Classification Techniques in Analyzing Surgical Outcomes Data

Peter J Fabri, MD, PhD, FACS

We live in a world that has been characterized by "big data." Data that have been collected for years and even decades are now seen as important inputs to understanding current processes, predicting outcomes, and prescribing actions. This movement is so prominent that the term *data mining* has been supplanted by *analytics* to describe the broadly based and sophisticated methodologies that have been created and validated to allow these massive databases to be analyzed. Every major university has at least one course entitled "Analytics" and often degree programs to address the needs of business, engineering, psychology, and others.

The NSQIP was initially introduced within the Veterans Affairs system in the mid-1990s as the first structured, risk-adjusted, prospective database of surgical outcomes.[1] A private-sector version was developed and then implemented by the American College of Surgeons (ACS) and has rapidly evolved and matured into a national surgical outcomes database.[2] This database now qualifies as "big data" and is available to investigators for analysis of specific surgical procedures and evaluation of surgical quality in hospitals and institutions.

During the same years that the ACS NSQIP has "matured," there has been considerable enhancement of the analytical techniques available to study large databases.[3-5] These newer techniques were developed to address many of the limitations of the earlier methods and have been made possible because of the rapid expansion of the computing power of modern desktop computers and servers. The purpose of this review is to provide an overview of the analytical techniques currently available to enhance the ability of the reader, either clinician or investigator, to interpret and apply predictive modeling in surgical care. This review is not a treatise on statistics and is not intended to be an in-depth description of how to perform database queries or to use modern statistical software. The interested reader is referred to the focused references and more general statistical texts, as well as to the help functions in modern statistical software, for assistance.

Disclosure Information: Nothing to disclose.

Received October 28, 2013; Accepted November 6, 2013.

From the Departments of Surgery and Industrial Engineering, University of South Florida, Tampa, FL.

Correspondence address: Peter J Fabri, MD, PhD, FACS, Departments of Surgery and Industrial Engineering, University of South Florida, Tampa, FL 33612.

MODELING

Predictive modeling is a broad mathematical and statistical methodology that attempts to associate a set of known input variables with an output by deriving a function or model that presents a "fit."

INPUTS ▶ FUNCTION / MODEL ▶ OUTPUT

The function or process is usually represented as a mathematical equation with variables and coefficients. It is easy to believe that the function/model has physical meaning and that the individual coefficients are "real," but it is, in fact, just a model, and there could actually be other models that are much better but have not been found.

George EP Box, one of the pre-eminent statisticians of the 20th century, is quoted as saying "All models are wrong, but some are useful."[6] This statement is extremely important when considering predictive models. Translated into pragmatic terms, it means that a published predictive model (even the NSQIP model!) is not accurate, but it is the best that we can do at the moment. In addition, this statement places a major burden on the investigator to do everything possible to assure that the model is useful. Box went on to say "Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful."[6] (See p. 74.)

Predictive modeling is a retrospective and inductive process and, as such, it cannot identify cause and effect. It is easy to assume that because the bilirubin level is in a model that bilirubin is actually causative but, in fact, bilirubin might very likely be a surrogate measure for something else not measured, which is the actual cause. For example, a database that included arm length and IQ would show a very strong correlation between them, suggesting that we should all start stretching our arms. But once you "fix" age, the relationship disappears. They are both the consequence of age, with no actual causal relationship.

DEFINITIONS

It is critical to establish a few definitions in advance. These terms were selected because they are often misunderstood.

ISSN 1072-7515/13/$36.00
http://dx.doi.org/10.1016/j.jamcollsurg.2013.11.005

Assumptions: A set of agreements about important relationships that are neither tested nor proven, but if incorrect would likely alter the result.

Bias: A systematic difference between what is measured and what is actual. In a simple x-y plot of data points, with a best-fit line (see Fig. 1), the y-intercept is actually the bias because, in a perfect world, if x = 0, y should also = 0. Predictive models with too few input variables often exhibit large bias.

Collinear and multicollinear: When data points lie along a single line (collinear), knowing the parameters of the line allows predicting the points, so additional points that are collinear do not actually add new information. Similarly, variables that are each collinear but have the same slope (multicollinear) do not add additional information but do increase the complexity of the model.

Dimensionality: The number of input variables that is included in a predictive model. Less complex models (fewer inputs) are subject to bias, and more complex models (more inputs) are subject to increased variance. This has been called the "bias-variance trade-off" and leads to the "curse of dimensionality." Much like Goldilocks, a predictive model should be not too big, not too small, but just right.

Dummy variable: To include qualitative nominal variables in an otherwise quantitative model, the variable must be recoded as a binary integer. Variables with more than 2 options are typically recoded as a set of binary integers.

Independent: This term has a number of meanings, even within statistics, and it is easy to confuse them. When used as in "independent variable," independent is a synonym for input or predictor. It does not imply that the variables are actually independent of each other. When used as in "the variables are independent," it means that there is no correlation or relationship between the variables, and there is no joint probability. This latter definition of independent requires proof and cannot be assumed. The first does not imply the second and vice versa.

Least squares: An analytical method of determining "best fit" by minimizing the sum of the squared difference between the actual data points and the predicted values. It assumes a normal distribution, linearity, additivity, uniformity of variance, and independence of variables. Minor deviations from these assumptions are acceptable, but major deviations can lead to incorrect conclusions.

Methods: A set of computer-based mathematical functions or algorithms that generate a model.

Outlier: A data point that is so extreme that it is either an error or it represents an observation from a different population. Outliers are extremely important in modeling, as they can exert an enormous (and inappropriate) influence on the final model. This creates a dilemma—should we delete the outlier as being spurious or should we focus on the outlier as being important?

Parametric and nonparametric methods: Parametric analytical method calculates parameters (eg, central tendency [mean] and dispersion [standard error]) and then uses only the parameters in subsequent analyses rather than the actual data. This requires that the parameters appropriately represent the data, which requires that the assumed probability distribution is reasonably correct. Nonparametric methods are also known as "distribution-free," as they make no assumptions about the underlying distributions of data, but rather estimate the likelihood of finding the degree of overlap between the data subsets.

Probability and odds probability: In the pure sense, probability is a known, generalizable measure of how often something occurs (eg, flipping a coin). Because true probability is infrequently known in medicine, likelihood is used and is estimated from a preliminary sample or set. Odds is the ratio of the probability of yes to the probability of no

$$\left(\frac{P(x)}{1-P(x)}\right).$$

This is equivalent to a likelihood ratio, and the log of the likelihood ratio is the same as log odds.

Significance: Put simply, how willing am I to be wrong? It does not mean important or useful. A p value <0.05 means I am willing to be wrong, at the most, 1 time in 20. Very often, observations that are statistically significant (unlikely to be wrong) are completely useless, such as a treatment that significantly increases survival by 5 minutes.

Small sample size: Unusual things can happen with small datasets and subsets because of the increased impact of any outliers. Small is often defined as <30 samples. Avoid using a subset with <30 samples.

Standardization: Transforming variables that are of markedly different magnitudes (eg, age and annual income) so that they are more comparable. One way to accomplish this is by subtracting the mean of the variable and dividing by the standard deviation $\left(\frac{x-\mu}{\sigma}\right)$. This creates the equivalent of a z score. Without standardization, such a model would be excessively influenced by the annual income.

Validation: A consequence of the "curse of dimensionality" is that models with too few variables exhibit large bias (not accurate), and models with too many variables