Statistics for the Rest of Us

Clinical-Pathologic Conference: Use and choice of statistical methods for the clinical study, "Superficial adenocarcinoma of the esophagus"

Eugene H. Blackstone, MD^{a,b}
Thomas W. Rice, MD^a

Drs Rice and Blackstone

See related article on page 1077.

From the Department of Thoracic and Cardiovascular Surgery^a and the Department of Biostatistics and Epidemiology,^b The Cleveland Clinic Foundation, Cleveland, Ohio.

Received for publication April 6, 2001; revisions requested May 23, 2001; revisions received Aug 24, 2001; accepted for publication Aug 30, 2001.

Address for reprints: Eugene H. Blackstone, MD, The Cleveland Clinic Foundation, 9500 Euclid Ave, Desk F25, Cleveland, OH 44195 (E-mail: blackse@ccf.org).

J Thorac Cardiovasc Surg 2001;122:1063-76

Copyright © 2001 by The American Association for Thoracic Surgery

0022-5223/2001 \$35.00 + 0 12/1/119858

doi:10.1067/mtc.2001.119858

Recently, the Editor of the *Journal* telephoned us with a "crazy idea." He read a few phrases from the "Patients and Methods" section of our paper "Superficial Adenocarcinoma of the Esophagus" (which appears in this issue¹) and thought most readers would understand the first phrase, perhaps 50% the second, maybe 10% to 25% the third, and but a handful the fourth. His idea was to call a time-out to bring readers up to speed on statistical methodology. He suggested we extract key phrases from our paper and explain them in the format of a Clinical-Pathologic Conference (CPC).

His selection of "Superficial Adenocarcinoma of the Esophagus" is interesting, because the intensity of statistical analysis required to unlock the meaning of the data is high. Further, the article appears in the General Thoracic Surgery section, introducing into that arena data analysis concepts and methods more frequently found in the cardiac surgery sections.

Before proceeding, please read the paper.

Each section of the CPC is introduced by quotations from the paper and followed by dialogue between Drs Rice (TWR) and Blackstone (EHB). Throughout the dialogue, key technical ideas are highlighted for discussion in marginal notes. We recommend two sources of supplemental information: chapter 71 of *Thoracic Surgery*² and chapter 6 of *Cardiac Surgery*.³

Essence of the Article

Surgery is the treatment of choice for superficial adenocarcinoma of the esophagus. The ideal patient has high-grade dysplasia found at surveillance, good pulmonary function, and undergoes a transhiatal esophagectomy. Discovery of N1 disease or development of postoperative pulmonary complications necessitating reintubation reduces the benefits of surgery. (Ultramini-Abstract)

EHB: Dr Rice, for readers unfamiliar with superficial adenocarcinoma of the esophagus, what instigated this study?

TWR: Adenocarcinoma arising in Barrett esophagus is occurring with increasing frequency, resulting in more patients presenting with cancers confined to the mucosa or submucosa—superficial adenocarcinoma. These patients are likely to be cured by operation. Thus, my initial motivation was to provide a gold standard to which experimental alternatives to esophagectomy for early-stage disease could be held.

While analyzing the data and writing the manuscript, we realized that death from cancer was not the primary determinant of outcome. Rather, it was comorbidity, surgical factors, and postoperative mortality and morbidity. This

Margin Notes

Ultramini-Abstract

The *Ultramini-Abstract* was introduced to convey the essence of a study's findings.⁴ It is generally two or three sentences long (50 words maximum for *The Journal of Thoracic and Cardiovascular Surgery*).

The maximum word length of the Ultramini-Abstract resulted from an experiment by the editorial office. A couple dozen manuscripts submitted to the *Journal* were reviewed, and 25-, 50-, 75-, and 100-word summarizing statements were generated and evaluated. Twenty-five words (one sentence) proved too few to capture the essence of most papers. Seventy-five words read like a condensed abstract. On occasion, the essence of a study could not be captured in 50 words. Such manuscripts contained too many ideas (information content overload); they needed to be split into two or more papers.

Although ostensibly intended for readers, an ultramini-abstract helps writers focus on the truest statements they can make from their understanding of a study's information, data, and analyses. It is the best preparation for writing a manuscript.

Prevalence, Incidence, Rate

Prevalence, *incidence*, and *rate* are used interchangeably. Perhaps common usage should prevail, because it rarely leads to confusion. But it is not accurate. We prefer selecting the specific word whose technical definition matches the context.

Prevalence is the frequency of occurrence of some factor, characteristic, event, or incident in a group. Of the three words being considered, it is the least commonly used but the most commonly meant! For example, Table 2 of the paper indicates that between 1985 and 2000 at The Cleveland Clinic, the *prevalence* of high-grade dysplasia among 122 patients undergoing esophagectomy was 38 patients, or 31%.

Incidence is frequency of occurrence *per unit of time*. It is expressed on a scale of inverse time (cases per year, deaths per year), or *rate* of occurrence. The prevalence of high-grade dysplasia in a population is governed by the rate of appearance of new cases (incidence) and the rate of removal of cases by death.

Rate as used in scientific contexts is a quantity per unit time. Speed is a rate: $\text{km} \cdot \text{h}^{-1}$; cardiac output is a blood flow rate: $\text{L} \cdot \text{min}^{-1}$. In the context of events, rate is synonymous with incidence. The hazard function is a rate ($\text{deaths} \cdot \text{year}^{-1}$) and incidence. In the paper, we used hazard functions and so did not want to confuse incidence and prevalence.

How, then, can we rephrase such common expressions as these?

"Incidence of hospital mortality was. . ."

"Hospital mortality rate was. . ."

"Five-year survival rate was. . ."

We could write, "Prevalence of hospital mortality was. . ."

However, in most instances, the words prevalence, incidence, and rate are superfluous. It is better to just write, "Hospital mortality was. . ." or "Five-year survival was. . ."

In other contexts, the word *occurrence* is a suitable substitute for prevalence. For example, "Pneumothorax occurred in. . ." is preferable to "Incidence of pneumothorax was. . ."

changed the focus of the paper and enriched its clinical applicability.

EHB: Good long-term outcome in these patients suggested that the goals of treatment be surgical mortality and morbidity approaching zero. These are goals more typical of coronary artery bypass grafting than cancer palliation.

Crafting a Road Map for the Reader

The purposes of this study were to (1) evaluate the results of surgical management of superficial adenocarcinoma of the esophagus and (2) identify predictors of long-term survival for (a) decision-making (preoperative factors), (b) prognostication (operative factors), and (c) hospital care (postoperative complications). (Introduction)

TWR: The statement of purpose grew out of our iterative work with the data and results. It was not a linear process from hypothesis to inference. Pursuits of many leads were abandoned. Insights gained generated new questions, which in turn dictated new analyses, producing new insights. These are the dynamics of a serious clinical study.

EHB: When we finally understood the meaning of the data from this iterative process, we distilled its essence into an *Ultramini-Abstract*.⁴ From that, we developed a statement of purpose. The "Results" section was organized and the "Patients and Methods" section structured in exact alignment with the statement of purpose. The words match exactly. Thus, the paper has an explicit and consistent road map to guide the reader as it guided the writers.

Defining the Study Group

From our prospective surgical database of 577 patients undergoing resection of esophageal carcinoma at The Cleveland Clinic Foundation beginning January 1983, 122 patients were found to have superficial adenocarcinoma of the esophagus. (Patients and Methods)

TWR: It is crucial to define the study group. This may seem simplistic, but it is not. When I came to The Cleveland Clinic, I started a registry of esophageal surgery that evolved into a prospective database. A registry prevents patients from falling through the cracks and from other biases of ascertainment.

EHB: Characterization of the study group includes context of care (the specific institution), time frame, and population from which the group was drawn.

Moving Target: Trends Across Time

*The number of patients operated on increased across time. . . . The surgical technique evolved from routine thoracotomy to transhiatal esophagectomy with lymph node sampling for those patients with a low risk of lymph node metastases. . . . These models include factors whose **prevalence** changed across*

Figure 1. Relation of an unlimited scale of risk, here expressed in logit units, to the probability of occurrence of an event. The logistic equation mapping logit units to probability is shown. *ln*, Natural logarithm.

time. Strategically, we believe that such models are desirable and more helpful than simply attributing the improvement in results to a so-called learning curve. . . . Because the surgical technique and decision-making changed across time (Appendix I) and simultaneously early mortality improved ($P = .01$), we analyzed the potentially confounding trends across time to identify if possible those changes that improved results. (Patients and Methods)

TWR: During the experience, marked changes occurred in epidemiology, presentation, preoperative evaluation, surgical technique, and postoperative care. Does this evolution negate analysis of the experience? Can you identify changes that were for better or for worse?

EHB: This was both an analytic and philosophic challenge. We faced a moving target. Inferences about the relation of evolutionary changes to patient outcome were not protected by a mechanism such as a randomized clinical trial. If we were “good modelers,” if each change was documented patient by patient, and if we had **sufficient data**, then multivariable analyses could quantify the impact of changes on outcome.^{5,6} That is a lot of “ifs!”

Logistic Regression for Time Trends

Management changes were represented by **dichotomous variables** (yes or no). A useful method for relating a dichotomous **event**, such as a management change, to one or more **explanatory variables** is **logistic regression**.

Logistic regression uses a mathematical equation known as the **logistic equation**.⁷⁻⁹ It is a **sigmoid** (S-shaped) curve like the oxygen dissociation curve (Figure 1) and therefore has intuitive medical relevance when used in risk factor

Sufficient Data

A common misconception is that the larger the study group (called the **sample** because it is a sample of all such patients, past, present, and future), the larger the amount of data available for analysis. However, in studies of outcome events, the effective sample size for analysis is proportional to the **number of events** that has occurred, not the size of the study group. Thus, a study of 200 patients experiencing 10 events has an **effective sample size** of 10, not 200.

Ability to detect differences in outcome is coupled with effective sample size. A statistical quantification of the ability to detect a difference is the **power** of a study. This is a complex subject, so only those few aspects of power that affect multivariable analyses of events will be mentioned.

The rule of thumb in multivariable analysis is that the ratio of events to risk factors identified should be about 10 to 1.^{5,6} However, the guideline is not specific enough. Many variables represent subgroups of patients, some of them few in number (such as 6 patients with T1b N1 disease). If a single patient in a small subgroup dies, multivariable analysis may identify that subgroup as one at high risk when, in fact, the variable represents only this specific patient, not a common denominator of risk. The purpose of a multivariable analysis is to identify general risk factors, not individual patients experiencing events!

Thus, more than 1 event needs to be associated with every variable considered in the analysis. For our group, sufficient data means at least 5 events associated with every variable. However, because variables may be correlated and subgroups overlap (T1b N1 patients are in the larger subgroup of N1 patients as well as the T1b group), in the course of analysis, the number of unexplained events in a subgroup may fall below 5, which is insufficient data.

This strategy could result in identifying up to 1 factor per 5 events. We get nervous at this extreme, but in small studies we are sometimes close to that ratio.

Thus, there is both an upper limit of risk factors that can be identified by multivariable analysis and a lower limit of events to allow a variable to be considered in the analysis. **Sufficient data**, then, implies having enough events available to test for all relevant risk factors.

Dichotomous Variables

Dichotomous variables are the simplest subset of categorical variables. They can take on only two different classes or values, such as yes or no, positive or negative, 0 or 1. A dichotomous outcome may be called **binary data** (eg, hospital death).

Outcomes and Events

Results of therapy are **outcomes**. A subset of outcomes is **events**. Events are expressed in analyses as **dichotomous variables** (see above). Outcomes may be related to **explanatory variables** (see below), such as death, recurrence of cancer, functional status after surgery, or postoperative FEV₁.

An outcome in one setting can be an explanatory variable in another. In the paper, management changes were an event in the context of examining therapy. They were explanatory variables in the context of an analysis of mortality.

Explanatory Variables

The set of variables examined in relation to an outcome is called *explanatory variables*, *independent variables*, *correlates*, *risk factors*, *incremental risk factors*, *covariables*, or *predictors*. These alternative names distinguish this set of variables from outcomes. No statistical properties are implied. The least understood name is *independent variable* (or *independent risk factor*). Some mistakenly believe it means the variable is uncorrelated with any other risk factor. All it actually describes is a variable that by some criterion has been found (1) to be associated with outcome and (2) to contribute information about outcome in addition to that provided by other variables considered simultaneously.

Logistic Equation

The *logistic equation* is $P = 1/[1 + e^{-z}]$, where P is probability, e is approximately 2.7183 and is known as the base for the natural system of logarithms (see below), and z is the logarithmic parameter, specifically, the power to which e is raised.

The logistic equation was devised to characterize population growth.⁷ Berkson and Hollander⁸ noted that it characterized a number of biologic phenomena, including the proportion of erythrocytes lysed as their suspension medium became increasingly hypotonic. Berkson⁹ made it the basis for bioassay.

We can rearrange the logistic equation as follows:

$$\begin{aligned} P + Pe^{-z} &= 1 \\ e^z &= P/(1 - P) \\ z &= \ln(P/[1 - P]) \end{aligned}$$

Thus, the logistic equation relates the *absolute* probability, P , of an event to an approximation of *relative risk* known as the *odds ratio*. The odds ratio is the proportion of patients experiencing an event divided by the proportion of patients not experiencing it ($1 - P$, the so-called *complement* of P): $P/(1 - P)$. To convert the odds ratio to a limitless scale (going from minus infinity to plus infinity), its logarithm is used, z . Dr Berkson called the units of this scale “logit units.”¹⁰

Logistic Regression

In the 1960s, Jerome Cornfield^{12,13} suggested the logarithmic odds ratio (log odds) parameter z of the logistic equation be the carrier of explanatory variables. The mathematical form of z he suggested was “logit linear”:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

where the β 's are regression coefficients and the x 's are risk factors, such as age or FEV₁. The β 's translate the measurement scale of the risk factors (x 's) onto the scale of risk (logit).

An increasing number of risk factors and a larger magnitude of the relation between a unit change in the value of a risk factor and risk “move a patient to the right” on the logit scale. This increments risk commensurate with where the patient started in the logit curve.

Since its introduction, *logistic regression* has become the most common form of multivariable analysis for non-time-related events such as hospital mortality, occurrence of postoperative events, or use of particular management techniques.

analysis. If a risk factor imparts 2 units of risk, a robust patient, far to the left on the graph in Figure 1, would have only a small probability of experiencing an event. In contrast, a fragile patient, near 0 on the graph, would have a large probability of experiencing the same event.^{10,11} In one form or another, all types of event analyses are based on a similar S-shaped relation.

Generally, many variables are examined in *logistic regression*,^{12,13} but for time trends, our attention was confined to the date of operation.

Events Occurring After Time Zero

Postoperative complications were recorded and assessed. (Patients and Methods)

TWR: *Outcome* of cancer surgery usually is dominated by cancer mortality. Because so few cancer deaths occurred in this study, other factors that could influence outcome were recorded and evaluated, including events occurring during postoperative care.

EHB: Events occurring after *time zero* (time of esophagectomy in this study) generally are not analyzed as potential *risk factors*. They are called *time-varying covariables* and are avoided for compelling reasons. First, because these events take place after time zero, some patients die before they occur; this affects the denominator for the analysis. Second, they themselves are outcomes, with their own risk factors that should be identified. Mortality and other complications following an occurrence should be studied. Third, the closer they occur to death, the more apt they are to be a surrogate for death (*confounding*).

We justified examining the influence of events occurring shortly after time zero as a way to gain insight into issues of postoperative management. Sequential analysis (discussed below) prevented our being fooled by confounding.

Formal, Systematic Follow-up

Patients were followed up by periodic clinic visits; however, cross-sectional systematic follow-up was made in January 2000. (Patients and Methods)

TWR: You insisted we attempt to contact all patients we believed were still alive. Why couldn't we depend on clinic notes, simply recording the date patients were last seen?

EHB: Complete, “active,” systematic follow-up of patients is a necessity. “Passive” follow-up through clinic visits or inquiries of patients' physicians is inadequate.

The following hypothetical explanation may help: 100 patients underwent operation on the same day. The goal was to determine their fate 2 years later. Data were assembled from clinic visits. Some patients were last seen 6 months after surgery, others at 10 months, a few at 15 months, and 2 at 2 years. One patient died 30 months after surgery. Imagine the impossibility of obtaining a meaningful answer to the status of these 100 patients at 2 years when the status

of only 3 people was known at that time! This is called numerators in search of denominators.¹⁴

In reality, patients undergo operations over a span of time. One good follow-up strategy is to determine the status of each patient at a fixed interval after surgery (as in the example). This is the *anniversary* method of follow-up.¹⁵ Another good method is to ascertain the status of all patients at a given point in time (called *cross-sectional* follow-up). This is the *common closing date* method, which we employed.¹⁶ Anything short of formal, systematic, complete follow-up by one of these two methods leads to uninterpretable survival estimates.

Descriptive Statistics

Descriptive statistics are summarized as the mean and standard deviation for continuous variables and as frequencies and percentages for categorical variables. (Patients and Methods)

TWR: Surprisingly, as a surgeon trained in technical details, I may miss the essence of the techniques of analysis that are introduced with this sentence. Do not be intimidated by statistics! You need to understand the methods without performing the statistics!

EHB: Nearly every phrase in this sentence has a technical meaning. Each also implies assumptions about the data that the reader is asked to take on faith!

Descriptive statistics means information that characterizes the study group. It allows readers to appreciate the composition of this specific group. There are often important geographic and institution referral differences in clinical studies. To avoid jumping to the conclusion, “That’s not my experience!,” readers should study the descriptive statistics carefully.

Ideally, patients’ data (stripped of informative identifiers) would be made available case by case. This is impractical. Instead, the study group is characterized by summarizing information. Summarizing information is different for different types of *variables*.

Continuous Variables

Some variables, like age, can take on a different value for every patient. This characterizes a *continuous* variable. One way to describe a continuous variable is to list each value. *Cumulative distribution plots* do just that. A description of how a cumulative distributed curve is constructed for age is given in the legend for Figure 2. The legend explains the *median* value (50% older and 50% younger), *percentiles*, and *quartiles*.

A more abstract way to summarize age is to imagine that the values fall into a pattern that can be represented by a bell-shaped *mathematical model*. Figure 3 is a *histogram* showing patient age grouped into 5-year intervals. It looks somewhat bell-shaped, like the smooth bell-shaped curve superimposed on it. The bell-shaped curve was constructed

Time Zero

In time-to-event (survival) analysis, *time zero* is the time at which every patient in the study becomes at risk of experiencing the event being examined. In this study, time zero was esophagectomy.

Fortunately, surgery is an unmistakable event that makes errors of defining time zero uncommon (although they occur in particular settings). In medical studies, time zero is often elusive. For example, we do not know time of onset of adenocarcinoma of the esophagus.

Time-varying Covariables

Time-varying covariables are factors, events, or measurements whose values change after time zero. Typical examples are respiratory failure occurring after operation, cancer recurrence, adjuvant therapy, development of a new medical condition, and change in blood pressure. Their proper analysis requires special mathematics. Their relation to other events, such as death, must be interpreted with care.

Confounding

A *confounder* is a variable related both to outcome and to groups being compared. This presents a challenge, because it is analogous to the researcher being required to answer the question, “Which came first, the chicken or the egg?”

Variable or Parameter?

A *variable* is an item that can take on different values for different patients. A *parameter* is a constant. The two terms are antonyms, yet they are commonly used as synonyms! We recommend their proper technical usage. Thus, age is a variable, but mean age is a parameter.

Mathematical Model

A *mathematical model* is an equation (or set of equations) representing real data. Equations contain symbols representing parameters whose values are estimated from the data (see “Parameter Estimates,” page 1069). Mathematical models may arise from a theory of nature or from empiric observation that they represent data reasonably. They are “compact” because an entire set of data is summarized by values of a small number of parameters in the mathematical model.

Histograms and Cumulative Distributions

A *histogram* is a type of bar graph that summarizes the distribution of values of a continuous variable. Categories of the variable are selected of equal width (eg, 5-year age groups), and the number of patients in each category is displayed on the vertical axis.

In contrast, *cumulative distribution curves* utilize every value, not categories of values, and increment monotonically upward (see Figure 2). The shape of the histogram is roughly the slope of the cumulative distribution function.

Gaussian (Normal) Distribution

The equation of the bell-shaped *Gaussian (normal) distribution* curve is

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where:

π is a constant, approximately 3.1415927 . . . , pi

e is a constant, approximately 2.7183 . . . , the base of the natural logarithms

σ is a parameter that represents the standard duration of the variable

μ is a parameter that represents the mean of the variable

x represents a value of the variable X , generally graphed on the horizontal axis

y represents the probability of occurrence of a particular value of x .

Because in medicine *normal* has several unrelated meanings, we have used the more technical term *Gaussian*.

Standard Deviation Versus Standard Error

Standard deviation is the Gaussian distribution parameter representing the scatter or deviation of individual values from the mean. It is a descriptive statistic.

Standard error is the standard deviation of the mean, an estimate of the *precision* of the mean (*precision* is related to scatter; *accuracy* is related to lack of *bias*—systematic deviation from the true value). Unlike the standard deviation, which is similar in value for large and small samples of data, the standard error decreases as n increases.

Because the Gaussian curve is symmetric around the mean, the two parameters of the Gaussian distribution are expressed by the shorthand *mean $\pm$ SD*, where SD is 1 standard deviation. This means 68% of patient ages fall between (mean – SD) and (mean + SD). This is one instance, not terribly common in statistics, in which the shorthand $\pm$ is used instead of *confidence limits*.

Misleading Means

Data may not be distributed symmetrically on both sides of the mean. Often, they are *skewed* to the right (see below). The typical postoperative stay may be 6 days, but a few patients stay 30, 200, or more days. The presence of a few long stay values inflates the estimate of the mean. This typically results in a standard deviation larger in magnitude than the mean, such as 10 ± 14 days. These parameter estimates imply that 68% of the stays will range from –4 days to +24 days! Yet, length of stay can take on only positive values, so –4 days alerts you to summarizing statistics that make no sense.

Mean and standard deviation are *parameters* of a specific model of data distribution. If the Gaussian model does not represent the data well, it is a bad model, and something else must be done.

One thing that can be done is to transform the data onto a scale that is less susceptible to skewness. For example, the data values might be transformed to logarithmic scale. Logarithms of positive numbers spread small values and bunch large ones. The mean value of logarithms may be more normally distributed and have a sensible standard deviation. The mean, mean –

Figure 2. Cumulative distribution of age at operation for superficial adenocarcinoma of the esophagus. Each patient's age is represented on the curve from youngest to oldest. Each unique age value increments to the curve by $1/n$, where n is the total number of patients. Notice that 50% of the patients were younger than 64.5 years and 50% were older. This is the *median* age, expressed by horizontal and vertical *solid straight lines*. The *coarse dashed lines* enclose the 25th and 75th percentiles (or *quartiles*), meaning that 25% of patients were younger than the 25th percentile and 75% were younger than the 75th percentile. For consistency with the standard deviation, which encloses about 70% of the ages, the 15th and 85th percentiles (15% of ages above and 15% below) are shown by *fine dashed lines*.

from a mathematical equation called the *Gaussian (normal) distribution*. The Gaussian distribution equation contains two constants that characterize its shape. These constants are *parameters*. The parameter representing the location along the horizontal axis of the peak of the curve is called the *mean*. The estimate of the mean is average age.

The other parameter identifies the *point of inflection* of the curve on each side of the mean. This is the point of transition from a steep ascent or descent to the shallower flange of the bell. The location of the inflection point is a parameter called the *standard deviation*. The closer patient ages are grouped together, the closer the standard deviation will be to the mean. Ages between 1 standard deviation below the mean and 1 standard deviation above it encompass 68% of the patients, as can be appreciated in Figure 2.

The mean and standard deviation are easily computed. But the computations are *misleading* if the distribution of values is asymmetric (*skewed*). This situation may be addressed by transformations, such as *logarithms*, or by nonparametric statistics, such as percentiles.

Categorical Variables

In contrast to continuous variables, variables such as sex, depth of tumor invasion (T), and regional lymph node status (N) have values representing one of two or one of a small

Figure 3. Histogram of age and superimposed *Gaussian distribution* curve. Ages of patients have been categorized and counted in 5-year intervals. The *smooth curve* is a parametric summary of the distribution of age, expressed as the *mean* (the *vertical line* at the peak of the curve) and *standard deviation* (enclosed by the *finely dashed vertical lines*). It is questionable whether the mean and standard deviation represent the distribution of these ages well; yet the Gaussian distribution is often used in this setting because of tradition or ease of calculation. On the other hand, notice that the mean on this figure and median on Figure 2 are similar; the ages enclosed within ± 1 standard deviation of this figure are similar to those within the 15th and 85th percentiles of Figure 2.

number of categories—hence the name *categorical* variable. The number of patients in each category is the *frequency*. Because this number varies widely from study to study, it is customary to express frequencies on a uniform scale, namely, the number per 100 patients (*percent*).

Distribution of Times to an Event (Survival Analysis)

Nonparametric estimates of survival were obtained by the method of Kaplan and Meier. The parametric method was used to resolve the number of phases of instantaneous risk of death (hazard function) and to estimate their shaping parameters (Patients and Methods). . . . The instantaneous risk of death was high immediately after the operation, then fell to a constant level of 4.2% per year. (Results)

EHB: *Nonparametric and parametric* are key technical terms. **Cumulative distribution curves** and histograms of age (Figures 2 and 3) require no mathematical models containing parameters. They are called *nonparametric* statistics. In contrast, the bell-shaped curve superimposed on the histogram of age (Figure 3) is an equation with parameters.

The Kaplan–Meier method produces nonparametric estimates of the distribution of times until death.¹⁷ It is analogous to the construction of Figure 2, except that, by convention, the cumulative distribution of times until death is turned upside down.

SD, and mean + SD of logarithms are then raised to the power of the base (called taking the antilogarithm), producing what is called the geometric mean and its asymmetric confidence limits. Another transformation is the inverse of each value, that is, its value divided into 1. If the inverse values are normally distributed, their mean and standard deviation can be found. Then, the mean, mean – SD, and mean + SD are transformed back to the original measurement scale, producing the harmonic mean and its asymmetric confidence limits.

An alternative is to forget about modeling the data altogether. Report the value for which half the patients have a greater number (median), and present various percentiles (eg, 25th and 75th percentiles or 15th and 85th to be consistent with the width of a standard duration) described in Figure 2.

Skewness

Skewness is a statistical measure of the asymmetry of distribution of values for a variable. In medical data, asymmetry is often characterized by a number of atypically large values for a variable. Because the number line proceeds from small numbers on the left to large ones on the right, asymmetry in the data distribution is called *right skewness*. (See “Misleading Means,” above.)

Logarithm

A logarithm is the exponent or power of a fixed number, called the base. When the base is raised to that power (the *antilogarithm*), the untransformed number is regenerated. Typical bases are 10 and the number *e*, whose value is 2.7183 (*e* is called the base of the natural logarithms). For example, the logarithms of the numbers 0.001, 0.01, 0.1, 1, 10, 100, and 1000 to the base 10 are –3, –2, –1, 0, 1, 2, and 3.

Cumulative Distribution Versus Survival Curve

If all patients in a study have died, the distribution of times until death can be depicted by a *cumulative distribution function*, as in Figure 2. We generally are unable to use this simple cumulative distribution method because at follow-up not everybody has died.

For living patients, the time of death is not yet known; nevertheless, we know they have lived a specific length of time after time zero. Thus, we have *incomplete* information about their length of life, not *missing* information. The Kaplan–Meier method (one of many such methods) uses both complete data (dead patients) and incomplete data (living patients) to estimate at least a portion of the distribution of time until death.

Patients with incomplete data (living) are called *censored*. This term comes from the way governments determine population survival from census figures, that is, by counting living people.

Parameter Estimates

Parameters in mathematical models are placeholders for numeric values. When the parameters take on numeric values, the model becomes an *equation* that can be solved, for example, for individual patients’ risks.

Numeric values are called *parameter estimates*. They are *estimates* because they are based on a finite sample of data. Just as a mean value (a parameter estimate) is associated with uncertainty proportional to both the standard duration and effective sample size, so any parameter estimate is associated with uncertainty.

Parameter values are estimated by means of statistical theory and procedures. The estimation process may be complex or as simple as counting and dividing (to estimate a probability).

Hazard Function

The *hazard function* is the instantaneous risk of death or other time-related event.

If the hazard function is steady across time, it is called a *constant hazard* or *linearized rate*. It is easily estimated by dividing the number of events by the total of follow-up time for that event. A constant hazard results in survival decreasing exponentially. This is analogous to exponential radioactive decay driven at a constant rate, called the half-life.

In most medical settings, the hazard function is not constant. The human population hazard function is high at birth, diminishes rapidly, is relatively flat for a few decades, and then rises with advanced age (sometimes called a bathtub-shaped hazard function).

The units of hazard are inverse time. Because it is instantaneous, the magnitude of the hazard function can be huge for a short while, such as immediately after surgery. If the duration of high hazard is brief, few deaths will ensue, however.

Multivariable Versus Multivariate

Multivariable analysis is an analysis of a set of explanatory variables with respect to a *single outcome* variable. *Multivariate* analysis is an analysis of several outcome variables simultaneously with respect to explanatory variables.

Before modern multivariate analysis was possible, the terms most used for a multivariable analysis were “multiple” or “multivariate.” Since the advent of methods to analyze multiple outcomes simultaneously, *multivariable* has come to be associated with simple outcomes analysis in the American literature. European literature groups these together as *multivariate*, perhaps because multivariable analysis is the degenerate form of multivariable analysis when number of outcomes is 1.

Strength of Association

The *strength of association* of a risk factor with outcome is expressed by a type of parameter called a *coefficient*. A coefficient is a multiplier in an algebraic expression. For example, in the expression $0.026 \times \text{age}$, 0.026 is the coefficient and multiplier of age. The coefficient translates units of age into units of age-associated risk.

Most multivariable models consist of an additive relation among risk factors, as shown for logistic regression. That is, each variable in the analysis, such as age, FEV₁, or type of cancer, is weighted by its coefficient (generally, the larger the weight, the stronger the association with outcome). Then, the product pairs of the coefficient and variable are added together with all other pairs to form a risk score.

A *parametric* method using a mathematical model can also characterize the distribution of times until death.¹⁸ Because such distributions are rarely bell-shaped, models more suited to survival data are used. Raw survival data are used to *estimate the parameters* (constants) of these models. The parametric method used in this paper was based on mathematical models of the birth–life–death process.¹⁹ Such models incorporate an expression for the rate of transition from life to death, called the *hazard function*.¹⁸ They are identical to biochemical kinetics models, with reaction rate analogous to hazard function.²⁰

In this study, the instantaneous rate of death was high immediately after surgery, then fell rapidly to a steady value after about 6 months. A steady hazard (*constant hazard*) results in survival decreasing exponentially.

Risk factors can modulate the hazard function. In this study, they raised and lowered the constant hazard rate.

Multivariable Analysis

Value of a Sequential Strategy

The strategy for the multivariable analysis used a sequential approach to variables that reflects the purposes of the study (Methods and Materials). . . . *Decision Model. . . Prognostic Model. . . Hospital Care Model* (Results).

TWR: I needed to know what elements of the data were important during successive phases of patient care. What information is important for decision-making before a planned operation? How is prognosis refined after esophagectomy by pathologic stage? What is the survival impact of unforeseen events occurring early postoperatively?

EHB: Providing information helpful in each phase of clinical care required a *sequential approach* to multivariable analysis. Initially, only preoperative variables and their relation to outcome were examined. Then, pathologic variables were added and superceded information removed (eg, pathologic stage for clinical stage). Finally, postoperative events were added to the analysis.

TWR: This is a “medical” approach to *multivariable data analysis*. It is an advantage to have a colleague who knows the statistical methodology and has participated in patient care.

Concepts

EHB: For many, *multivariable analysis* is a mystery. We know intuitively that a patient’s outcome is related to many variables. We measure or observe and record variables, some of which may be *associated* with outcome, even if they are not directly *causal*. One goal of multivariable analysis is to identify, from among the many recorded variables, those most related to outcome (risk factors).

Risk factor identification is challenging in medicine, because many variables are correlated with one another. For

example, women on average are shorter and have a smaller body surface area than men; sex, height, and body surface are correlated. Risk factors are identified in a context that accounts for correlated information by evaluating all variables *simultaneously*. The **strength of association** with outcome of each variable is adjusted for all other variables in the analysis. Thus, it is correct to think of this strength as the *incremental risk* the variable adds beyond that contributed by all other simultaneously considered variables.

The number of variables that can be in a model simultaneously is limited by the number of events, not total n . (See “Sufficient Data.”) Thus, although we might like to consider all variables at once and then trim down the list (called a *backward* variable selection strategy), in this study, with a limited number of events, we built the model gradually from simple (few variables) to more complex (greater number of variables) using a *forward* variable selection strategy.

When the number of events is small, we recommend developing a parsimonious multivariable model (the simplest model that adequately explains the data).¹⁹ Thus, the analysis is directed toward finding the common denominators of the event.³

Understanding the Variables

Initial screening of variables possibly related to survival used the log-rank test and the Cox proportional hazards model. (Patients and Methods)

TWR: Because many factors influence patient survival, it is necessary to use multivariable analysis. So what good is screening variables one at a time and presenting univariable results?

EHB: We screen individual variables to answer a couple of questions. First, are there sufficient data for analysis? As noted earlier, if there are fewer than about 5 events associated with a subgroup of patients, we cannot use this subgroup for multivariable analysis. Second, is there a **proportional hazards** relationship between a variable and outcome? By proportional hazards, we mean the ratio of hazard when a risk factor is present to that when it is absent is constant across time. This assumption of Cox proportional hazards modeling must be verified if that method of risk factor analysis will be used.²¹

Truthfully, other than “weeding out” variables and testing assumptions, I pay little attention to univariable survival tests. What is important is the multivariable relations. Thus, I do not prescreen to get rid of otherwise perfectly good but not univariably statistically significant variables. There are instances in which the relation of a variable to outcome is hidden in univariable analyses, and not until other factors have been accounted for is it revealed. These are called *lurking* variables.²²

A controversial use of screening is to restrict the number of variables examined in the multivariable analysis.²³ This

Checking the Proportional Hazards Assumption

Whenever *Cox proportional hazards* analysis is performed, the assumption of proportional hazards must be verified. The Cox model is formulated for a single dichotomous variable as follows:

$$\Lambda(t) = \Lambda_0(t)e^{\beta_1 x_1}$$

where $\Lambda(t)$ is the cumulative hazard function, $\Lambda_0(t)$ is the underlying cumulative hazard (not specified explicitly), e is 2.7183. . . , the base of the natural logarithm, β_1 is the Cox regression coefficient, and x_1 is the dichotomous variable.

The ratio of cumulative hazard with the factor present ($x_1 = 1$) to that with it absent ($x_1 = 0$) is

$$\frac{\Lambda(t, x_1 = 1)}{\Lambda(t, x_1 = 0)} = e^{\beta_1}$$

Taking logarithms:

$$\beta_1 = \ln[\Lambda(t, x_1 = 1)] - \ln[\Lambda(t, x_1 = 0)]$$

Notice that the logarithm of the two cumulative hazard curves is separated across all time by β_1 . If separation is not constant, the proportional hazards assumption is violated.

Cumulative hazard is estimated from the survival curve $S(t)$ by taking the logarithm:

$$\Lambda(t) = -\ln[S(t)]$$

can lead to restrictive prespecifying of variables to be examined, which may preclude generation of new knowledge.

Organizing Variables

The potential risk factors (variables) were organized for analysis. . . . (Patients and Methods)

TWR: The key to your analyses is grouping similar risk factors. Is this a more powerful strategy than considering each factor as it appears in an unordered list?

EHB: Organization of well-understood, high-quality variables is key to successful, medically informed modeling of outcomes. To the casual statistical consultant, all variables are equal. Under such circumstances, chances are reduced that the analyses will “turn out right.” In a collaborative effort, those analyzing the data become familiar with each variable, what it means, how its values were gathered, its quality in terms of **accuracy and precision**, and other knowledge and understanding of the variables, patients, and goals of the study. From this intimate knowledge of the variables, we group them into medically meaningful classes.

We consider the *class of variables* as “the” variable and the individual variables within the class as minor differences in specification. To illustrate, we consider “patient size” as the variable, but it may be represented by height, weight, body surface area, or body mass index.

Accuracy Versus Precision

Accuracy is the absence of systematic error of measurement (bias) from the “truth.” *Precision* is the ability to provide the same answer in repeated measurements. These terms are commonly interchanged, but in data analysis they are different. Scales may be inaccurate because of an offset of weight or incorrect calibration. However, they may yield repeatable (precise), inaccurate readings. A measurement may be imprecise because of inability to obtain consistent results, because the scale may be too coarse, or because of interobserver error.

Linearizing Transformations

To linearize the relation between the measurement scale of a continuous or ordinal variable and the scale of risk may require transformation of the measurement scale. Transformations of scale might include inverse, logarithm, power, root (such as square root), and so on. The right transformation produces a scale linearly related to risk.

Other techniques can be used to ensure a linear relationship between risk and measurement scales that, together, we call *calibration*. Calibration is extra work! Busy statisticians may not be given (or take) the time necessary to explore calibration. It is worth the time!

Calibration

Continuous and ordinal variables were assessed univariably by decile risk analysis to suggest transformations of scale to incorporate into the multivariable analyses to ensure that the relation of these variables to outcome was well calibrated with respect to model assumptions. (Patients and Methods)

TWR: Many investigators stratify continuous variables into two (or a few) groups and analyze the resulting categories. I notice that you always analyze continuous variables as such. Is this just a difference in style?

EHB: Continuous variables contain information unique to each patient. Creating categorical variables from continuous variables wastes precious information. Generally, the cut points (points of categorization, such as age > 70 years) are arbitrary. This practice flies in the teeth of a philosophical idea: continuity in nature.¹⁹ A 69.9-year-old is more like a 70.1-year-old than a 59-year-old or an 85-year-old. We nearly always find that continuously valued risk factors follow a smooth gradient of risk that supports the idea of continuity in nature.

There is a scientific argument as well. We are interested in knowing the shape of the relationship of the variable to outcome. You cannot characterize the shape if you *begin* by categorizing continuous variables.

TWR: I remember your asking me, “At what value is FEV₁ [1-second forced expiratory volume] associated with reduced survival?” I said, “About 2 L.” As plotted in the paper’s Figure 5, this is indeed the case. However, this rela-

tionship is not continuous. It is flat to about 2.2 L; then it is associated with decreasing survival.

EHB: This particular shape was suggested by a calibration process that took the form of *linearizing transformations*. Figure 4, A, shows a scale of risk along the vertical axis and FEV₁ on the horizontal axis. The relation of FEV₁ to the scale of risk is not perfectly linear. Figure 4, B, shows a transformed scale FEV₁, and the points now line up straighter. This is what is meant by a linearizing transformation.

When we “unwound” the transformation for FEV₁, it became evident that above about 2.2 L there was little increment in risk and below it, a substantial and steep gradient of risk. Thus, we had discovered the shape of the relationship.

Managing Missing Values for Variables

Informative imputation for missing values of pulmonary function tests used a multiple regression model based on available function tests, age, and sex. (Patients and Methods)

TWR: A number of patients did not have pulmonary function tested preoperatively. If these patients were discarded, their other data would be wasted.

EHB: Most investigations of missing data have been in social science, where it makes sense to discard from analysis individuals who fail to return their survey. Less attention has been given to sporadic missing data, characteristic of clinical studies.

For sporadic missing data, we usually *impute* (substitute) the mean value of patients with nonmissing data. We verify the imputed data are *noninformative* (that is, they do not add information that biases the results of analysis) by forming indicator variables. These identify patients in whom values for a particular variable have been imputed. The indicator variables are incorporated into analyses to test whether patients with missing data behave differently with respect to outcome than patients with available data.

In the case of pulmonary function tests in this study, more than a small amount of data was missing. Therefore, knowing that medical data contain correlated variables, we performed *informative imputation*. Specifically, we substituted a value based on other variables correlated with pulmonary function, rather than the mean for the whole group. To do this, we performed a multivariable analysis of pulmonary function tests from patients with nonmissing measurements. This generated an equation to predict pulmonary function of those patients based on age and sex.²³

Identifying the Risk Factors

Multivariable survival analysis was performed for each hazard phase using a directed technique of entry of variables into the multivariable models. (Patients and Methods)

Figure 4. Calibration of 1-second forced expiratory volume (FEV_1) to risk. A scale of risk is given on the *vertical axis* (akin to the logit units of Figure 1), and 8 groups of equal numbers of patients according to the value for FEV_1 along the *horizontal axis*. Their mortality, converted to the risk scale, is shown at each *closed circle*. (The 8th closed circle cannot be shown because there were no deaths in the 8th group with the highest FEV_1 s.) A, Linear scale of FEV_1 . Clearly, there is a decreasing (more negative) value of risk at higher FEV_1 (simple regression line shown, with explained scatter for these points of 80%). B, Inverse scale of FEV_1 . Because of the inverse transformation, the lower FEV_1 s are to the right of the scale and the higher FEV_1 s to the left. Risk falls from left to right, unlike in Figure 4, A. There is now tighter correspondence of risk to this rescaling of FEV_1 (85% of scatter explained) than the conventional scale of Figure 4, A.

TWR: When you use words like “directed variable selection,” I get nervous. It sounds like multivariable analysis is art, not science.

EHB: My former colleague, Dr David Naftel of the University of Alabama at Birmingham, enumerated the reasons why different investigators might obtain different models using the same data set.²⁴ One source of difference is the approach to model building.

We do a lot of “hand work,” directed by extensive statistics about variables not yet in the model, but adjusted for those that are. I pay particular attention to the cluster of variables in each organized category, entering that variable from each that seems to best represent the category. There is an art to this. It is an art that employs knowledge about both the data and the medical condition.

Part of the hand work is sorting out correlations between variables and possible compensation of one variable for another variable that incompletely or inadequately relates to outcome. For example, if age is inappropriately managed at its extremes, a variable associated with the elderly or the young may be identified as a risk factor; however, this factor is merely an adjustment for inadequately calibrating age.

What Is Magic About $P < .05$?

However, the early hazard phase, determined from the data, was calculated to contain only 5 events; thus, there was limited ability to identify early-phase risk

factors. A $P = .1$ criterion for retention of variables in the final models was used. (Patients and Methods)

TWR: I thought statistically significant meant $P < .05$.

EHB: The requirement of at least 19:1 odds ($P < .05$) to reject the idea that the relationship of a variable to outcome is unlikely to be due to chance is attributed to Sir Ronald Fisher.²⁵ Actually, he selected this value for a specific agricultural experiment, warning the reader that each new situation requires establishing appropriate odds to distinguish a relationship from chance.

P is highly dependent on effective sample size. If there is not much data, it is hard to find risk factors based on P ! To avoid overlooking risk factors in small studies, we may choose $P < .1$ or $P < .2$ for inclusion of variables in the multivariable analysis. This is called avoiding a *type II* statistical error. On the other hand, a spurious variable may be identified as a risk factor by chance. This is a *type I* statistical error. So there is danger of both type I and type II errors that must be balanced.

Bootstrap Bagging—What It Can and Cannot Do

Because of small study size, bootstrap resampling was used to validate the models. . . . Thus, the risk factors were not only identified as statistically significant by traditional analysis, but also occurred the most frequently in bootstrap analysis. The tables of risk factors include frequency of occurrence from multivariable

bootstrap modeling, as well as conventional magnitude and certainty of the association. (Patients and Methods)

TWR: When you introduced me to bootstrapping, my hope was that it would multiply the data, eliminating the limitation of n . That is not how it works, and its role is different.

EHB: Actually, it is the proverbial answer to the maiden's prayer, but a different prayer than you had hoped for! Remember the dilemma that using P value criteria exposes the investigator to the chance of both spurious risk factor detection and failure to detect? Remember your accusation of "art, not science" in variable selection?

Recently, a technique has been introduced that is similar in concept to visual evoked potentials or signal-averaged electrocardiograms.²⁶ The entire analytic process of variable selection is subjected to repeated resampling and reanalysis.

In practice, a patient is drawn at random (using a random number generator) from the original data set. This begins the formation of a new data set. Another patient is drawn at random; it might be the same patient or a different one. This goes on until a new data set is built with either the same number of observations as the original or somewhat fewer. An automated process is then used to select variables. Once a model is obtained, it is stored in the computer. This entire process of selecting patients and performing an analysis is repeated 100 to 1000 times. As the results are averaged, a "signal" gradually emerges.²⁷ Some variables are repeatedly found to be risk factors, others only occasionally. The few that stand out as consistent are reliable risk factors.²⁸

Let me try to put this process into your domain. Imagine a space alien trying to figure out what a thoracic surgeon is. If the alien watches randomly throughout the day, it may find the surgeon asleep, eating, playing baseball with children, examining a patient, or performing an operation in the thorax. After repeated examinations of a randomly selected group of thoracic surgeons, the picture gradually emerges that this is a person who performs operations for diseases of the lungs, esophagus, and chest wall. If the alien is observing differences between thoracic surgeons and people at random, factors like sleeping and eating and playing with children disappear into the background and the professional profile emerges.

Presenting Results

Confidence Limits: Expressing Uncertainty of Inferences

Confidence limits (CL) of proportions are also equivalent to 1 standard error (68% CL) (Patients and Methods). . . . Two patients died in the hospital after the operation and 1 within 30 days, for an operative mortality of 2.5% (CL 1.1%-4.9%). (Results)

TWR: You and Dr John Kirklin introduced confidence limits into our literature in the late 1960s. I have not seen

many papers recently that utilize them as extensively as you suggested.

EHB: Their need and utility are as compelling today as 30 years ago. There were 2 deaths in the hospital and 1 out of the hospital within 30 days in your study. The fact is that mortality was 2.5%. There is nothing uncertain about this. However, confidence limits translate an experience of the past into an estimate of results in future patients. Intuitively, the smaller the experience, the less certainty that results will be similar in the future. In this experience, 2.5% mortality (called the *point estimate*) is consistent with mortality ranging from about 1% to 5%.

I do not know why surgeons have not found this information useful. Even the general public expects pollsters to give them a "margin of error."

Multivariable Results

Tables of risk factors identified in the hazard domain are presented with their regression coefficients rather than hazard ratio, because the model is not one of proportional hazards. (Patients and Methods)

TWR: Some years ago you used bullets to indicate risk factors and possibly a P . Now you use complex tables with multiple footnotes. In addition, I am accustomed to obtaining hazard ratios from our statistician, but you give me regression coefficients. Why?

EHB: A multivariable analysis generates an enormous amount of information about (1) the model's structure and estimates of model structural parameters (if one is using parametric modeling); (2) risk factors identified; (3) magnitude of the association of risk factors with outcome (expressed as coefficients, odds ratios, or hazard ratios); (4) direction of relation (positive, negative); (5) uncertainty of association (standard deviation); (6) score on which P is based; (7) P ; (8) covariance structure (documenting interrelations among variables); and recently, (9) bootstrap reliability. There is no room to print all of this information! Therefore, some triage is nearly always necessary (a complete transcription of a multivariable model is also sometimes needed²⁸); bullet points were one approach to triage.

As to why we do not use *hazard ratios*, the answer is simpler. Hazard ratios are meaningful under assumptions of proportional hazards. When we use transformations of scale and nonproportional hazards modeling, hazard ratios are not readily interpretable.

A Picture Is Worth 1000 Words

. . . because the hazard function multivariable analyses are completely parametric (generate an equation), "nomograms" from the analyses are presented in which specific values are entered into the equations, the equations solved, and the results presented graphically with confidence limits. (Patients and Methods)

TWR: The value of a parametric analysis is that it produces an equation that can be solved for any patient with any risk factor. It is about more than just identifying risk factors.

EHB: The solution, moreover, can be presented graphically in what we call nomograms. This was one of the motivations for our developing a completely parametric hazard function methodology.¹⁸ Thus, I can show you the relationship of survival and FEV₁, or of survival and age, by solving an equation. I can plot a graph of a *patient's specific prognosis* from the equation.²⁹ This information is ideal for understanding disease and its treatment, for making individual patient decisions, and for obtaining informed consent.

Nomograms require only simple high school level algebra. Values for all variables in the model are multiplied by their respective coefficients, the products are summed, the rest of the equation is solved, and a plot is generated.

Internal Verification of Model Adequacy

The accuracy of this model is corroborated by the comparison to actual deaths (Results). . . . Adequacy of the prognostic model (Table 5)

TWR: If a person has pN1 disease, prognosis is grim. Increasing depth of tumor invasion is also related to poorer survival, by univariable analysis. However, depth of tumor invasion (T) is related to the probability of having N1 disease.³⁰ Yet I do not see T in the prognostic model. Why not?

EHB: Patients with greater depth of tumor invasion have poorer survival than those with more superficial disease. However, greater tumor invasion is accompanied by other even more prognostically important factors, such as pN1 disease. After accounting for other factors, depth of tumor invasion contributed too little additional prognostic information to be retained in the multivariable model.

It is possible that small effective sample size precluded detecting an additional increment of risk related to T or that the study was too restrictive in the spectrum of T (confined to superficial carcinomas) to detect a more general trend of increasing risk with increasing depth of invasion. One of the beauties of a completely parametric model is that we can check this out! Using the multivariable model (see "Patient-specific Prediction"), we calculated expected survival for each level of tumor invasion.

As Appendix Figure I (paper) shows, there was good correspondence with Kaplan–Meier survival estimates stratified by T. Even though T was not directly represented in the model, it was adequately accounted for by other variables, such as pN1.

It would be a mistake to conclude that T is not a risk factor. Certainly, the greater the depth of tumor invasion, the worse the survival. However, the poorer prognosis is accounted for by other factors correlated with T.

Patient-specific Prediction

Parametric models permit the calculation of patient-specific survival curves as in Figure 6 (paper).³ These curves can be generated for alternative treatments and compared with that actually given.²⁹

Perhaps unappreciated is that a multivariable analysis reveals differences in survival unsuspected by average survival expressed by Kaplan–Meier curves. In Figure 6 (paper), the low-risk and high-risk patient-specific predictions are quite different. Both differ substantially from the average Kaplan–Meier curve. This is why we should calculate individual survival probabilities on the basis of information we know.

Patient-specific predictions also have a role in interval validation of model accuracy. We use two methods. First, at the actual time of follow-up or death for each patient, we calculate predicted survival. Survival is transformed to cumulative hazard. The sum of cumulative hazards across patients will equal the number of events observed. We then subgroup patients and verify that the number of predicted deaths is similar to the number observed in each subgroup.

Another way to verify a model is to generate a patient-specific survival curve for each patient. The patients are then subgrouped. We verify that the average of these individual curves corresponds to actual subgroup Kaplan–Meier survival estimates.

Interpreting Results: Importance of an External Standard

After accounting for pathologic stage, age at operation became a risk factor. No sharp age cutoff was identified: the older the patient, the shorter the survival. However, patients younger than 55 years had poorer survival than their US population counterparts, whereas patients aged 55 to 75 and those more than 75 years lived about as long as expected. (Results)

TWR: Before you started the analysis, I believed that we should not be operating on older patients. You changed my mind. Certainly, older patients have a more complex hospital course and poorer survival than younger patients, as you show in the multivariable model. You have convinced me that the prognosis of older patients is better and the prognosis of younger patients is actually worse. Explain this.

EHB: The problem with age is that it is a risk factor for mortality for all of us. So I inquired whether the relation of advanced age to survival was different after surgery from that expected in the general population. I used government life tables to construct a survival curve for each patient based on age, sex, and ethnicity. These curves were then averaged within age groups for convenience of comparison.

Although elderly patients had an increased early mortality, overall they fared about as well as predicted for the gen-

eral population. Younger patients had a distinctly worse prognosis than their counterparts in the general population, even though their survival after surgery was better than for older patients.

EHB and TWR: Epilogue

This CPC illustrates important facets of clinical investigation. It shows that collaboration between the clinical investigator and analyzers of the data is crucial. The knowledge of these individuals is not mutually exclusive, but shared. This facilitates a clinically pertinent data analysis and presentation that has clinical inferences for future patient care. It also leads to questions for further investigation. Finally, it maximizes the extraction of useful information from the data. However, this requires application of ever-changing technology in data analysis, statistics, and informatics.

References

1. Rice TW, Blackstone EH, Goldblum JR, DeCamp MM, Murthy SC, Falk GW, et al. Superficial adenocarcinoma of the esophagus. *J Thorac Cardiovasc Surg.* 2001;122:1077-90.
2. Piantadosi S, Kirklin J, Blackstone E. Statistical terminology and definitions. In: Pearson FG, Deslauriers J, Ginsberg RJ, Hiebert CA, McKneally MF, Urschel HC Jr, editors. *Thoracic surgery*. New York: Churchill Livingstone; 1995. p. 1649-77.
3. Kirklin JW, Barratt-Boyes BG. Generation of knowledge from information, data and analyses. In: *Cardiac surgery*, Vol 1, 2nd ed. New York: Churchill Livingstone; 1993. p. 249-82.
4. Kirklin JW, Blackstone EH. Notes from the Editors: Ultramini-abstracts and abstracts. *J Thorac Cardiovasc Surg.* 1994;107:326.
5. Harrell FE Jr, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. *Stat Med.* 1984;3:143-52.
6. Marshall G, Grover FL, Henderson WG, Hammermeister KE. Assessment of predictive models for binary outcomes: an empirical approach using operative death from cardiac surgery. *Stat Med.* 1994;13:1501-11.
7. Verhurst PF. Notice sur la loi que la population suit dans son accroissement. *Math Physique.* 1838;10:113-21.
8. Berkson J, Hollander F. On the equation for the reaction between invertase and sucrose. *J Wash Acad Sci.* 1930;20:157-72.
9. Berkson J. Application of the logistic function to bio-assay. *J Am Stat Assoc.* 1944;39:357-65.
10. Berkson J. Why I prefer logits to probits. *Biometrics.* 1951;7:327-39.
11. Kirklin JW. A letter to Helen (presidential address). *J Thorac Cardiovasc Surg.* 1979;78:643-54.
12. Cornfield J. Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure: a discriminant function analysis. *Fed Proc.* 1962;21:58-61.
13. Gordon T. Statistics in a prospective study: the Framingham study. In: Gail MH, Johnson NL, coordinators. *Proceedings of the American Statistical Association: Sesquicentennial Invited Paper Sessions*. Alexandria (VA): American Statistical Association; 1989. p. 719-26.
14. Spodick DH. Numerators without denominators: there is no FDA for the surgeon. *JAMA.* 1975;232:35-8.
15. Drolette M. The effect of incomplete follow-up. *Biometrics.* 1975;31:135-44.
16. Elveback L. Estimation of survivorship in chronic disease: the "actuarial" method. *J Am Stat Assoc.* 1958;53:420-40.
17. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc.* 1958;53:457-81.
18. Blackstone EH, Naftel DC, Turner ME Jr. The decomposition of time-varying hazard into phases, each incorporating a separate stream of concomitant information. *J Am Stat Assoc.* 1986;81:615-24.
19. Blackstone EH. Black death, smallpox, and continuity in nature: philosophies in generating new knowledge from clinical experiences. *Thorac Cardiovasc Surg.* 1999;47:279-87.
20. Blackstone EH. Outcome analysis using hazard function methodology. *Ann Thorac Surg.* 1996;61:S2-7.
21. Cox DR, Oakes D. *Analysis of survival data*. London: Chapman and Hall; 1984.
22. Joiner B. Lurking variables: some examples. *Am Stat.* 1981;35:227-33.
23. Harrell FE Jr, Lee KL, Mark DB. Tutorial in biostatistics: multivariable prognostic models—issues in developing models, evaluating assumption and adequacy, and measuring and reducing errors. *Stat Med.* 1996;15:361-87.
24. Naftel DC. Do different investigators sometimes produce different multivariable equations from the same data? *J Thorac Cardiovasc Surg.* 1994;107:1528-9.
25. Fisher RA. *Statistical methods and scientific inference*. Edinburgh: Oliver and Boyd; 1956. p. 42.
26. Breiman L. Bagging predictors. *Machine Learning.* 1996;26:123-40.
27. Blackstone EH. Breaking down barriers: helpful breakthrough statistical methods you need to understand better. *J Thorac Cardiovasc Surg.* 2001;122:430-9.
28. Piehler JM, Blackstone EH, Bailey KR, Sullivan ME, Pluth JR, Weiss NS, et al. Reoperation on prosthetic heart valves. *J Thorac Cardiovasc Surg.* 1995;109:30-48.
29. Lytle BW, Blackstone EH, Loop FD, Houghtaling PL, Arnold JH, Akhrass R, et al. Two internal thoracic artery grafts are better than one. *J Thorac Cardiovasc Surg.* 1999;117:855-72.
30. Rice TW, Zuccaro G Jr, Adelstein DJ, Rybicki LA, Blackstone EH, Goldblum JR. Esophageal carcinoma: depth of tumor invasion is predictive of regional lymph node status. *Ann Thorac Surg.* 1998;65:787-92.