# 8

# Broader Implications and a Bit of Craft Lore

It is difficult to arrive at any overall assessment about the state of statistical learning. Statistical learning is currently undergoing very rapid growth and change. Efforts to evaluate the relative merits of various procedures are further complicated by little agreement on what the performance yardsticks should be: goodness-of-fit, forecasting accuracy, robustness, interpretabilty, computational feasibility, consistency proofs, and so on.

One consequence is that the number of alternative procedures is large, growing, and changing. Another consequence is that there are no clear winners among the list of credible contenders. A final consequence is that because there is little policing of claims made, one must evaluate all assertions about performance with considerable care. This may be especially true about claims about statistical learning procedures made by for-profit enterprises.

Nevertheless, it is probably useful to assemble what craft lore there may be and to provide some general observations and suggestions. Both are offered with some trepidation. They could easily change as more experience is gained.

## 8.1 Some Fundamental Limitations of Statistical Learning

It is important to start at the top. Ideally, what is statistical learning supposed to get done? What are its goals? Recall that in most formal expositions of statistical learning, especially those within a regression framework, there exists in nature an explicit data-generation mechanism. Key features of that mechanism can be represented in a function linking a set of predictors to a response. Then, the function, symbolized by $f(X)$, can be assigned one of two roles. If the conditional distribution of the response is assumed to be the result of a causal process, $f(X)$ represents the causal machinery. It depicts how independent manipulation of each predictor alters the value of the response. Alternatively, the $f(X)$ can be used to describe how the conditional

R.A. Berk, *Statistical Learning from a Regression Perspective*,
DOI: 10.1007/978-0-387-77501-2_8, 

distribution of the response varies with different $x$-values, but with no causal interpretation. One has the conditional distribution, but nothing more.

If either of these accounts is credible, the goal of statistical learning can be to determine the function linking predictors to the response. One must have in the dataset all of the predictors, and each must be very well measured. Then, if the training data can be plausibly treated as a random sample from an appropriate population or a random realization from the relevant stochastic process, many statistical learning procedures can be seen as an effort to obtain a good estimate of the $f(X)$. "Good" can be defined in several different ways, but minimizing generalization error is a common yardstick.

Even under ideal circumstances, however, statistical inference can be problematic. If the goal is to construct confidence intervals, very little may be known about the relevant sampling distribution. If a priori ignorance prevails about the $f(X)$ as well, there may be no sensible hypotheses to test that were posed before the data were examined. Hypotheses generated as part of the data analysis process violate a key assumption of statistical tests. The computed $p$-values will likely be too small.

More fundamentally, one must be clear that there are no formal mathematical results stating that statistical learning procedures will accurately capture the $f(X)$. There also are no proofs of unbiased estimation, and the consistency proofs that exist to date address generalization error for some population model. There is no guarantee that the population model is in any sense "right" or even useful. Moreover, there remains the usual difficulty of figuring out what asymptotic results convey about the results from the data on hand.

In practice, matters are usually worse. One will rarely have all of the required predictors, and it will be rare indeed for all of them to be well measured. Therefore, statistical learning applications will more typically be exploratory and descriptive, and occasionally be the basis for forecasting.

Perhaps most important, one must not see statistical learning as the magic bullet of data analysis. Over the past several decades great promises have been made for any number of statistical procedures, which sometimes proved to be useful, but hardly the revolutions that many of their advocates claimed: SEM, ARIMA, HLM, robust regression, raking, specification tests, sliced inverse regression, Rasch models, ARCH, meta-analysis, life history analysis, log-linear models, latent class analysis, and on and on. So that there be no mistake, no such claims are being made here about statistical learning.

## 8.2 Some Assets of Statistical Learning

The benefits of statistical learning can be organized into three large categories. Some of the benefits boil down to an attitude adjustment. Others come from an ability to better address certain kinds of data analysis tasks. A final set involves improving other statistical procedures.

### 8.2.1 The Attitude Adjustment

Statistical learning may be seen as implicit criticism of conventional causal modeling that has for several decades dominated much research in the social and life sciences. There are certainly many overt critiques of business as usual (Box, 1976; Leamer, 1978; Rubin, 1986; Freedman, 1987; Manski, 1990; Heckman, 1999; Breiman, 2001b: Berk, 2003: Freedman, 2005). Statistical learning, by contrast, delivers its message by example. And this message has several related themes.

First, statistical learning can proceed quite happily without worrying much about cause and effect. Inputs are associated with outputs, but the inputs are not necessarily causes and the outputs are not necessarily effects. One can use causal insights to help determine inputs and outputs when prior information about cause and effect is available. But, within the framework emphasized in this book, outputs are nothing more than variables whose conditional distributions are of interest. Inputs are nothing more than the variables to be used in the conditioning. Consequently, the inputs may be selected to help in classification or forecasting even if they play no causal role whatsoever.

Second, the data do not have to be generated by a real intervention. This means that randomized experiments, for example, usually deliver no special leverage. There is also no formal need to proceed as if the predictors in an observational dataset were treatments that could be manipulated. It is often said that experiments are the gold standard for causal inference. Unless determining the effects of causes is a key goal of a statistical learning data analysis, the underpinnings and tools of causal inference are likely to be irrelevant

Third, the most important statistical benchmark for a successful data analysis is successful forecasting. If fitted values correspond well to observed values one may be on the right track. But a good fit is not good in and of itself. Rather, it implies that the forecasts may be good as well. This means that goodness-of-fit statistics and tests are in statistical learning rarely of much interest. The broader message is that one must look beyond the data used to build a model in order to determine if that model has genuine merit.

Fourth, with forecasting skill as the standard, many statistical learning procedures perform well, and very often substantially better than conventional causal modeling. One reason can be that a statistical learning procedure may use more of the information in a dataset than a conventional model does. Another reason can be that the conventional models represent the associations between the predictors and the response less well than an inductive model does. Finally, statistical learning procedures are often designed to maximize forecasting skill. Conventional models rarely are.

At the same time, there is nothing in statistical learning that precludes causal thinking. With conventional causal modeling, the regression coefficients associated with each predictor are supposed to reveal what the average change in the response would be if the predictor values for a given observation were actually altered (e.g., a person with no high school degree obtained one). In statistical learning, partial response plots make no such claim. They simply show how, on the average, the response variable varies depending on the value of a given predictor, with the values of all other predictors fixed at their current levels. Likewise, plots of predictor importance are not plots of regression coefficients; they do not represent causal importance. And one must not think that tree diagrams are anything like path diagrams.

Partial plots, importance plots, classification plots, and tree diagrams can, however, provide some ideas about how the response variable might change if a predictor were manipulated. These can be very useful for designing future studies to get directly at possible causal relationships (e.g., with a randomized experiment). Partial plots, importance plots, classification plots, and tree diagrams can also generate new ideas and even theoretical insights. In short, statistical learning results can be useful for understanding possible causal relationships but unlike causal models, they are not meant to be surrogates for real experiments.

Finally, perhaps the most important attitude adjustment is that description is a noble and useful scientific enterprise. One can do good science and not do causal modeling or experiments.

### 8.2.2 Selectively Better Performance

Although statistical learning of the sort described in this book is certainly more than a niche player, it performs far better at some tasks than others. Here are some tasks at which statistical learning can excel.

1. Determining Functional Forms—When the response functions are not known but are likely to be highly nonlinear, statistical learning procedures can shine. Even if no claims are made that the true $f(X)$ has been determined, important information about that function may be revealed. For example, there may be good evidence that the function is roughly S-shaped and for which $x$-values the response is changing especially rapidly. This can be extremely instructive, even if the particular S-shaped curve (e.g., cumulative normal v. cumulative logistic) cannot be identified. Recall the many earlier examples in which unanticipated nonlinear functions were found.
2. Discovering Unexpected Predictors—The ability of many statistical learning procedures to exploit a large number of inputs means that some predictors, or transformations of predictors, that might have unanticipated relationships with the response variable can be found. This is true for CART, but especially boosting and random forests. For example, in research on racial bias in decisions to charge with a capital crime, some unexpected two-way and three-way interactions are sometimes needed (Berk et al., 2005a).
3. Discovering Which Predictors Matter—Even if there is a broad consensus about the predictors that need to be included in the training data, there

may be little agreement about how important each of the predictors is. If one finds contribution to the fit or to forecasting skill a useful definition of "importance," several statistical learning procedures can provide instructive measures of predictor importance. For example, particularly for the most serious form of prison misconduct, sentence length delivers the most forecasting skill (Berk et al., 2006).

4. Providing Useful Regression Diagnostics—By being able to find useful predictors whose roles were unanticipated and by being able to reveal unanticipated response functions, statistical learning procedures can serve as very useful diagnostic tools for parametric regression models (Berk et al., 2005a). More is said about this shortly.
5. Avoiding or Compensating for Overfitting—In exploratory data analysis, overfitting is almost unavoidable. But bagging can correct in part for overfitting, and random forests does not overfit. Boosting can overfit in principle, but as commonly used does not seem to overfit significantly. Moreover, statistical learning is embedded in a statistical tradition where concerns about overfitting and tools to counter overfitting are very salient. There are several measures of fit adjusting for degrees of freedom, cross-validation, the use of test data, and a variety of regularization methods.
6. Forecasting—Especially if the data do not constitute a time series, random forests and boosting are probably state-of-the-art forecasting tools. For example, there seems to be some success to be had using random forests to forecast future incidents of serious domestic violence (Berk et al., 2005b)
7. Responding to asymmetric costs—All of the procedures discussed require a commitment to a particular costs/loss function. But when the response variable is categorical, CART and random forests are especially able to take account of asymmetric costs. For many applied problems, asymmetric costs are critical because the costs of false negatives can be very different from the costs of false positives. Many examples were provided in earlier chapters. Asymmetric costs for quantitative variables are at a very early stage. But quantile random forests has some promise as do methods being developed for stochastic gradient boosting.
8. Exploiting Many Predictors—When there are a very large number of predictors, most procedures attempting to link inputs to outputs will stumble, especially if the number of predictors exceeds the number of observations. CART and random forests are particularly strong in this regard and can handle hundreds of predictors even if there are far more predictors than cases. The main limitation is not the statistical procedure, but computing power. Some have claimed that because SVM works with the cross-products between observations, the curse of dimensionality is lifted. This turns out to be at least an exaggeration (Hastie et al., 2001: 384–385).

### 8.2.3 Improving Other Procedures

The results of a statistical learning analysis can be a useful intermediate step in another statistical procedure. The key attribute that statistical learning brings to these tasks is an ability to construct fitted values that can correspond especially well to the actual values. When a good fit is very important, statistical learning procedures can be especially effective.

A useful example is in the construction of propensity scores (Rosenbaum, 2002; McCaffrey, 2004). If one is interested in the causal impact of some categorical intervention, such as a school program to enhance reading, and if that intervention is not provided through random assignment, there will be the substantial likelihood of selection bias. In the school illustration, students who participate in such a program will probably differ on the average from students who do not. These "pre-existing" differences are then confounded with estimates of the impact of the reading program.

However, if an unbiased estimate of the probability of program participation can be constructed, that estimated probability can be used as an effective statistical control through matching or other means (Imbens, 2004). The probability of program participation is called a "propensity score." In principle, using propensity scores to adjust for selection bias in estimates of the treatment effects can be an effective tool.

For example, propensity scores can be used as weights in what is called the "difference-in-differences" estimator (Heckman et al., 1998). The data analysis problem is still the same: selection bias into the experimental and control groups. But now there is a pretest and a posttest. The difference-in-differences estimator compares the members of the experimental group to members of the control group through the change in their performance between the pretest and the posttest, with those changes weighted by propensity scores. If those propensity scores are estimated in a consistent manner, the difference-in-differences estimator will in principle deliver a consistent estimate of the treatment effect. (But see Freedman and Berk, 2008).

Usually propensity scores are estimated with logistic regression, and one always has to wonder how close to unbiased the fitted values really are. Statistical learning procedures can improve the quality of those fitted values by making them closer to the actual values. Then, instructive estimates of the treatment effects may be more likely. But the issues are a little tricky. Conventional boosting procedures applied to classification problems would seem to be a tool of choice, but it risks pushing any estimated probabilities away from .50 toward 0.0 and 1.0. Recall that an alternative approach was discussed.

Another useful asset of statistical learning procedures is the very flexible functional forms that are often constructed. These, in turn, can be instructive as conventional regression models are built. As noted earlier, statistical learning procedures may reveal predictors or functional forms not anticipated by conventional regression models. These models might then be revised accordingly. Conversely, predictors that seem to be important within a conventional regression model may vanish when more appropriate functional forms are used. This might suggest dropping such regressions from the regression analysis.

Yet another example is improving covariance adjustments. Recall that covariance adjustments for a given predictor depend on residualizing that predictor and the response with respect to all other predictors. That is, any linear dependence between the response and the other predictors is removed and any linear dependence between the given predictor and the other predictors is removed. But the quality of this "purging" depends on using the right predictors with the correct functional forms. Many statistical learning procedures can exploit hundreds of predictors and search for the functional form that fits the data best. Then, one option is applying statistical learning tools to residualize the response and key predictors for the "nuisance" covariates before the important relationships are examined.

## 8.3 Some Practical Suggestions

Just as for any other set of statistical procedures, practice is guided significantly by craft lore. In that spirit, we turn to a bit of craft lore about the use of statistical learning. It is important to keep in mind, however, the craft lore can change dramatically with experience, and the experience with statistical learning to date is somewhat spotty.

### 8.3.1 Matching Tools to Jobs

To begin, it can be useful to reconsider which procedures are likely to be most effective for which data analysis tasks. The smoothers discussed in Chapter 2 are primarily visualization aids that can be applied in a wide variety of settings. They can be taken as standalone tools, as when one smooths a two-dimensional or three-dimensional scatterplot. They can be used as a component of other procedures, such as the generalized additive model. Their main strength is providing information in a very accessible manner about how predictors are related to a response variable.

If the primary goal is good fit, accurate classification, and/or accurate forecasting, the procedures discussed in Chapters 4 through 7 are likely to be a better choice: random forests, boosting, and support vector machines. CART can be a handy way to hunt for possible interaction effects and can serve as a intermediate step for more powerful statistical learning procedures. But otherwise, CART has largely been superseded.

Random forests, boosting, and support vector machines can all perform well. It is not clear yet which perform better for which kinds of datasets, or even if the differences in performance are likely to matter a great deal in practice. It is easy to get caught up in differences of a few percent in forecasting accuracy, which are too small to matter and may not hold up.

For example, suppose the goal is to forecast which high school students are at the greatest risk for dropping out. A classifier might be trained on data from several cohorts of high school students and evaluated with test data that are a random sample from the same population. But as a practical matter, the forecasting tool developed would be applied to new cohorts of students that would likely differ from the training and test samples by more than random sampling error. One might well expect a gradual drift in the background of incoming freshmen and the mix of incentives to remain in school. As a result, forecasting accuracy could decline by an amount that could easily swamp the performance differences between competing classifiers, and a new analysis might show that the classifier that had previously performed best no longer did. In short, differences in performance that may be of methodological interest may also be of no practical importance.

Therefore, one key factor in choosing between statistical learning tools can be the quality of the output. To date, there are important differences among random forests, boosting, indexboosting and support vector machines beyond the predicted values and a confusion table. If one needs to examine response functions and evaluate predictor importance, an implementation of support vector machines may not have what is needed. There are implementations of random forests and boosting that do, although at the moment, the measures of variable performance in random forests, which exploit the out-of-bag data for forecasting, is probably more desirable.

Another key factor in choosing among statistical learning tools is their ability to address in a flexible manner the relative costs of false negatives and false positives. Currently, random forests is likely to do this better for classification problems than either boosting or support vector machines. More generally, asymmetric loss functions can be important. There are very recent developments for random forests and stochastic gradient boosting that have the promise of allowing for certain special cases.

One should also consider the relationship between the sample size and the number of predictors. If the number of predictors exceeds the number of observations, random forests may be the only viable choice among the better-performing statistical learning procedures. There may even be reason to prefer random forests if the number of predictors is large compared to the number of observations, even if there are fewer predictors than observations.

A final factor is the range of response variables that can be properly analyzed. At this point, boosting may be the most flexible, at least within the gradient boosting approach. But there seems to be no reason in principle why random forests and support vector machines cannot be made more broadly applicable, and it is likely that the range of response variable types that can be handled by these procedures will increase over the next several years.

### 8.3.2 Getting to Know Your Software

There is not yet, and not likely to be in the near future, a consensus on how any of the various statistical learning procedures should be implemented in software. For example, a recent check on software available for support vector machines found working code for over a half dozen procedures. There is, as well, near anarchy in naming conventions, and notation. Thus, the term "cost," for instance, can mean several different things and a symbol such as $\gamma$ can be a tuning parameter in one derivation and a key feature in another derivation.

One cannot assume that a description of a procedure in a textbook or journal article corresponds fully to software using the very same name. Consequently, it is very important to work with software for which there is good technical documentation on the procedure and algorithms being used. There also needs to be clear information on how to introduce inputs, outputs, and tuning parameters. Two computer programs can use the same name for different items, or use very different names for the same item. And in either case, the naming conventions may not correspond to the naming conventions in the technical literature.

Even when the documentation looks to be clear and complete, a healthy dose of skepticism is useful. There are sometimes errors in the documentation, or in the software, or in both. So, it is usually important to "shake down" any new software with data that have previously been analyzed properly to determine if the new results come out as expected. In addition, it is usually helpful to experiment with various tuning parameters to see if the results make sense. In short, *caveat emptor*.

It is also very important keep abreast of software updates, which can come as often as two or three times a year. As a routine matter, new features are added, bugs fixed and documentation rewritten. These changes are often far more than cosmetic. Working with an older version of statistical learning software can lead to unnecessary problems.

Finally, a key software decision is whether to work primarily with shareware such as R or with commercial products. The tradeoffs have been discussed earlier at various points. Cost is certainly an issue, but perhaps more important is the tension between having the most current software and having the most stable software and documentation. Shareware is more likely to be on the leading edge, but often lacks the convenience and stability of commercial products. One possible strategy for individuals who are unfamiliar with a certain class of procedures is to begin with a good commercial product, and then once some hands-on skill has been developed, migrate to shareware.

### 8.3.3 Not Forgetting the Basics

It is very easy to get caught up in the razzle-dazzle of statistical learning and for any given data analysis, neglect the more simple fundamentals. All data explorations must start with an effort to get "close" to the data. This requires a careful inspection of elementary descriptive statistics: means, standard deviations, histograms, cross-tabulations, scatterplots and the like. It also means understanding how the data were generated and how the variables were measured. Moving into a statistical learning procedure without this groundwork can lead to substantial grief. For example, sometimes numeric values are given to missing data. Treating these values as legitimate measures can seriously distort any data analysis, including ones undertaken with statistical learning.

It will usually be helpful to apply one or more forms of conventional regression analysis before moving to statistical learning. One then obtains an initial sense of how good the fit is likely to be, the likely signs of key relationships between predictors and the response, and hints of problems that might be more difficult to spot later (e.g., high correlations among some predictors). An important implication is that it will often be handy to undertake statistical learning within a larger computing environment in which a variety of statistical tools can be applied to the same data. This can weigh against single-purpose statistical learning software.

To take one simple example, a tuning parameter in random forests may require a distinct value for each response class. But the order in which those arguments are entered into the function for the tuning parameter may be unclear. In the binary case, for example, which category comes first? Is it $\omega = c(1,0)$ or $\omega = c(0,1)$? It is easy to make the wrong choice. Random forests runs just the same and generates sensible-looking output. But the analysis has not been tuned as it should have been. It can be difficult to spot such an error unless one knows the marginal distribution of the response variable and the likely sign of relationships between each predictor and the response. A few cross-tabulations and a preliminary regression analysis can help enormously.

To take a little more complicated example, one of the few graphical displays of output from support vector machines depends on specifying a slice of the control variables used to the subset of the data so that a plane can be plotted. If there are several control variables, it is easy to choose a slice in which there are no data, or too few observations to construct a useful plot. Yet, no error message may be produced, and misleading interpretation can follow. A series of cross-tabulations or conditional plots can help a lot.

### 8.3.4 Getting Good Data

As noted many times, there is no substitute for good data. The fact that boosting, for example, can make a weak classifier much stronger, does not mean that boosting can make weak data stronger. There are no surprises in what properties good data should have: a large number of observations, little measurement error, a rich set of predictors, and a reasonably well-balanced response variable distribution. The clear message is that it is very important to invest time and resources in data collection. One cannot count on statistical learning successfully coming to the rescue. Indeed, some forms of statistical learning are quite fragile and easily pulled off course by noisy data, let alone data that have systematic measurement error.

The case for having legitimate test data is a bit more ambiguous. Statistical learning procedures that use out-of-bag data or the equivalent do not formally need a test dataset that is a random sample from the same population as the training data. The out-of-bag observations serve that purpose. But most statistical learning procedures currently are not designed to work with random samples of the data, even when that might make a lot of sense. Therefore, having access to test data is usually very important.

Even for random forests, test data beyond the out-of-bag observations can come in handy. Comparisons between how random forests performs and how other approaches (including conventional regression) perform are often very instructive. Yet such comparisons cannot be undertaken unless there are test data shared by all of the statistical procedures applied. Finally, having a true test dataset can help a great deal if random forests is applied repeatedly to the same training data after changes in the tuning parameters. At the very end of the tuning process, there is still the opportunity to get an honest measure of performance from data that until that moment have not been used.

### 8.3.5 Being Sensitive to Overtuning

We have discussed several summary measures of model performance that can be used to help tune models. Tuning can be an important process in model development. However, if the tuning process goes on for very long, the desirable properties of the summary measures can be badly diluted.

For example, neither the AIC or BIC take into account the number of parameters estimated for earlier models that were considered and rejected. Cross-validation can be compromised when the same dataset is used over and over. The independence between the training data and the test data is gradually lost.

These concerns suggest a strategy, just noted, in which there is a hold-out sample, or another random sample from the same population, that is used at the very end to evaluate the final model. Ideally there would be a large number of hold-out or random samples that could be used for tuning as well. In addition, it will generally be a good idea to show great restraint when there is an opportunity to tune, particularly when the number of observations is relatively small and the number of predictors is relatively large.

### 8.3.6 Matching Your Goals to What You Can Credibly Do

Much of the literature on statistical learning is formulated around some $f(X)$. There is a real mechanism by which the data were generated. A key goal of a data analysis is to recover the data-generation function from a dataset. It can be very tempting, therefore, to frame all data analyses in a similar manner.

But, one of the themes of this book has been that in reality, more modest goals are likely to be appropriate. Perhaps most important, one will not have access to all of the requisite predictors, let alone predictors that are all well measured. In addition, various kinds of data snooping will often be difficult to avoid, and even the best adjustments for overfitting may prove insufficient. For these and other reasons, description will be what the data analysis is really about.

This does not mean that the stability of one's results cannot be usefully addressed. It also does not mean that causal thinking is unimportant. And it certainly does not mean that one cannot achieve an improved understanding of what the $f(X)$ might be. For example, entire classes of functions can sometimes be effectively eliminated.

But what it does mean is that much of the formal rationale for any statistical procedure, including statistical learning procedures, cannot be relied upon. It can be very difficult to know, for example, what use to make of proofs of consistency. It also means that packaging one's results as function estimation or as a model of how the data were generated can be false advertising. If description is the enterprise, it needs to be labeled as such.

## 8.4 Some Concluding Observations

Statistical learning has considerable potential, and its reach and power will likely increase in the next several years. But with that potential comes almost certain misuse. There are already some instructive examples.

As just noted, it can be very seductive to proceed as if the goal were to estimate $f(X)$ even when one does not have the requisite predictors or predictors of the requisite quality. Then, the actual work being undertaken is description. Concepts such as bias and consistency no longer apply and cannot be appealed to. And the work cannot properly be packaged as function estimation.

Another error is to undertake statistical inference as part of a statistical learning analysis when the $p$-values are not likely to make much sense. The $p$-values may be wildly misleading because the data are not a random sample or random realization of anything, because the statistical learning procedure invalidates the required assumptions, and/or because the necessary sampling distributions are unknown or not credibly estimated.

Still another error is to accept statistical learning results uncritically. The very flexibility with which fitted values are constructed can lead to results that are factual nonsense. There is also the possibility that software will malfunction or be fundamentally flawed. In general, all results must pass the sniff test of subject matter credibility.

Finally, data snooping can lead to significant data analysis errors. Even data analysts who are well aware of its risks can inadvertently lower their guard and allow data snooping errors to affect their results. For example, a number of different applications of random forests, using different tuning parameters, may be examined before selecting a single model and the particular values of its tuning parameters.

The concluding message, therefore, is to users of statistical learning results. At the very least, demand that all results to be taken seriously rest on test data or their equivalent. And if the results do not make subject matter sense, skepticism is a sensible stance. Ask that each step in the data analysis, including how the data were collected, be reviewed. If anomalies persist, consider getting an independent third party involved. As with any new and complicated procedure, there is lots of room for mistakes and even a substantial amount of fakery.