" routinely confronted by judges. This monumental task, the Court said, would be left to the discretion of lower court judges.

In his first opinion, Judge Pollak laid out the Daubert framework and found that fingerprinting had not been adequately tested, had no known error rate, had not produced research that was published in peer-reviewed journals and was only generally accepted among a group of like-thinking professionals who are all in the fingerprint identification business. At the same time, it should be noted, Judge Pollak also held that there was no dispute that each person's fingerprint was unique so that identification could be readily accomplished when two full prints could be compared. He took judicial notice of this fact. Moreover, he held that the government expert would be permitted

to introduce exhibits illustrating the similarities between the partial latent print found at the scene and the defendant's print. These two holdings are not terribly controversial in themselves. The holding that ignited the firestorm was Judge Pollak's order that the expert would be foreclosed from offering an opinion on the import of those similarities or that they indicated a "match."

In reversing the first decision, Judge Pollak was obligated to explain how these factors were now met or why they were no longer relevant. Remarkably, he stated, "I concluded in the January 7 opinion that Daubert's testing factor was not met, and I have found no reason to depart from that conclusion." Yet, somehow, he now found that the other three factors mentioned in Daubert, error rate, peer review and publication, and general acceptance, were satisfied. How this was possible, without testing, is a great mystery of the decision. For him, this mystery was solved by his observation that fingerprint identification "is not, in my judgment, itself a science." He likened forensic scientists to "accountants, vocational experts, accident reconstruction experts, [and] appraisers of land or of art." Forensic science was a specialty, not a science.

Judge Pollak's conclusion has been echoed by a great number of federal judges. Judge Crow similarly held that fingerprinting, though as yet untested, is admissible in court. He was persuaded by the technology's success over the past 100 years, in which it "has withstood the scrutiny and testing of the adversarial process." Scientists undoubtedly will find such an assertion laughable. Judge Crow, however, answered their snickering as follows:

"Those of a 'scientific' bent certainly can take issue with whether the judges and lawyers have the education or training to engage in 'scientific' testing and with whether the courtrooms provide the sterile, laboratory-like and objective atmosphere associated with and probably conducive to scientific analysis. Even so, it seems an unreasonable stretch simply to discard this experiential testing as wholly unreliable and to relegate the testifying opinion of all these fingerprint examiners to *ipse dixit*. Moreover, this court joins others who do not read Daubert and Kumho as elevating the scientific method to the touchstone by which all [expert] evidence is to be judged."

In doubting the value of the scientific method as the touchstone by which expert evidence is to be evaluated, judges like Pollak and Crow fail to say

what should replace it. Presumably, it is some combination of "years of personal experience" and general acceptance among members of some well-meaning guild. As a matter of law, I believe this is an incorrect interpretation of Daubert and Kumho Tire. More troubling though, it reflects a basic misunderstanding of the subject of empirical expertise. Contrary to Judge Crow's belief, this overreliance on undifferentiated experience does indeed relegate the opinions of testifying experts to *ipse dixit* – a Latin phrase that roughly translates as, "because I said so."

Judge Crow's statement is remarkable for both its candor and its utter failure to appreciate the culture attending scientific testing of hypotheses. Science does not "exist" categorically or in some concrete encyclopedia of knowledge that passes muster by, say, some committee of the National Academies of Science. Science is a process or method by which factual statements or predictions about the world are devised, tested, evaluated, revised, replaced, rejected, or accepted. There are as many methods of testing as there are hypotheses – indeed, probably more. Courts make a fundamental error when they try to divide the world into science and specialty categories. In truth, every expert who appears in court has "specialized" knowledge of one sort or another. At best, it is specialized knowledge based upon good applied science; at worst, it is specialized knowledge based upon "years of personal experience." The question is, for all specialized knowledge proffered in court, how much and what kind of testing should be necessary before it is admitted? This is a policy question that should depend on two factors. The first concerns the difficulties inherent in studying the phenomenon. The second involves the legal stakes present in cases in which the specialized knowledge might be employed. In the specialty area of fingerprinting, both factors indicate that the courts are getting it wrong.

On the difficulties of studying the phenomenon of fingerprint identification, consider the following hypothesis: There is a vanishingly small statistical likelihood that some set of ridge characteristics (say 8, 10, or 12) on a given fingerprint will be found in a random sample of the population. This hypothesis depends on the factually testable question of what proportion of the population (or relevant subpopulations) has particular ridge characteristics. This is a question of base-rates. To be admissible, fingerprint iden-

tification need not be powerful enough to show identity, but the fact-finder should be given some idea whether one person in 5, or 100, or 1000, could have left the partial print. Of course, other hypotheses could be imagined, including especially proficiency rates among practitioners of the specialty. Proficiency testing would provide data regarding the accuracy of forensic examiners in applying the technology, with possible comparisons to lay-people or even computers. But these hypotheses are hardly daunting and, indeed, a modestly bright graduate student could design research to test many of them. Over time, however, as the science of fingerprint identification progressed, more difficult and sophisticated hypotheses might emerge, requiring greater ability and resources to study. But the most basic work has yet to be done. The other forensic sciences, including bite-mark analysis, handwriting identification, firearms analysis, and so on, are similarly amenable to test. Unfortunately, like fingerprints, most have not been seriously tested.

The second factor concerns what legal risks are involved in the case, or cases, presenting the expertise in question. Fingerprint identification is offered by prosecutors in thousands of trials each year in which defendants are in jeopardy of their liberty and sometimes their lives. In addition, as a practical matter, prosecutors have the institutional wherewithal, especially through the Justice Department, to invest in the research. It is true that testing fingerprinting, not to mention the surfeit of other forensic sciences, is an expensive proposition. Cost should affect courts' expectations regarding what research is done and when it gets done. But the government has so far not claimed that the costs would be prohibitive in this area – a claim likely to receive a skeptical response in any event. Indeed, failure to put the testing burden on the government creates perverse incentives. If courts admit untested speculation, what incentive does the Justice Department have to do the research? The greater the costs in liberty, lives, and property, the greater should be the expectation that good-quality work be done.

In the context of fingerprinting, the amenability of the subject to test and the gravity of the legal stakes involved make it an easy case. Admittedly, other kinds of specialized knowledge will present harder cases. Particularly difficult for lawyers and judges is the first factor, whether an empirical subject is amenable to test. Scientists can be enormously helpful in addressing

this question in concrete cases ranging from alcohol and drug testing to polygraphs. Organizations such as the National Academies of Science and the AAAS have already entered partnerships with legal institutions, such as the Federal Judicial Center and the National Institute of Justice. These and other science organizations should be encouraged to do more. Most judges and lawyers have little creativity when it comes to conceptualizing how certain empirical statements might be examined. Topics such as predictions of violence, the battered woman syndrome, post-traumatic stress disorder, fire and arson investigation, birth defects, and repressed memories, all challenge the legal imagination. If lawyers and judges are going to improve their comprehension of these and other subjects, bridges must be built between the legal and scientific communities.

Daubert initiated a scientific revolution in the law. Although it has taken more than 200 years, the law is ever so slowly coming to embrace the scientific culture of empirical testing. Yet some courts remain in a prescientific age. When Galileo announced that he saw moons around Jupiter through his telescope, the Pope declared that he was mistaken, for the Bible did not allow it. But the moons are there. Similarly, courts can decree that fingerprinting is reliable, but this does not make it true. Only testing will tell us whether it is so.

### 0.0.29 Scientific Evidence and Public Policy, David Michaels (*American Journal of Public Health*), Supplement 1, 2005

Supplement 1, 2005

David Michaels (*American Journal of Public Health*)

In June 1993, the US Supreme Court ordered federal trial judges to become "gatekeepers" of scientific testimony. Under the Daubert v. Merrell Dow Pharmaceuticals, Inc. decision and two related Supreme Court rulings, trial judges are now required to evaluate whether any expert testimony is both "relevant" and "reliable." What began as a well-intentioned attempt to improve the quality of evidentiary science has had troubling consequences. The picture is disturbing: on the basis of a lay judge's ruling, respected scientists have been barred from offering expert testimony in civil cases, and corporate defendants have become increasingly emboldened to cavalierly accuse any adversary of practicing "junk science." Such a challenge requires a wealth of resources. Thus, in a striking contrast to civil actions where defendants are usually corporations, scientific standards are not rigorously applied in criminal trials, sometimes costing innocent and impoverished defendants their freedom, if not their lives. Increasingly, scientific evidence, which is relied upon by federal regulatory agencies charged with protecting public health, is being subjected to Daubert-like challenges. All three developments – in civil actions, criminal trials, and rule-making – favor the powerful in our society over the weak and vulnerable.

Close analysis of the Supreme Court decision reveals a series of concerns. The requirements Daubert imposes on federal judges are unreachable – no absolute criteria exist for assessing the validity of scientific evidence. Scientific reasoning is no more susceptible to a mechanical approach than legal reasoning. Checklists of criteria, although appealing in their convenience, are inadequate tools for assessing causation. Alternatively, judges may rely on their own experience and "common sense," which has inherent biases and limitations.

Compounding this problem, the Daubert decision provides no philosophical tool to help judges identify "good science." The ruling itself is an amalgam of two incompatible philosophies of science, Popper's and Hempel's, neither

of which is capable of supplying the criterion of reliability the Court seeks. It is, therefore, not surprising that judges are no better able to evaluate scientific evidence than groups of jurors, who use a deliberative process to pool their collective wisdom and assess the evidence presented to them.

When scientists evaluate scientific evidence in regulatory agencies, consensus committees, or even on the witness stand, they commonly apply a weight-of-evidence approach, a process or method in which all of the scientific evidence that is relevant to the status of a causal hypothesis is taken into account. A second, related Supreme Court decision, General Electric Co. v. Joiner, encourages judges to evaluate separate elements of scientific evidence individually rather than by assessing the totality of the evidence. This approach, methodologically questionable when applied in civil liability cases, runs directly counter to the precautionary policies built into most health, safety, and environmental statutes.

Daubert has also resulted in judges arbitrarily selecting one scientific discipline (e.g., epidemiology) over another (e.g., toxicology) in positing scientific validity. Uncertainty in science, which is the norm not the exception, does not mean the science is flawed, nor do disagreements among scientists mean that one of the parties is wrong or is using "junk science."

Furthermore, the criteria that make scientific claims valid within scientific settings are not easily transferable to legal settings. As Sheila Jasanoff writes, "the grand question for the law is not how judges can best do justice to science; the more critical concern is how courts can better render justice under conditions of endemic uncertainty and ignorance."

The likelihood that questions of scientific validity are raised in a legal proceeding is related to the wealth of the parties involved. Indigent defendants in criminal trials, for example, are rarely capable of hiring experts to counter questionable science that purports to link them with a crime. In contrast, corporate defendants often hire teams of lawyers and scientific experts to use Daubert to make it difficult and costly for plaintiffs to put on their scientific cases through expert witnesses. The tobacco industry, for example, used its extensive resources to challenge the testimony of numerous expert witnesses; it recognized that driving up the costs to plaintiff attorneys would help insulate the industry from legal accountability for producing a

dangerous product. Overall, it appears that Daubert has likely discouraged plaintiffs without scientifically sound claims from pursuing them in court, whereas others with strong claims but insufficient resources have also been prevented from pursuing just compensation for injury.

Manufactured Uncertainty:

Magnifying or manufacturing scientific uncertainty is another tactic used to delay or prevent public health and environmental protection. The approach was used with great success by the tobacco industry and other manufacturers of dangerous products; now, it is rare for proposed regulations not to be challenged with claims that the scientific evidence is flawed or otherwise imperfect. Manufactured uncertainty has achieved a new level of official respectability in the Data Quality Act, which requires federal agencies to establish procedures to ensure the quality of information disseminated by government. Promoted by tobacco and other opponents of regulation, this largely unknown statutory provision was slipped into a thick federal appropriations bill and passed without debate. It allows parties subject to regulation to challenge every piece of evidence considered by regulators.

Opponents of regulation have deceptively promoted the Data Quality Act and the application of Daubert in regulatory proceedings as a plea for "sound science." In reality, while these "sound science" reforms "sound like science," they have little to do with the way science and scientists work. Instead, they are yet another tactic to delay or halt the imposition of requirements to protect the public's health and environment.

The Coronado Conference Papers:

Concerned about these developments, a group of scientists, several of whom had directed federal public health regulatory programs, came together in 2002 to form the Project on Scientific Knowledge and Public Policy (SKAPP). With support from the Common Benefit Trust, a fund established pursuant to a court order in the Silicone Gel Breast Implant Products Liability Litigation, SKAPP has examined the use and misuse of science in two forums in which public policy is shaped: the courts and the regulatory arena. SKAPP is currently based at The George Washington University School of Public Health and Health Services.

The core papers in this special issue of the *American Journal of Public*

*Health* are the product of a conference on "Scientific Evidence and Public Policy" convened by SKAPP in March 2003. Our objective was to bring new perspectives and academic disciplines to the discussion of science in public policy and to examine why polluters and manufacturers of dangerous products have been so successful in influencing our systems of justice and regulation. In particular, we felt that whereas the Supreme Court's Daubert decision had been the focus of much legal scholarship, its philosophical underpinnings and impacts had not been examined in the scientific community. The 2-day symposium, held in Coronado, CA, provided a forum for scientists, philosophers, cognitive linguists, and science studies scholars to have a dialog with legal scholars and federal judges. Early drafts of several papers were first presented at the Coronado Conference and greatly strengthened by lively debate. Other papers were written subsequent to the conference, informed by, and in response to issues raised there. We hope the Coronado Conference papers in this issue will inform public thinking and policy around the use of science in courts and in protecting the public's health. To facilitate this, the papers are available for download at the SKAPP Web site, www.DefendingScience.org, and have been deposited at PubMed Central, the National Institute of Health's free digital archive of biomedical and life sciences journal literature.

The Coronado Conference papers provide an important assessment of Daubert. Science is more subtle and less rigid than Daubert characterizes it. Whether applied in the courts or by regulatory bodies, Daubert's demand for scientific certainty runs counter to the workings of science, as well as to the basic principle that policy decisions should be made with the best available evidence and must not wait until every piece of evidence is in and until every conceivable doubt is erased.

About the Author:

The author is with the Department of Environmental and Occupational Health at The George Washington University School of Public Health and Health Services, Washington, DC, and is director of the SKAPP Planning Committee. The author wrote this article with the concurrence of the SKAPP Planning Committee.

### 0.0.30 Doubt Is Their Product, David Michaels (*Scientific American*), June, 2005

June, 2005

David Michaels (*Scientific American*)

Industry groups are fighting government regulation by fomenting scientific uncertainty.

Few scientific challenges are more complex than understanding the health risks of a chemical or drug. Investigators cannot feed toxic compounds to people to see what doses cause cancer. Instead laboratory researchers rely on animal tests, and epidemiologists examine the human exposures that have already happened in the field. Both types of studies have many uncertainties, and scientists must extrapolate from the evidence to make causal inferences and recommend protective measures. Because absolute certainty is rarely an option, regulatory programs would not be effective if such proof were required. Government officials have to use the best available evidence to set limits for harmful chemicals and determine the safety of pharmaceuticals.

Uncertainty is an inherent problem of science, but manufactured uncertainty is another matter entirely. Over the past three decades, industry groups have frequently become involved in the investigative process when their interests are threatened. If, for example, studies show that a company is exposing its workers to dangerous levels of a certain chemical, the business typically responds by hiring its own researchers to cast doubt on the studies. Or if a pharmaceutical firm faces questions about the safety of one of its drugs, its executives trumpet company sponsored trials that show no significant health risks while ignoring or hiding other studies that are much less reassuring. The vilification of threatening research as "junk science" and the corresponding sanctification of industry-commissioned research as "sound science" has become nothing less than standard operating procedure in some parts of corporate America.

In 1969 an executive at Brown & Williamson, a cigarette maker now owned by R. J. Reynolds Tobacco Company, unwisely committed to paper the perfect slogan for his industry's disinformation campaign: "Doubt is our product since it is the best means of competing with the 'body of fact' that exists

in the mind of the general public." In recent years, many other industries have eagerly adopted this strategy. Corporations have mounted campaigns to question studies documenting the adverse health effects of exposure to beryllium, lead, mercury, vinyl chloride, chromium, benzene, benzidine, nickel, and a long list of other toxic chemicals and medications. What is more, Congress and the administration of President George W. Bush have encouraged such tactics by making it easier for private groups to challenge government-funded research. Although in some cases, companies may be raising legitimate arguments, the overall result is disturbing: many corporations have successfully avoided expense and inconvenience by blocking and stalling much needed protections for public health.

The Taxicab Standard:

A good example of the current battles between industry and science is the controversy over beryllium. This lightweight metal is vital to the production of nuclear warheads because it increases the yield of the explosions; throughout the cold war, the U.S. nuclear weapons complex was the nation's largest consumer of the substance. Beryllium and its alloys are now used to make electronics equipment and even golf clubs. But the metal is also extremely toxic – breathing in tiny amounts can cause chronic beryllium disease (CBD), a debilitating ailment that scars the lungs. Victims have included not just the machinists who worked directly with the metal but others simply in the vicinity of the milling and grinding processes, often for very short periods. One accountant developed CBD after working for a few weeks each year in an office near where beryllium was being processed. CBD has also been diagnosed in people living near beryllium factories.

As assistant secretary of energy for environment, safety and health from 1998 to 2001, I was the chief safety officer for the nuclear weapons complex, responsible for protecting the health of workers at production and research facilities as well as for safeguarding the surrounding communities and environment. When President Bill Clinton appointed me, the Department of Energy's exposure standard for beryllium had not changed since 1949, some years after the substance's health dangers had become clear. In response to a crisis involving many sick workers and community residents, two scientists working with the Atomic Energy Commission estimated what they thought

to be a safe level – two micrograms of beryllium per cubic meter of air – while they were riding in a taxicab on their way to a meeting. The commission, the predecessor of the DOE, then implemented the so-called taxicab standard.

When the Occupational Safety and Health Administration (OSHA) was established in 1971 to protect the health of workers in the private sector, it, too, adopted the taxicab standard for beryllium. Over the following decades, however, it became clear that workers exposed to beryllium levels well below the standard were falling sick. In the 1990s' the DOE and OSHA began the time-consuming legal process of changing their exposure limits for beryllium. Brush Wellman, the nation's leading producer of the metal, hired Exponent, a Menlo Park, Calif., consulting firm specializing in product defense. Sharing authorship with Brush Wellman's scientists, these consultants wrote a series of papers suggesting it was possible that the size, surface area and number of beryllium particles may be more important than previously thought in the development of CBD. They also raised the hypothesis that skin exposure could play a larger role in disease risk. The consultants concluded that the current standard for beryllium might not be protective but that more research was required before changing it.

After reviewing all the studies and taking testimony from industry and independent scientists, the DOE leadership in the later years of the Clinton administration decided that although more research is always desirable, the department had more than enough information to warrant immediate implementation of a stricter standard for beryllium. We issued a new rule, reducing the acceptable workplace exposure level by a factor of 10. Although we could not prove that the lower limit would eliminate the health risks, we chose a level that we believed would prevent most cases of CBD and that was also technologically feasible. This new standard, however, applies only to DOE workers; workers in the private sector, who fall under OSHA's umbrella, do not enjoy the same protection. In 1998 OSHA declared its intention to follow DOE's lead, but three years later the agency dropped that initiative. In November 2002 the agency implicitly accepted the industry's arguments by issuing a call for additional data on the relation of beryllium disease to, among other things, the size, surface area and number of particles and the extent of skin contact. That is where matters stand today.

As it happens, most scientists believe that beryllium also increases the risk of lung cancer; several studies conducted by epidemiologists at the Centers for Disease Control and Prevention support this conclusion. In 2002, however, statisticians from another product-defense firm, Roth Associates in Rockville, Md., and the University of Illinois published a reanalysis of a 10-year-old CDC study. By changing some key parameters, the authors raised the estimates for the background rate of lung cancer so that the elevation caused by beryllium was no longer statistically significant. (This procedure is rather easily accomplished, whereas the opposite – turning insignificance into significance – is extremely difficult.) Brush Weilman and NGK Metals, a producer of beryllium alloys, had funded the research. The new analysis was published in *Inhalation Toxicology*, a peer-reviewed journal – not one primarily focused on epidemiology but peer-reviewed nonetheless – and the industry now touts its study as evidence that everyone else is wrong.

This pattern is not unique to the beryllium industry. Many other companies that produce hazardous chemicals have hired researchers to dispute and reanalyze data showing adverse health effects. Their conclusions are almost always the same: the evidence is ambiguous, so regulatory action is unwarranted. Out of the almost 3,000 chemicals produced in large quantities (more than one million pounds annually), OSHA enforces exposure limits for fewer than 500. In the past 10 years the agency has issued new standards for a grand total of two chemicals; the vast majority of the others are still "regulated" by voluntary standards set before 1971, when the newly created agency adopted them uncritically and unchanged. New science has had no impact on them. I conclude that successive OSHA administrators have simply recognized that establishing new standards is so time- and labor-intensive, and will inevitably call forth such orchestrated opposition from industry, that it is not worth expending the agency's limited resources on the effort.

Emphasizing uncertainty on behalf of big business has become a big business in itself. The product-defense firms have become experienced and successful consultants in epidemiology, biostatistics and toxicology. In fact, it is now unusual for the science behind any proposed public health or environmental regulation not to be challenged, no matter how powerful the evidence. Currently representatives of indoor tanning salons are hard at work

disparaging the designation of ultraviolet radiation as a cause of skin cancer. Furthermore, the denial of scientific evidence and the insistence on an impossible certainty are not limited to business interests. For instance, some zealous environmentalists remain adamantly opposed to food irradiation – the use of gamma rays, x-rays or electron beams to kill microbes in meats and produce – even though the benefits of the practice greatly outweigh the risks.

PPA and Vioxx:

The power of companies to influence and distort research is also strong in the pharmaceutical industry. Consider the Food and Drug Administration's belated clampdown on phenylpropanolamine (PPA), the over-the-counter drug that was widely used as a decongestant and appetite suppressant for decades. Reports of hemorrhagic strokes in young women who had taken a PPA-containing drug began circulating in the 1970s. Over the next 20 years, the FDA raised questions about PPA's safety, but the trade association representing the drug's manufacturers – including Bayer, Sandoz (now part of Novartis), Wyeth and GlaxoSmithKline – rejected the agency's concerns, employing scientists and lobbyists to keep PPA on the market. Eventually a compromise was reached that allowed the companies to select an investigator and fund an epidemiological study whose design would be approved by both the manufacturers and the FDA. They chose the Yale University School of Medicine; in 1999 the study confirmed that PPA causes hemorrhagic stroke.

Did the manufacturers withdraw the drug, which by then had annual sales of more than $500 million? No. Instead they turned to the Weinberg Group, a product-defense consulting firm based in Washington, D.C., to attack the study and had their attorneys put the researchers through grueling legal depositions. David A. Kessler, former head of the FDA and now dean of the University of California at San Francisco School of Medicine, said, "With the amount of hassle and harassment that [the Yale scientists] had to endure, I'm sure the next time they're asked to undertake something like this, they'll wonder if it's worth the cost." The FDA finally advised manufacturers to stop marketing PPA in November 2000. The agency estimates that the chemical caused between 200 and 500 strokes a year among 18- to 49-year-old people.

Or consider rofecoxib, more commonly known as Vioxx, the once popular

pain reliever made by Merck. Even before the FDA approved Vioxx in May 1999, the agency had reviewed data suggesting that the drug could increase the risk of heart disease. Several independent scientists (that is, ones not on Merck's payroll) also raised red flags, but for the most part the FDA ignored them. Then, in early 2000, the results of a clinical trial showed that participants who took Vioxx for an average of nine months had five times the risk of heart attack as those taking the comparison painkiller, naproxen (sold under the brand name Aleve).

Merck's scientists faced a dilemma. They could interpret this finding to mean either that Vioxx increased heart attack risk by 400 percent or that naproxen reduced the risk by an astounding 80 percent, making it about three times as effective as aspirin in protecting the cardiovascular system. Unsurprisingly, the company's researchers chose the latter interpretation. But Merck abruptly turned about and took Vioxx off the market last September when another trial found that: participants taking the drug for more than 18 months suffered twice as many heart attacks and strokes as those taking a placebo. One FDA analyst estimated that Vioxx caused between 88,000 and 139,000 heart attacks – 30 to 40 percent of which were probably fatal – in the five years the drug was on the market.

Although the *Wall Street Journal* has reported that certain documents suggest that Merck executives were aware of the increased risk of heart attacks, it is hard to imagine that the company's scientists were deliberately promoting a drug they knew was unsafe. At the same time, it is hard to imagine they honestly thought naproxen reduced the risk of heart attack by 80 percent. If they did, they should have urged the government to pour it straight into the water supply. It seems more likely that their allegiances were so tightly linked with the products they worked on, as well as the financial health of their employers, that their judgment became fatally impaired. And the FDA? That agency has neither the legal authority nor the resources to effectively identify the adverse outcomes caused by drugs already on the market.

As a result, civil lawsuits have become the primary means for protecting the public from unsafe drugs and chemicals. Recent rulings of the U.S. Supreme Court, however, have made it harder for plaintiffs to introduce sci-

entific testimony to support their cases. Under the precedents set by Daubert v. Merrell Dow Pharmaceuticals and two related rulings, federal trial judges are now required to determine whether the testimony is reliable and relevant. What began as a well-intentioned effort to improve the quality of scientific evidence has had troubling consequences: according to an analysis published in 2002 in the *Journal of the American Medical Association*, federal judges have barred respected researchers from testifying in drug lawsuits because their evidence – such as medical case reports and toxicological studies on animals – did not meet the strict new standards. Corporate defendants have become increasingly emboldened to challenge any expert testimony on the grounds that it is based on "junk science."

Data Quality:

Industry groups have tried to manipulate science no matter which political party controls the government, but the efforts have grown more brazen since George W. Bush became president. I believe it is fair to say that never in our history have corporate interests been as successful as they are today in shaping science policies to their desires. In 2002, for example, the Bush administration remade a committee that advises the CDC on the issue of childhood lead poisoning. Secretary of Health and Human Services Tommy Thompson replaced prominent researchers with individuals more likely to side with the lead industry. (One new member had testified on behalf of the lead paint industry in a suit brought by the state of Rhode Island to recover the costs of treating children with lead poisoning and cleaning contaminated homes.) Since then, the CDC has not moved to strengthen the federal standards for lead poisoning despite research showing that even very low levels of lead in the blood can sharply reduce a child's IQ.

What is more, this administration has tried to facilitate and institutionalize the corporate strategy of manufacturing uncertainty. Its most significant tool is the Data Quality Act (DQA), a midnight rider attached to a 2001 appropriations bill and approved by Congress without hearings or debate. The DQA authorized the development of guidelines for "ensuring and maximizing the quality, objectivity, utility, and integrity of information." This sounds harmless, even beneficial; who wouldn't want to ensure the quality of government-disseminated information? In practice, however, industry groups

use the DQA to slow or stop attempts at regulation by undercutting scientific reports. The law gives corporations an established procedure for killing or altering government documents with which they do not agree. It has been used by groups bankrolled by the oil industry to discredit the National Assessment on Climate Change, a federal report on global warming; by food industry interests to attack the World Health Organization's dietary guidelines, which recommend lower sugar intake to prevent obesity; and by the Salt Institute to challenge the advice of the National Institutes of Health that Americans should reduce their salt consumption.

Even better for industry would be a way to control information before it becomes part of an official government document. To accomplish this tantalizing goal, in August 2003 the Office of Management and Budget (OMB) rolled out a new proposal entitled "Peer Review and Information Quality." Under the plan, all covered information would undergo some form of peer review before being issued by a government agency, and any information that might affect major regulations or that could have a "substantial impact" on public policies or private-sector decisions would be put through a cumbersome system in which the information was reviewed by experts independent of the agency. Because the proposed peer-review process would exclude all scientists receiving grants or contracts from the agency, it seemed designed to maximize the ability of corporate interests to manufacture and magnify scientific uncertainty.

Enough was enough. In November 2003 the usually quiescent science community finally rose up in protest at a meeting sponsored, at the OMB's request, by the National Academy of Sciences. In the face of this opposition – dozens of organizations fired off scathing letters to the White House – the OMB retreated and implemented a less onerous program that did not exclude the most qualified scientists from the peer-review process.

A new regulatory paradigm is clearly needed, but the Bush administration is heading in the wrong direction. Instead of encouraging industry groups to revise the reports of government scientists, agencies should be focusing more scrutiny on the data and analyses provided by corporate scientists and product-defense firms. And instead of allowing uncertainty to be an excuse for inaction, regulators should return to first principles: use the best science

available but do not demand certainty where it does not exist.

A good example of such an approach is the program to provide compensation for weapons workers sickened after exposure to radiation or chemicals at DOE sites. (I helped to design the initiative, which was enacted by Congress in 2000.) Because it is impossible to definitively determine whether a particular cancer has been caused by radiation exposure, the program estimates probabilities based on the cancer rates among survivors of the nuclear blasts at Hiroshima and Nagasaki. The model is not perfect, but the estimates are as accurate as the available data and methods allow.

In that case, we did the right thing. Now it is time for industry to do the right thing. We need a better balance between health and money.