Applied Probabilistic Reasoning: Part II, Bayes Theorem and Beyond

It is remarkable that a science which began with the consideration of games of chance should have become the most important object of human knowledge. ... The most important questions of life are indeed, for the most part, really only problems of probability – Pierre-Simon Laplace (1812)

September 30, 2016

Part II: Topic Areas

1) Bayes Rule (Theorem) and Diagnostic Testing

2) The Fallacy of the Transposed Conditional (equating $P(A|B)$ and $P(B|A)$)

3) The Probability of Causation (for example, how to prove that some agent (e.g., asbestos) caused a disease (e.g., mesothelioma) in a particular individual

4) The Interpretation of Probability and Risk (particularly in a medical context, and possibly one that is personal)

5) The Odds Ratio (and the confusion with Relative Risk (RR))

6) Probabilistic Reasoning and the Prediction of Human Behavior

Bayes Rule (Theorem)

Bayes Theorem has been known for several hundred years.

The simplest form of Bayes' theorem:

$$P(A|B) = P(B|A)(\frac{P(A)}{P(B)})$$

Thus, if we wish to connect the two conditional probabilities $P(A|B)$ and $P(B|A)$, the latter must be multiplied by the ratio of the marginal (or prior) probabilities, $\frac{P(A)}{P(B)}$.

Noting that $P(B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A})$, the simplest form of Bayes' theorem can be rewritten in a less simple but more common form of

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\bar{A})P(\bar{A})}$$

Diagnostic Testing Terminology

Suppose we have a test that assesses some relatively rare occurrence (for example, disease, ability, talent, terrorism propensity, drug or steroid usage, antibody presence, being a liar [where the test is a polygraph]).

Let B be the event that the test says the person has "it," whatever that may be;
A is the event that the person really does have "it."

Two "reliabilities" are needed to characterize test performance:

(a) the probability, $P(B|A)$, that the test is positive if the person has "it"; this is referred to as the *sensitivity* of the test;

(b) the probability, $P(\bar{B}|\bar{A})$, that the test is negative if the person doesn't have "it"; this is the *specificity* of the test.

Positive and Negative Predictive Values

The conditional probability used in the denominator of Bayes' rule, $P(B|\bar{A})$, is merely $1 - P(\bar{B}|\bar{A})$, and is the probability of a "false positive."

The quantity of prime interest, the *positive predictive value* (PPV), is the probability that a person has "it" given that the test says so, $P(A|B)$, and is obtainable from Bayes' rule using the specificity, sensitivity, and prior probability, $P(A)$:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + (1 - P(\bar{B}|\bar{A}))(1 - P(A))} \; .$$

Similarly, the *negative predictive value* (NPV) is the probability that a person doesn't have "it" given that the test says he doesn't $(P(\bar{A}|\bar{B}))$ and can be obtained in the same way.

The Downside of Diagnostic Tests

To understand how well the test does, the facilitative effect of B on A needs interpretation; that is, a comparison of $P(A|B)$ to $P(A)$, plus an absolute assessment of the size of $P(A|B)$ by itself.

Here, the situation is usually dismal whenever $P(A)$ is small (such as when screening for a relatively rare occurrence), and the sensitivity and specificity are not perfect.

Although $P(A|B)$ will generally be greater than $P(A)$, and thus B facilitative of A, the absolute size of $P(A|B)$ is commonly so small that the value of the screening may be questionable.

How to Evaluate Diagnostic Tests

There is some debate as to how a diagnostic test should be evaluated; for example, are test sensitivity and specificity paramount or should our emphasis instead be on the positive and negative predictive values?

Sensitivity and specificity, being properties of the test itself and obtained on persons known to have or not to have the condition in question, would be of primary interest when deciding whether to use the test.

But once the diagnostic test results are available, and irrespective of whether they are positive or negative, sensitivity and specificity are no longer relevant.

For clinical or other applied uses, the main issue is to determine whether the subject in question has the condition given the observed test results, and this is measured by the positive and negative predictive values.

Mammogram Screening

Our first numerical example considers the efficacy of mammograms in detecting breast cancer.

In the United States, about 180,000 women are found to have breast cancer each year from among the 33.5 million women who annually have a mammogram.

Thus, the (prior) probability of a tumor is about $180{,}000/33{,}500{,}000 = .0054$.

Mammograms are no more than 90% accurate, implying that

$P(\text{positive mammogram} \mid \text{tumor}) = .90$ (test sensitivity)
$P(\text{negative mammogram} \mid \text{no tumor}) = .90$ (test specificity)

Because we do not know whether a tumor is present, all we know is whether the test is positive, Bayes' theorem must be used to calculate the probability we really care about, the positive predictive value (PPV), which is .047:

$$P(\text{tumor} \mid \text{positive mammogram}) =$$

$$\frac{.90(.0054)}{.90(.0054) + .10(.9946)} = .047,$$

This is obviously greater than the prior probability of .005 (so the event of a positive mammogram is facilitative of the event of breast cancer) but still very small in magnitude;

again, as in the Fecal Occult Blood Test example, more than 95% of the positive tests that arise turn out to be incorrect.

Natural Frequencies

Gigerenzer and colleagues have argued for the importance of understanding the PPV of a test, but suggest the use of "natural frequencies" and a simple 2 × 2 table of the type presented earlier, rather than actual probabilities substituted into Bayes' rule.

Based on an assumed population of 10,000, the prior probability of A, plus the sensitivity and specificity values, we have the following 2 × 2 table:

	tumor	no tumor	Row Sums
$+$ mammogram	49	995	1044
$-$ mammogram	5	8951	8956
Column Sums	54	9946	10,000

The PPV is then simply $49/1044 = .047$, using the frequency value of 49 for the cell ($+$ mammogram, tumor) and the $+$ mammogram row sum of 1044.

Mood Disorders Questionnaire (MDQ) Evaluation

The second example is from clinical psychology and uses data from the Mark Zimmerman et al. article entitled "Performance of the Mood Disorders Questionnaire in a Psychiatric Outpatient Setting." (*Bipolar Disorders*, 2009, *11*, 759–765).

The data reported in the article can be given in the form of the following 2×2 contingency table.

The row attribute is a classification by the Mood Disorders Questionnaire (MDQ); the column attribute is a clinical classification according to a Structured Clinical Interview for DSM Disorders (SCID), which is the supposed "gold standard" for bipolar disorder diagnosis.

	SCID:BP	SCID:NBP	Row Sums
MDQ:Bipolar (BP)	33	65	98
MDQ:Not Bipolar (NBP)	19	363	382
Column Sums	52	428	480

Various MDQ test characteristics can be computed from the frequencies given in the table:

sensitivity $= .635$ ($= 33/52$);
specificity $= .848$ ($= 363/428$);
positive predictive value $= .337$ ($= 33/98$);
negative predictive value $= .950$ ($= 363/382$).

Several comments are in order about these rather dismal values.

First, the diagnostic accuracy of the MDQ (the proportion of correct diagnoses using the SCID to indicate the "true" diagnosis) is 82.5% $(= (33 + 363)/480)$, but this value is less than simple prediction by the base rates which would consistently predict someone to be "not bipolar" (these predictions would be correct 89.2% of the time $(= 428/480)$).

Second, the event (B) of receiving an MDQ diagnosis of "bipolar" is facilitative of an SCID diagnosis of "bipolar" (event A); that is,

$P(A|B)\ (= 33/98 = .337) > P(A)\ (= 52/480 = .108)$.

But because the PPV of .337 is below 1/2, **a person testing "bipolar" with the MDQ is more likely than not to be assessed as "not bipolar" with the supposedly more accurate SCID**.

In fact, 2 out of 3 diagnoses of "bipolar" with the MDQ are incorrect. This is a clear indictment of the MDQ as a reasonable screening device for the diagnosis of being bipolar.

These kinds of anomalous situations where prediction by base rates outperforms prediction by a diagnostic test and where positive predictive values are less than one-half, are discussed in greater detail in Module 4 on Probabilistic Reasoning and Diagnostic Testing.

The Fallacy of the Transposed Conditional

The simplest form of Bayes' theorem relates $P(A|B)$ and $P(B|A)$ by multiplying this later conditional probability by the ratio of the prior probabilities:

$$P(A|B) = P(B|A)(\frac{P(A)}{P(B)})$$

Given this form of Bayes' theorem, it is clear that for $P(A|B)$ and $P(B|A)$ to be equal, the two prior probabilities, $P(A)$ and $P(B)$, must first be equal.

When the prior probabilities, $P(A)$ and $P(B)$, are not equal, to assert equality for $P(A|B)$ and $P(B|A)$, is to commit the "fallacy of the transposed conditional," the "inverse fallacy," or in a legal context, the "prosecutor's fallacy."

Four Examples of the Inverse Fallacy

(1) in the (mis-)interpretation of what a p-value signifies in statistics;

(2) returning to the Sally Clark case from an earlier part of this module, her ultimate conviction is partly attributable to the operation of the "prosecutor's fallacy";

(3) in deciding when to be screened for colon cancer by a colonoscopy rather than by the simpler, less invasive, and less expensive sigmoidoscopy;

(4) the confusion between test sensitivity (specificity) and the positive (negative) predictive value.

Misinterpreting *p*-values

In beginning statistics, a "*p*-value" is defined as follows: assuming that some given null hypothesis, H_o, is true, the *p*-value is the probability of seeing a result (in your data) as or more extreme than what was actually observed.

It is not the probability that the null hypothesis is true given what was actually observed.

Explicitly, the probability of seeing a particular data result conditional on the null hypothesis being true, $P(\mathrm{data} \mid H_o)$, is confused in the transposition fallacy with $P(H_o \mid \mathrm{data})$, the probability that the null hypothesis is true given that a particular data result has occurred.

Sally Clark Again

We return to the Sally Clark conviction where the invalidly constructed probability of 1 in 73 million was used to successfully argue for Sally Clark's guilt.

Let A be the event of innocence and B the event of two "cot deaths" within the same family.

The invalid probability of 1 in 73 million was considered to be for $P(B|A)$;

a simple equating with $P(A|B)$, the probability of innocence given the two cot deaths, led directly to Sally Clark's conviction.

Continuing the Royal Statistical Society Press Release

Aside from its invalidity, figures such as the 1 in 73 million are very easily misinterpreted. Some press reports at the time stated that this was the chance that the deaths of Sally Clark's two children were accidental. This (mis-)interpretation is a serious error of logic known as the Prosecutor's Fallacy.

The Court of Appeal has recognised these dangers (*R v. Deen* 1993, *R v. Doheny/Adams* 1996) in connection with probabilities used for DNA profile evidence, and has put in place clear guidelines for the presentation of such evidence.

DNA Evidence and the Prosecutor's Fallacy

The exact same circumstances can occur in the (mis)use of DNA evidence.

Here, the event B is the existence of a "match" between a suspect's DNA and what was found, say, at the crime scene; the event A is again one of innocence.

The value for $P(B|A)$ is the probability of a DNA match given that the person is innocent.

Commission of the "prosecutor's fallacy" would reverse the conditioning and say that this latter (presumably small) probability is actually for $P(A|B)$, the probability of innocence given that a match occurs.

Colonoscopy Versus Sigmoidoscopy

Edward Beltrami's book, *Mathematical Models for Society and Biology* (Academic Press; 2013), includes a chapter called: "A Bayesian Take on Colorectal Screening ..."

We begin with several selective quotations from an article in the *New York Times* by Denise Grady (July 20, 2000), "More Extensive Test Needed For Colon Cancer, Studies Say":

The test most commonly recommended [a sigmoidoscopy] to screen healthy adults for colorectal cancer misses too many precancerous growths and should be replaced by a more extensive procedure [a colonoscopy] that examines the entire colon, doctors are reporting today.

The more common test, sigmoidoscopy, reaches only about two feet into the colon and is generally used to screen people 50 and older with an average risk of colon cancer. The more thorough procedure, colonoscopy, probes the full length of the colon, 4 to 5 feet ...

Beltrami's Analysis

We have the following conditional probabilities involving the two events U: there are advanced upper colon lesions, and L: there are no lower colon polyps: $P(U|L) \approx .02$ and $P(L|U) \approx .50$.

A doctor wishing to convince a patient to do the full colonoscopy might well quote the second statistic, $P(L|U)$, and say "50% of all upper colon cancerous polyps would be missed if only the sigmoidoscopy were done."

Although this statement is true, it might not be as convincing to undergo the much more invasive colonoscopy compared to a sigmoidoscopy if the first statistic, $P(U|L)$, were then quoted: "there is a very small probability of 2% of the upper colon showing cancerous lesions if the sigmoidoscopy shows no lower colon polyps."

Confusing the 2% in this last statement with the larger 50% amounts to the commission of the transposition fallacy.

Confusing Test Sensitivity and the Positive Predictive Value

Consider the generic diagnostic testing context where B is the event of testing "positive" and A is the event that the person really is "positive."

Equating sensitivity and the positive predictive value requires $P(A|B)$ to be equal to $P(B|A)$;

or in words, the probability of having "it" given that the test is positive must be the same as the test being positive if the person really does have it.

Consider our example on breast cancer screening: if the base rate for having cancer is small (as here: $P(A) = .0054$), and differs from the probability of a positive test (as here: $P(B) = .90(.0054) + .10(.9946) = .1044$), the positive predictive value can be very low ($P(A|B) = .047$), which is nowhere near the assumed test sensitivity ($P(B|A) = .90$).

The Probability of Causation

In mass (toxic) tort cases (i.e., for civil “wrongs,” such as for asbestos, breast implants, and Agent Orange) there is a need to establish, in a legally acceptable fashion, some notion of causation.

First, there is a concept of *general causation* concerned with whether an agent can increase the incidence of disease in a group;

because of individual variation, a toxic agent will not generally cause disease in every exposed individual.

Specific causation deals with an individual’s disease being attributable to exposure from an agent.

Cohort Study

The establishment of general causation (and a necessary requirement for establishing specific causation) typically relies on a *cohort study*.

This is a method of epidemiologic study where groups of individuals are identified who have been or in the future may be differentially exposed to agent(s) hypothesized to influence the probability of occurrence of a disease or other outcome.

The groups are observed to assess whether the exposed group is more likely to develop disease.

One common way to organize data from a cohort study is through a simple 2 × 2 contingency table, similar in form to those seen earlier:

	Disease	No Disease	Row Sums
Exposed	N_{11}	N_{12}	N_{1+}
Not Exposed	N_{21}	N_{22}	N_{2+}

Relative Risk

Here, N_{11}, N_{12}, N_{21}, and N_{22} are the cell frequencies; N_{1+} and N_{2+} are the row frequencies. If we let p_E and p_{NE} denote the two underlying probabilities of getting the disease for particular cases within the conditions, respectively, the ratio $\frac{p_E}{p_{NE}}$ is referred to as the relative risk (RR), and may be estimated with the data as follows:

estimated relative risk $= \frac{N_{11}/N_{1+}}{N_{21}/N_{2+}}$.

The common legal standard used to argue for both specific and general causation is an RR of 2.0 (or greater).

Attributable Risk

A measure commonly referred to in tort litigations is attributable risk (AR), defined as

$\text{AR} = \frac{p_E \; - \; p_{NE}}{p_E}$, and estimated by $1 - \frac{1}{\mathrm{RR}}$.

Attributable risk, also known as the "attributable proportion of risk" represents the amount of disease among exposed individuals assignable to the exposure.

The common legal standard used to argue for both specific and general causation is an RR of 2.0, or an AR of 50%. At this level, it is "as likely as not" that exposure "caused" the disease (or "as likely to be true as not,")

Genetics Again

Besides toxic tort cases, genetics is an area where the idea of attributable risk is continually discussed in informed media outlets such as the *New York Times*.

The “penetrance” of a particular genetic anomaly or mutation was briefly explained earlier in the context of Angelina Jolie’s decision to undergo a preventive mastectomy.

But there now seems to be a stream of genetic studies reported on regularly where an informed understanding of attributable and relative risk would be of benefit for our own personal medical decision making.

To give one such example, we have the recent article in the *New York Times* by Nicholas Bakalar (August 6, 2014), entitled “Study Shows Third Gene as Indicator for Breast Cancer.”; several quotes follow:

Quotes From the Times

Mutations in a gene called PALB2 raise the risk of breast cancer in women by almost as much as mutations in BRCA1 and BRCA2, the infamous genes implicated in most inherited cases of the disease, a team of researchers reported Wednesday. Over all, the researchers found, a PALB2 mutation carrier had a 35 percent chance of developing cancer by age 70. By comparison, women with BRCA1 mutations have a 50 percent to 70 percent chance of developing breast cancer by that age, and those with BRCA2 have a 40 percent to 60 percent chance. **The lifetime risk for breast cancer in the general population is about 12 percent**.

The breast cancer risk for women younger than 40 with PALB2 mutation was eight to nine times as high as that of the general population. The risk was six to eight times as high among women 40 to 60 with these mutations, and five times as high among women older than 60.

The Energy Employees Occupational Illness Compensation Program (EEOICP)

During the Clinton administration, an initiative was enacted into law to compensate workers in the nuclear weapons industry who developed cancer or lung disease as a consequence of exposure to radiation, beryllium, and other toxic hazards.

The EEOICP was signed into law on December 7, 2000 by President Clinton, along with Executive Order 13179 reproduced in your readings in Module One.

The EEOICCP is one of the most successful and well-administered Federal compensation programs. It has its own non-profit advocacy group called "Cold War Patriots" (submotto: We did our part to keep America Free!)

This advocacy group provides informational meetings and help for those who might be eligible under the program.

New Mexican Ad

Below is part of an ad that appeared in the *New Mexican* (Santa Fe, New Mexico; June, 2014) announcing informational meetings in Penasco, Los Alamos, and Espanola:

Attention Former LANL (Los Alamos National Lab), Sandia Labs, and Uranium Workers:

— Join us for an important town hall meeting

— Learn if you qualify for benefits up to $400,000 through the Energy Employees Occupational Illness Compensation Program Act (EEOICPA)

— Learn about no-cost medical benefit options

— Learn how to apply for consequential medical conditions and for impairment re-evaluation for approved conditions

Federal Register

The EEOICP represents an implementation of the "as likely as not standard" for attributing possible causation (and compensation).

An extensive excerpt is given in your Module One reading from the *Federal Register* concerning the Department of Health and Human Services and its *Guidelines for Determining the Probability of Causation and Methods for Radiation Dose Reconstruction Under the [Energy] Employees Occupational Illness Compensation Program Act of 2000*.

This material should give a good sense of how the modeling principles of probability and statistics are leading to ethically defensible compensation models;

here, the models used are for all those exposed to ionizing radiation through an involvement with the United States' nuclear weapons industry.

The Interpretation of Probability and Risk

The Association for Psychological Science publishes a series of timely monographs on *Psychological Science in the Public Interest*. One recent issue was from Gerd Gigerenzer and colleagues, entitled "Helping Doctors and Patients Make Sense of Health Statistics" (Gigerenzer et al., 2007).

It discusses aspects of statistical literacy as it concerns health, both our own individually as well as societal health policy more generally.

If an overall admonition is needed, it is that context is always important, and the way data and information are presented is absolutely crucial to an ability to reason appropriately and act accordingly.

We review several of the major issues raised by Gigerenzer et al. in the discussion to follow.

Rudy Guiliani Quotation

We begin with a quotation from Rudy Guiliani from a New Hampshire radio advertisement that aired on October 29, 2007, during his run for the Republican presidential nomination:

I had prostate cancer, five, six years ago. My chances of surviving prostate cancer and thank God I was cured of it—in the United States, 82 percent. My chances of surviving prostate cancer in England, only 44 percent under socialized medicine.

Not only did Guiliani not receive the Republican presidential nomination, he was just plain wrong on survival chances for prostate cancer.

The problem is a confusion between survival and mortality rates. Basically, higher survival rates with cancer screening do not imply longer life.

Survival Rate/Mortality Rate

Define a five-year survival rate and an annual mortality rate:

five-year survival rate = (number of diagnosed patients alive after five years)/(number of diagnosed patients);

annual mortality rate = (number of people who die from a disease over one year)/(number in the group).

The inflation of a five-year survival rate is caused by a *lead-time bias*, where the time of diagnosis is advanced (through screening) even if the time of death is not changed.

Overdiagnosis Bias

Moreover, such screening, particularly for cancers such as prostate (or for breast “cancer” – *ductal carcinoma in situ*), leads to an *overdiagnosis bias*, the detection of a pseudodisease that will never progress to cause symptoms in a patient’s lifetime.

Besides inflating five-year survival statistics over mortality rates, overdiagnosis leads more sinisterly to overtreatment that does more harm than good (for example, incontinence, impotence, and other health-related problems).

See, for example, the book by H. Gilbert Welch, *Overdiagnosed: Making People Sick in the Pursuit of Health* (2012)

Relative Risk Reduction/Absolute Risk Reduction

A major area of concern in the clarity of reporting health statistics is in how the data are framed as relative risk reduction or as absolute risk reduction, with the former usually seeming much more important than the latter.

We give examples that present the same information:

Relative risk reduction: If you have this test every two years, your chance of dying from the disease will be reduced by about one third over the next ten years.

Absolute risk reduction: If you have this test every two years, your chance of dying from the disease will be reduced from 3 in 1000 to 2 in 1000, over the next ten years

The Number Needed to Treat (NNT)

A useful variant on absolute risk reduction is given by its reciprocal, the *number needed to treat* (NNT);

if 1000 people have this test every two years, one person will be saved from dying from the disease every ten years.

(Numerically, the NNT is just the reciprocal of the absolute risk reduction, or in this case, $1/(.003 - .002) = 1/.001 = 1000$.)

Some criminal justice variants on the NNT idea discuss about the Number Needed to Incarcerate to prevent one instance of a violent act in the future – or in states that still allow the death penalty, the Number Needed to Execute

Times Article by Tara Parker-Pope on Prostate Screening

In informed media outlets such as the *New York Times*, the distinction between relative and absolute risk reduction is generally highlighted whenever there is also a downside to the medical procedure being reported.

An example of this caution is present in the article by Tara Parker-Pope (August 6, 2014), entitled “Prostate Cancer Screening Still Not Recommended for All.”

The article gives a lifetime risk of dying of prostate cancer of 3 percent and a drop to 2.4 percent under a PSA testing regime.

Although the absolute risk reduction of .6 percent does represent a 20 percent lower relative risk of dying, it is highly questionable whether this drop is worth the over-diagnosis and over-treatment that it requires.

Mismatched Framing

Because bigger numbers garner better headlines and more media attention, it is expected that relative rather than absolute risks are the norm.

It is especially disconcerting, however, to have potential benefits (of drugs, screening, treatments, and the like) given in relative terms, but harm in absolute terms that is typically much smaller numerically.

The latter has been referred to as "mismatched framing" by Gigerenzer and colleagues – remember that context always counts and that it counts crucially.

Fecal Occult Blood Test Revisited

The issues involved in presenting two probabilities or proportions either as an absolute difference or relatively as a ratio reappears continually when there is a need to assess and report magnitudes.

Remember the Fecal Occult Blood Test illustration: the absolute difference between $P(+CC \mid +FOBT)$ and $P(+CC)$ was a small value of $+.045$ (but still would be one way of stating the degree of facilitation of $+FOBT$ on $+CC$).

As a ratio, however, with respect to the prior probability of .003 for $P(+CC)$, this absolute difference does represent a fifteen-fold change.

So, a relative measure again appears much more impressive than an absolute difference.

The exact same story is told in the illustration for breast cancer screening with mammography.

The Odds Ratio: A Statistic that Only a Statistician's Mother Could Love

It is common in teaching beginning statistics to introduce the terminology of probability by saying that an event, A, occurs with probability, $P(A)$, with the latter represented by a number between zero and one.

An alternative way of stating this fact is to say that the "odds" of A occurring is a ratio, $P(A)/(1 - P(A)) = P(A)/P(\bar{A})$; that is, the probability of the event A occurring to the event not occurring (or equivalently, to $\bar{A}$ occurring).

So, if $P(A) = 2/5$, then the odds of A occurring is $(2/5)(3/5)$ or $(2/3)$, which is read as "2 to 3."

Another interpretation is to note that there are $2 + 3 = 5$ chances for A to occur; and that A occurs in 2 out of the 5 for a probability of $2/5$ $(= P(A))$

Confusing Odds and Relative Risk

Now, consider another event B with $P(B) = 4/5$.

Here, the odds of B occurring is $P(B)/(1 - P(B)) = (4/5)(1/5) = 4/1$, or "4 to 1".

When we take the ratio of the odds of B occurring to the odds of A occurring (that is, $(4/1)(2/3)$), the value of 6 is obtained.

In words, the odds of B occurring is six times greater than the odds of A occurring.

But the real question should be one of how this odds ratio relates to a relative risk of B to A given by $P(B)/P(A) = (4/5)/(2/5) = 2$.

Generally, the odds ratio will be larger than the relative risk; moreover, the odds ratio, because it is such a nontransparent statistic, is consistently (mis)identified in the literature as a relative risk statistic.

To indicate the widespread confusion that exists between relative risk and the odds ratio, part of the abstract of an article is given below that appeared in *Obstetrics & Gynecology* (2001, *98*, 685–688), entitled “An Odd Measure of Risk: Use and Misuse of the Odds Ratio” (William L. Holcomb, Tinnakorn Chaiworapongsa, Douglas A. Luke, & Kevin D. Burgdorf):

OBJECTIVE: To determine how often the odds ratio, as used in clinical research of obstetrics and gynecology, differs substantially from the risk ratio estimate and to assess whether the difference in these measures leads to misinterpretation of research results.

CONCLUSION: The odds ratio is frequently used, and often misinterpreted, in the current literature of obstetrics and gynecology.

Confusion in the Times

A final cautionary tale illustrates the damage that can be done when the media picks up on a story and confuses an odds ratio with relative risk.

We give a short article in the Module One reading that appeared in the *New York Times* (February 25, 1999), entitled "Doctor Bias May Affect Heart Care, Study Finds"; it begins: Unconscious prejudices among doctors may help explain why women and blacks complaining of chest pain are less likely than men and whites to receive the best cardiac testing, a study in today's issue of *The New England Journal of Medicine* suggests.
Sometime later (August 17, 1999), the *Times* published the following "Correction":

A brief report by The Associated Press on Feb. 25 about a study of bias in heart care cited a statistic incorrectly. The study, published in *The New England Journal of Medicine*, showed that doctors were 7 percent less likely to order cardiac catheterization tests for female or black patients than for male or white patients – not 40 percent less likely. The error is discussed in the current issue of the journal. Editors of the journal told the A.P. that they “take responsibility for the media’s overinterpretation” of the study, which used an unusual statistical method.

The “unusual statistical method” referred to is the use of odds ratios.

The article in *The New England Journal of Medicine* (1999, *341*, 279–283) that critiqued the Schulman et al. piece was entitled “Misunderstandings About the Effect of Race and Sex on Physicians’ Referrals for Cardiac Catheterizations” (Lisa M. Schwartz, Steven Woloshin, & H. Gilbert Welch). The abstract and two explanatory paragraphs from this later article are in your Module One reading.

Probabilistic Reasoning and the Prediction of Human Behavior

Module Two that we will touch on tomorrow is devoted to clinical (that is, expert judgement) and actuarial (that is, statistical) predictions of dangerous behavior;

such prediction is of interest for various legal purposes such as civil commitment, or the granting of parole or bail.

As will be shown in that module, and no matter how much society would wish it to be otherwise, we don't do very well in predicting dangerous behavior – or in the vernacular, we generally "suck" at behavioral prediction irrespective of whether it is done clinically or actuarially.

This unfortunate fact remains true in the face of all the "risk assessment" instruments offered and touted in the literature.

Evidence-Based Sentencing

There is now ample evidence that the reliable prediction of human behavior that might be of interest to the criminal justice system is extremely difficult (if not damn near impossible).

There is now the push for evidence-based sentencing (EBS) that depending on the prediction of a future recidivism might change an individual's length of sentence.

The reason given for this push is an analogy to the build-up of the Oakland Athletics baseball team in the early 2000s; here, the argument goes something like the following:

"well, if Billy Beane can get a great team with predictive analytics, we obviously can do the same in the criminal justice context."

For a particular uniformed (by any data) TED talk on this fraught analogy, see Anne Milgram's "Why Smart Statistics Are the Key to Fighting Crime" (filmed October 2013).

A recent and highly informative *Stanford Law Review* (2014, *66*, 803–872) article by Sonja Starr entitled "Evidence-Based Sentencing and the Scientific Rationalization of Discrimination," discusses in some detail the constitutional issues involved in EBS.

In your Module One reading, we also give part of an Attorney General Eric Holder speech (delivered at the National Association of Criminal Defense Lawyers 57th Annual Meeting and 13th State Criminal Justice Network Conference; August 1, 2014) that issues appropriate cautions about EBS.

Approaches to Predicting Human Behavior

There are two general approaches to the prediction of human behavior.

One is through the use of data that pertains to only one specific individual such as age, previous criminal history, and mental status.

The second concerns what particular groups a person might belong to, such as having the BRCA1 genetic mutation, race, sex, and ethnicity.

In legal contexts, the prediction of a specific person's behavior through individual variables like past criminal behavior is typically permissible;

but when prediction is made based on the group(s) one is in, such as race or gender, that usage is usually unconstitutional (see the Federal Rules of Evidence and the distinction between evidence that may be relevant but inadmissible [Rule 403]).

Possible Inadmissible Testimony

There are other methods of prediction that even if not inadmissible in a court of law, should nevertheless be excluded.

One good example would be the labeling done by so-called (clinical) experts that by itself supposedly predicts behavior reliably.

There is the notorious example of James Grigson discussed in Module Two who justified imposing a death sentence under Texas law by simply assigning the label of "sociopath" to a defendant;

in Grigson's view this meant that a perfect prediction of violent behavior was possible, and thus, the defendant should be executed.

A second current example involves evidence-based-sentencing which contends that we can obviously predict recidivism extremely well because of *Moneyball*.

Superpredators

Besides the pernicious assignment of a label such as "sociopath" to a single individual (which at one time in Texas allowed that individual's execution to proceed), a group of criminologists and sociologists in the 1980s engaged in the faulty labeling of a large swath of upcoming teenagers as "superpredators" without any credible evidence whatsoever.

The latter assertion that superpredators were about to emerge, led many states to enact laws permitting the incarceration of children to life without parole.

We give parts of an article in your Module One reading by Gail Garinger from the *New York Times* (March 14, 2012), entitled "Juveniles Don't Deserve Life Sentences," which provides some of the background for these misguided "get tough on crime" efforts:

Unreasonable Expectations for Prediction from Experts

Even though experts might be tasked with the prediction of rare events, there also seems to be the unreal expectation that this should be done perfectly, irrespective of the available evidence.

The reasoning goes that if Billy Beane can do it for the Oakland Athletics, it also must be possible to do close-to-perfect prognostications throughout the criminal justice system.

As an extreme case of such twisted reasoning, there is the Italian judge who convicted seven seismologists of manslaughter when they failed to predict or give a warning for a specific earthquake that occurred on April 6, 2009.

Several paragraphs about this incident are given in your Module One reading from an article by Florin Diacu in the *New York Times* (October 26, 2012) entitled "Is Failure to Predict a Crime?":

Nate Silver Quote

We might end these ideas on predicting human behavior with a clever twist on Reinhold Niebuhr's Serenity Prayer given by Nate Silver in his well-received book, *The Signal and the Noise*:

Prediction is difficult for us for the same reason that it is so important: it is where objective and subjective reality intersect. Distinguishing the signal from the noise requires both scientific knowledge and self-knowledge: the serenity to accept the things we cannot predict, the courage to predict the things we can, and the wisdom to know the difference. (p. 453)