Bias, Biostatistics, and Prognostic Factors

Mithat Gönen, PhD

Key Words: Prognosis, Prognostic factors, Statistics, Outcomes, Bias.

(J Thorac Oncol. 2011;6: S1705–S1709)

Prognostic factors are essential for patient management and counseling. Most cancer patients receive treatment based on the presence or absence of several risk factors. The most prominent example of this is tumor, node, metastasis staging, but many diseases are moving beyond the simple one-size-fits-all stage classification and using stylized prognostic factors. A notable difficulty in the development of prognostic factors in today’s clinical research environment is the lack of implemented guidelines and accepted practices. As a result, it is not clear what constitutes a prognostic factor. This article reviews some of this practice, identifies the various biases that arise from it, and points to solutions, many of which are discussed in detail at other forums but have been somewhat ignored, to this point, in the thymoma literature.

The clinical importance of reliable prognostic factors requires little justification. If one can predict the course of disease with reasonable precision, treatment choice, follow-up, and patient counseling are greatly enhanced. As a result, prognostic markers abound in all diseases including thymoma.1–7 There is no question that our ability to accurately prognosticate has improved by leaps and bounds during the past 2 decades. Yet, only a handful of markers have made it into international guidelines for diagnosis, treatment, or follow-up. This article will discuss some of the reasons behind the failure of many prognostic markers to deliver and offer some suggestions to avoid similar disappointments in the future.

Several examples I will use, and the arguments that will follow have been made previously elsewhere,8–13 although not necessarily in the context of thymoma. Some of the examples in this study are from diagnostic factor studies; this is because statistical concerns, especially about bias, have much in common between diagnostic and prognostic factor studies. Furthermore, diagnostic factor examples often have the advantage that they are simple and available, and there is more widespread knowledge about the factor or disease or both. For the same reason, some of the general references that discuss methodological issues in biomarker development and early detection of cancer are also relevant in this study.

ISSUES IN DEFINING PROGNOSTIC FACTORS

Bias in Prognostic Factor Studies

Bias is one of these overused terms that ceased to have a precise definition. For the purpose of this article, bias refers to a systematic difference between the sample and the population. If one were to use patients on Medicare to study the prognosis for papillary thyroid cancer, there would be a substantial bias as the median age at diagnosis is less than 45 years, yet the sample is, by definition, restricted to those aged 65 years or older. This difference is called systematic in the sense that it is not because of sampling error. Differences because of sampling error tend to get smaller as the sample size increases. Bias, conversely, does not dissipate in larger samples.

The thyroid cancer example suffers from the malady of all didactic analogies: it is too obvious and risks overshadowing the mechanism for more subtle but still important biases. Consider a hypothetical example of “New Prognostic Factor” (NPF), a novel tissue marker that is believed to be over expressed in certain malignancies. After several in vitro studies, a retrospective clinical study in thymoma reported on its potential as a prognostic factor. The investigators found that the presence of NPF in resected tissue was correlated with poor survival (p < 0.05). This was followed by three other similar studies from other institutions. All the three studies were also conducted retrospectively on patients who underwent surgery. The largest of these three studies (also larger than the original study) confirmed the significant survival difference between patients who expressed NPF and those who did not. The other two studies did not report statistically significant findings, but they also had smaller sample sizes. Dismissing these two studies as underpowered and claiming that the larger study can be taken as a validation, the investigators of the original study planned a prospective study where NPF was measured first and correlated with outcome subsequently. They were disappointed to see that the presence of NPF was not significantly correlated with survival.

Could this be due to a bias in the original retrospective study? One can only speculate, but the description of the patients gives a clue. The retrospective studies were carried...
out from stored tissue. It is entirely possible that the tumors that were available in the tissue bank were larger than the typical tumor to have been stored at the first place and then to have survived the onslaught of the demands on research specimens from multiple studies. If this is indeed correct, it would introduce a systematic difference between the sample (tumors in the tissue bank) and the population (all patients with thymoma). Because the prospective study mandated tissue collection on all enrolled patients, it had less (or no) bias. The difference in conclusion between the original retrospective studies can possibly be due to bias.

Although hypothetical, the scenario has many realistic elements: most prognostic factors are born out of retrospective studies, which have findings that are replicated by some, but not all, other investigations. Most such prognostic factors fail the more stringent criterion of prospective confirmation. All retrospective studies are prone to bias—in fact, to multiple sources of bias. Although bias is a problem in prospective studies as well, its magnitude tends to be smaller, and its mechanism is better understood.

Types of Bias Common in Prognostic Studies

Selection of Cases

The bias introduced by case selection is perhaps more appropriately named patient selection or selection bias. This bias refers to inclusion of patients in such a way that the resulting sample is extreme (either too good or too bad) in terms of risk of disease or risk of failure.

An example might help with the concepts. Carcinoembryonic antigen (CEA), by most measures an accepted prognostic marker for patients with colon cancer, attracted widespread clinical attention when it was reported that 35 of 36 patients with colon cancer who were studied had increased levels of CEA resulting in 97% sensitivity. This suggested some role for CEA as a diagnostic marker. A decade later, the picture was much less promising, with sensitivities of 5, 25, 45, and 65% reported for stages I, II, III, and IV, respectively. This effectively ruled out the use of CEA as a diagnostic marker. Even among stage IV patients, the difference between 97 and 65% is simply too big to be attributed to a single factor. One reason, however, that is likely to explain this substantial difference is the selection of patients used in the original study by Thomson et al. Although this original PNAS publication is short on details of who these patients were, it is likely that they all had very advanced cases of colon cancer. This is one of the common sources of selection bias.

Another common source of selection bias is already exemplified in the previous section. In the discussion of the availability of tissue for analyzing the NPF, I have already pointed out the possibility that tumors that were included in the study were bigger than those that are not included. At a very basic level, the source of the bias in the CEA and NPF examples is the same. Yet at another level, they are quite different. More often than not, proof of principle studies like the CEA study deliberately choose an extreme sample. Yet, there was no such intention in the NPF study—the investigators gratefully took whatever was available in the tissue bank. It is this aspect of selection bias that makes it dangerous: the fact that there is no other option for the source of study material does not mean that the sample is not biased.

Selection of Controls

Another type of selection bias can arise in determining who is eligible as a control in a case-control study. Ideally, the only difference between a case and control should be the factor under study (exposure, disease, treatment, etc). In practice, it is nearly impossible to find such controls in observational settings. An example from the use of serum peptides to detect prostate cancer might illuminate the difficulties. Cases were 25 men with biopsy-verified prostate cancer, and controls were men who are thought to be free of prostate cancer because either they were younger than 40 years or had undetectable prostate specific antigen (PSA) levels in their blood. On the one hand, the level of evidence required to be a control is much weaker than that of a case. On the other hand, it is impossible to mandate biopsies on the controls to rule out occult malignancies. The authors supplemented the controls by including young men who are very unlikely to have prostate cancer, but by doing so, they introduced another factor that is different between the two groups: age. To attribute the differences between the serum peptide characteristics of the cases and controls now requires either the knowledge or the assumption that age is unrelated to serum peptide measurements.

This example also points out how frustrating it can be to try to select controls without bias. This is one reason why most studies prefer to perform some sort of matching. Although this does not guarantee removal of all bias, it certainly makes an attempt to decrease it.

Double Dipping

A more scientific name for this kind of bias could be overfitting, but double dipping summarizes the source of the bias so clearly that it is preferable in an expository article. Double dipping refers to the bias created by using the same dataset for multiple, related analyses. Although there are many ways one can double dip, one is especially relevant for prognostic factor studies happens when continuous variables are categorized by thresholds selected from the same dataset.

It is difficult to find details of published examples of this type of bias; hence, an unpublished example will be used in this study. Although the data and the scenario are real, the analyses shown in this study were carried out only for the purposes of demonstration in this article. The clinical question posed is whether a change in uptake from baseline to midcycle positron emission tomography scans is prognostic for pathological response in the setting of neoadjuvant therapy. The clinical utility of such a prognostic factor is clear: a patient predicted to have poor pathological response at the end of the neoadjuvant treatment may switch treatments or proceed to surgery early, sparing valuable time and side effects.

The investigators collected data as a part of an ongoing clinical trial and reported their findings as seen in Tables 1 and 2. Most of the patients with a large decrease in SUV
responded and vice versa. In fact, there was only a single patient with a decrease in SUV more than 35% who nevertheless ended with a response of less than 60%. Summary statistics reported in Table 2 are very encouraging: SUV is 100% sensitive and 90% specific for pathological response. In addition, the negative predictive value is estimated to be 100%, that is, if a patient’s SUV does not decrease, there is a virtual guarantee of lack of response.

Most readers of Tables 1 and 2 object to these conclusions by pointing to the small sample size and the width of the subsequent (unreported) confidence intervals. This is certainly a concern. But even with a large sample size, there are other reasons to object to this style of thinking. Consider the underlying (noncategorized) data given in Figure 1. Ignoring the dotted lines for a moment, there is some correlation between percent decrease in SUV and treatment response. The upper left corner of the plot is mostly blank, and in very general terms, a higher decrease in SUV seems to correspond to a higher treatment response. The correlation between decrease in SUV and treatment response is 0.50, a respectable number by most standards. Fitting a line to this set of data reveals that a stable SUV (no change from baseline) corresponds to a predicted response of 26%, and each additional 10% decrease in SUV translates to an estimated increase of 6% in treatment response. There is a reason to think that a midcycle positron emission tomography scan contains some predictive value for the ultimate treatment response.

Nevertheless, the analysis in Table 2 is an overstatement of the evidence provided by the data. The thresholds, plotted with the dotted lines in Figure 1, reveal that they are selected to minimize the number of misclassified patients. This would be the obvious way to select thresholds, yet Figure 1 makes clear how sensitive the numbers in Table 2 are to the selected thresholds. These thresholds are very unlikely to survive an evaluation on an independent dataset and so are the estimates of sensitivity, specificity, and positive and negative predictive value. Even a minor change in the thresholds, as shown in Table 3, has a substantial effect on the reported parameters (and could significantly alter the study conclusions). Yet, reporting numbers such as the ones in Tables 1 and 2 instead of Figure 1 and associated analyses of correlation, intercept, and slope pave the way to strong conclusions about the prognostic value of SUV. This is partly not only the result of a small dataset but also the consequence of double and triple dipping and therefore being overly optimistic.

Double dipping can happen in other ways too. If one is fitting a model with many variables with the idea of coming up with a prognostic score, the resulting model usually is too much tuned to the data at hand (the term overfit is usually reserved for this situation). Evaluating the predictive properties of a model on the same dataset that it is derived from is sure to lead to optimistic conclusions.20 Statistical methods have been devised to decrease these optimistic conclusions,21,22 but they have not been in wide-spread use and are definitely not routinely expected by journal editors and referees. Even the best of these methods is unlikely to remove all the optimism because of overfitting, and the ultimate arbiter should be validation in an independent dataset.

**Statistical Analysis in Presence of Bias**

It is very important to make this point repeatedly: it is very difficult, if not outright impossible, to remove bias by clever statistical analysis. At best, bias can be reduced with statistical modeling at the cost of making additional assumptions, which could, at times, be unverifiable from the observed data. Some of these are discussed in various guidelines for biomarker studies23 and influential textbooks.24

As an example, consider the hypothetical example of NPF and suppose that at the time of the original retrospective study, the investigators suspected that they are dealing with a sample biased by tumor size. Instead of simply associating

---

**TABLE 1.** Correlation of a Decrease in SUV after Neoadjuvant Therapy with Pathologic Response

<table>
<thead>
<tr>
<th>Response</th>
<th>SUV &gt;35%</th>
<th>SUV ≥60%</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decrease in SUV &gt;35%</td>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Decrease in SUV &lt;35%</td>
<td>0</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>Total</td>
<td>3</td>
<td>10</td>
<td>13</td>
</tr>
</tbody>
</table>

SUV, standardized uptake value.

**TABLE 2.** Outcome Parameters for a Decrease in SUV of >35% in Predicting Pathological Response to Neoadjuvant Therapy

<table>
<thead>
<tr>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Positive Predictive Value</th>
<th>Negative Predictive Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>100%</td>
<td>90%</td>
<td>75%</td>
<td>100%</td>
</tr>
</tbody>
</table>

SUV, standardized uptake value.

---

**FIGURE 1.** Individual data points of changes in SUV after neoadjuvant therapy and pathologic response.
NPF with outcome in what is commonly called a univariate analysis, they have the option of including size and NPF in a multivariable analysis. This allows them to claim “size-adjusted” results and possibly draw stronger conclusions. Few such analyses draw attention to the required assumptions for such a model. First and foremost, one has to choose an appropriate multivariable regression model. Although Cox model has become standard in oncology mostly because of the modest set of assumptions it requires, it still forces the proportional hazards (PH) assumption to hold. PH can be thought of as the condition that the magnitude of the correlation between a factor and outcome (often represented as a hazard ratio) does not change over time. In my experience, most clinical investigators have difficulty grasping the meaning of this assumption, and most data analysts, perhaps for lack of a widely accepted alternative method, tend to accept the assumption without critical review.

In addition to proportional hazards, there is the functional form of the model to consider. Size, by virtue of being a continuous variable, can be used as is in a model, that is, as a continuous covariate. Although it is standard to assume a linear effect for size, which implies that the incremental effect of 1 Unit (say centimeter) increase in size is the same regardless of what the base size is, it is more realistic to think that the actual effect is more S shaped than a line with a relatively flat profile for very small or very large tumors. Most studies, without acknowledging the consequences, accept the linear form with the (perhaps optimistic) thought that the middle part of the S is what counts and that it can be approximated by a line. If, however, the concern was that the study originally included tumors that were bigger than what would be expected from a random sample, then the part of the curve that is most influential in an adjusted analysis would be the flat portion on the left (i.e., the part that corresponds to small tumors). In anticipation of this critique, some investigators choose to categorize size. Another argument in favor of categorization is that it makes it easier to interpret and communicate the results. The devil, as demonstrated in the previous section, is in the details: choosing a cutoff point(s) for a threshold and then carrying out the analysis from the same dataset will result in overfitting (double dipping).

**TABLE 3.** Effect of Changes in Thresholds on Outcome Parameters

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Positive Predictive Value</th>
<th>Negative Predictive Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUV &gt;35%; response &gt;60%</td>
<td>100%</td>
<td>90%</td>
<td>75%</td>
<td>100%</td>
</tr>
<tr>
<td>SUV &gt;30%; response ≥60%</td>
<td>75%</td>
<td>78%</td>
<td>60%</td>
<td>87%</td>
</tr>
</tbody>
</table>

SUV, standardized uptake value.

**FIGURE 2.** Phases in the identification and evaluation of a robust clinically applicable prognostic factor.

In this article, I have outlined some of the most common sources of bias in prognostic factor studies and their potential effects on the conclusions. This is not the first time these biases are identified or discussed in detail. There are numerous articles that serve the same purpose. However, the continued presence of these biases in many published works and the failure to discuss the implications of these issues point out to the difficulty of reaching the clinical investigators who carry out these studies.

These biases, strictly speaking, are not purely statistical issues. Yet, they are identified with statistics, partly because statisticians tend to identify them more often and try to represent or perhaps correct them by modeling their sources. In my opinion, dealing with biases post hoc is inefficient, insufficient, and frustrating. It is better to be thoughtful at the design stage and bring into the discussion various members of the team including statisticians, pathologists, radiologists, etc. This will likely delay the data collection but will speed up the analysis and interpretation of the results, not to mention the credibility of the conclusions.

The responsibility to notice bias should not be shoul-dered only by the investigators themselves. Editors, referees, and readers should demand high-quality reports in this field that follow the principles of good clinical science. One simple point (low-hanging fruit) in this respect is the timing of when a factor is crowned with the adjective “prognostic.” Figure 2 shows the typical stages of development for a prognostic factor. Current practice is to label the factor as prognostic at point A, right after the clinical development and certainly before validation. Some of these prognostic factors are never validated, some fail validation but their titles are not revoked, and some are validated only half heartedly. It is our collective responsibility to demand careful validation before we call something prognostic and even think about using it in clinical practice. Indeed, the factors called prognostic only after reaching point B are much more likely to make a difference.
clinically. Of course the number of factors reaching point B will be much smaller than those reaching A. But this by itself is desirable—we have nothing to lose but our unvalidated false-prognostic factors.

CONCLUSION
There is a great interest in being able to accurately predict the future, and a general sense that the key to this is identification of enough prognostic factors. However, there are many biases inherent in much of the data we have to work with, and there are issues buried in the details of how analyses are done that may result in overly optimistic results. The great pressure to identify prognostic factors and the complexity of the details in appropriately evaluating these is a dangerous combination. This article counters some of the lack of understanding by pointing out many of the more common pitfalls, so that investigators and readers can avoid being led astray. We must be careful in our studies, critical in our evaluation, and have the patience to seek validation in our quest to accurately predict outcomes.

REFERENCES