Statistical Challenges in Identifying Risk Factors for Aortic Disease

John A. Rizzo, PhD[1,2]*, Jie Chen, PhD[3], Hai Fang, PhD[4], Bulat A. Ziganshin, MD[1,5], John A. Elefteriades, MD[1]

[1]Aortic Institute at Yale-New Haven Hospital, Yale University School of Medicine, New Haven, Connecticut, USA; [2]Department of Economics and Department of Preventive Medicine, Stony Brook University, Stony Brook, New York, USA; [3]Department of Health Services Administration, University of Maryland, College Park, Maryland, USA; [4]China Center for Health Development Studies, Peking University, Beijing, China; and [5]Department of Surgical Diseases #2, Kazan State Medical University, Kazan, Russia

Abstract

Being largely asymptomatic, thoracic aortic aneurysms pose a challenge for the physician to identify and intervene in time to prevent death or a major complication. Knowing how to accurately analyze the available clinical data is vital to informing the proper management of these patients. This paper seeks to provide an overview of the statistical methods most commonly used to analyze clinical outcomes with a special focus on research related to aortic disease. Copyright © 2014 Science International Corp.

Introduction

The latest data from the Centers for Disease Control and Prevention show that aortic aneurysms in various locations are the 18th leading cause of death in the United States. Moreover, in individuals older than 65 years, aortic aneurysms are the 15th most common cause of death [1]. These numbers are striking, because it appears that aortic aneurysms cause more deaths than the human immunodeficiency virus. However, many argue that even these numbers are a significant underestimate of the true impact that aortic disease has on public health, because most cases of sudden cardiac death are considered to be coronary-related, while in reality many might be aneurysm-related. Thoracic aortic aneurysm is a silent disease, which in the strong majority of cases does not produce any symptoms [2]. The aorta grows slowly for many years until it reaches a critical point, at which it either dissects or ruptures—two complications that are bound to produce death unless treated immediately.

Being largely asymptomatic, thoracic aortic aneurysms pose a challenge for the physician to identify and intervene in time to prevent death or a major complication. Therefore, timely detection of patients at risk of developing a thoracic aneurysm is critically important. Such timely detection can be achieved by identifying risk factors, clinical conditions, and biomarkers that have been associated with thoracic aortic disease [2]. For example, in recent years such conditions as bicuspid aortic valve, [3] intracranial aneurysm, [4] and bovine aortic arch [5], as well as a strong family history of aortic disease [6] have all been shown to be associated with thoracic aortic aneurysm and dissection. Nevertheless, a large percentage of newly identified thoracic aortic aneurysms are incidental findings revealed during imaging studies (echocardiography, computed tomography, MRI) performed for unrelated reasons. Once a patient has been diagnosed with a thoracic aneurysm, it is vitally

*Corresponding author:
John A. Rizzo, PhD
Stony Brook University
403 Myrtle Ave
Port Jefferson, NY 11777
Tel: +1 631 741 8539, Fax: +1 631 828 6517, E-Mail: rizzologic@gmail.com

important to closely monitor the progression of the aneurysm until a critical size is attained, at which time surgical treatment is considered appropriate. The estimated growth rate of thoracic aneurysms is approximately 0.1 to 0.15 cm/yr [2]. If the aorta is rapidly increasing in size, that is an indication for early surgical intervention.

Current and future success in combating the "Silent Killer" (i.e. thoracic aortic aneurysm) is largely dependent on high-quality clinical research that is being performed at centers with large numbers of patients with thoracic aortic disease. However, it must be emphasized that appropriate data collection and analysis pose a serious challenge. Clinical data that are collected retrospectively and/or prospectively typically have a nonexperimental design and also bear other common imperfections, presenting challenges in estimating and identifying risk factors, early and late mortality and morbidity, as well as long-term survival and other outcomes for this disease. For this reason, knowing how to accurately analyze the available clinical data is vital for clinical researchers. Fortunately, there are many statistical techniques and tools that are currently available to a clinical scientist to aid in data analysis. It is just a matter of knowing which statistical method is appropriate for analyzing a specific outcome, and how to use that method.

Therefore, this paper seeks to provide an overview of the statistical methods most commonly used to analyze clinical outcomes with a special focus on research related to aortic disease. In this paper we will offer recommendations for dealing with complex statistical issues and will also provide references for further reading on each of the covered topics. We hope this paper will be "one-stop shopping" for many cardiovascular investigators working in the field of aortic disease.

Clinical Outcomes and Risk Factors

Clinical outcomes are the variables we wish to analyze and predict. These may include such measures as mortality, aneurysm size, and/or survival. Clinical outcomes may be binary as in the case of mortality, or continuous as in the case of aneurysm size. They may be static, i.e., measured at a point in time; or longitudinal, i.e., measured over time. Survival is an example

of a clinical outcome that is longitudinal because survival is measured over time.

Predictor variables (also called independent variables, explanatory variables, control variables, or risk factors) can be used to predict the value of outcome variables. They may be static or longitudinal as well. An example of a static risk factor is aneurysm size, while a longitudinal one is change in aneurysm size. Other examples of risk factors for clinical outcomes of aortic disease are age, gender, hypertension, smoking, high cholesterol, diabetes, or family history of aortic disease.

Predictor variables can be continuous or binary/categorical variables. Continuous variables can include an infinite number of different values within a given range. Variables such as blood pressure, height, weight, aortic aneurysm size, and growth rate are usually measured in the continuous form. Binary variables, sometimes called dummy variables, are used to sort data into mutually exclusive categories. Binary variables assume the value 0 or 1 to indicate the absence or presence of some qualitative effect. For example, we can construct a binary variable of FEMALE, which only takes the value of 0 or 1. The binary variable FEMALE $= 1$ if the patient's sex is female, and FEMALE $= 0$ if the sex is male.

Mathematically, one may express the relationship between an outcome variable Y and a number of predictor variables represented by the vector $\textbf{\textit{X}}$ as a functional relationship:

$$Y = f(\textbf{\textit{X}}) \tag{1}$$

The above equation specifies the relationship between Y and $\textbf{\textit{X}}$. The value of Y depends on the Xs, the predictor variables. Y is expected to change if the Xs take different values. The precise relationship between Y and $\textbf{\textit{X}}$ can be linear, nonlinear, or some other functional form.

Linear functional forms.   The most common relationship between a single predictor variable X and Y can be specified as

$$Y = \alpha + \beta X, \tag{2}$$

which indicates a *linear* relationship between Y and X. Mathematically, the estimated relationship between X and Y is thus $\Delta Y/\Delta X = \beta$, where β is a constant number.

Quadratic functional forms. Quadratic functions are used to capture decreasing or increasing marginal effects of a predictor variable X.

$$Y = \alpha + \beta_1 X + \beta_2 X^2. \qquad (3)$$

To illustrate, suppose that Y = TAA size, and X = AGE. The above model specification shows that the relationship between X and Y depends on the value of X. The estimated relationship is approximately the following: $\Delta Y/\Delta X = \beta_1 + 2\beta_2 * X$. If $\beta_1 > 0$ and $\beta_2 > 0$, i.e., the estimated values of β_1 and β_2 are positive, the quadratic functional form shows that TAA size *increases* with age, and the size increases *faster* among the elderly population. But if $\beta_1 > 0$ and $\beta_2 < 0$, TAA size would increase *more slowly* with age.

Step functional forms. The step function is a function that increases or decreases abruptly from one level to another. For example, let Y represent the annual risk of aortic dissection or rupture and X denote aneurysm size. The relationship between Y and X may be best described as a step function, where

$$Y = 0.5\% \text{ if } X < 4 \text{ cm.} \qquad (4)$$

$$2\% \text{ if } 4 \leq X < 6 \text{ cm}$$

$$10\% \text{ if } X \geq 6 \text{ cm.}$$

The step function in this example says that the annual risk of dissection or rupture is 0.5% if the aneurysm is less than 4 cm, but jumps to 2% for aneurysms between 4 cm and 6 cm, and to 10% for aneurysms 6 cm or higher.

Statistical Testing

Once outcome and predictor variables have been defined and the functional relationship between them considered, statistical tests of association must be applied to quantify the assumed relationships.

Bivariate Approaches

When investigating the effects of a single explanatory variable on an outcome, the analysis is said to be *bivariate.* Which statistical approach to employ depends on the nature of the dependent variable as well as the predictor variable. In particular, the appropriate test will differ depending on whether these variables are continuous, like aortic aneurysm size, or categori-

Table 1. Appropriate Statistical Approaches Depending on Nature of Outcome and Predictor Variables

Predictor Variable	Outcome Variable	
	Categorical	Continuous
Categorical	Chi-square or Logistic regression	*t*-test or Linear regression
Continuous	Logistic regression	Linear regression

The list of statistical tests here is not exhaustive. Other tests are available. However, we wish to focus our discussion on the four commonly used statistical approaches included in this table.

cal, like mortality. There are four cases to consider, summarized in Table 1.

Case #1: Outcome variable continuous, predictor variable categorical. When the dependent variable is a continuous measure like aortic aneurysm size and the predictor variable is categorical, the t-test or simple linear regression (ordinary least squares) is appropriate. The t-test determines whether the means of two groups are *statistically* different, and is appropriate whenever you want to compare the means of two groups. For example, one could employ a t-test to examine whether mean aneurysm size differed between patients with and without Marfan syndrome. Simple linear regression also determines whether differences in the outcome variable are statistically different employing a t-test.

Case #2: Outcome variable categorical, predictor variable categorical. When both the outcome and predictor variable are categorical, the chi-square (χ^2) test statistic and/or logistic regression are appropriate. For example, if one wanted to investigate whether having a bovine aortic arch anomaly (a categorical variable) was related to having a bicuspid aortic valve, one could employ a chi-square test or estimate a logistic regression model.

Case #3: Outcome variable continuous, predictor variable continuous. In this case, simple linear regression (ordinary least squares) may be used. Linear regression would be appropriate, for example, if one wanted to examine the relationship between a patient's age and aortic aneurysm size.

Case #4: Outcome variable categorical, predictor variable continuous. Logistic regression analysis may be used to examine the relationship between a continuous predictor variable and a categorical outcome. For example, if one wished to study

the relationship between aortic aneurysm size and risk of dissection or rupture, one could employ logistic regression analysis.

Case-Control Studies

Case-control studies are commonly employed in clinical research in order to infer treatment effects. These studies identify a treatment group (the cases) and a comparator group (the controls). The difference in outcomes between the cases and the controls provides an estimate of the treatment effect. When no attempt is made to adjust for differences between the cases and controls other than the treatment, the case-control study is a bivariate analysis. But cases and controls are often selected to match on a variety of other variables, such as age, gender, and race, in order to remove differences between the two groups along these other dimensions. In these instances, case-control studies have some similarity to multivariable analyses, in that they attempt to adjust for confounders.

Case-control studies may also include pre and post periods. In the pre period, both cases and controls do not receive any treatments. In the post periods, only the cases receive the treatments, but the controls do not. Even if cases and controls are dissimilar, the differences-in-differences estimation approach will net out these factors. The so-called differences-in-differences method estimates treatment effects as follows:

$$\text{Treatment effect} = (\text{Cases\_Post} - \text{Cases\_Pre}) - (\text{Controls\_Post} - \text{Cases\_Pre}). \quad (5)$$

This approach implicitly nets out time-invariant factors that differ between treatments and controls [7,8]. The differences-in-differences approach may also be used in multivariable regression.

Multivariable Regression

While bivariate analyses can often shed light on associations between variables, they fail to establish causal relationships [9]. This is because simple associations between two variables fail to account for other, confounding factors. To take an extreme example, suppose that one collected data on two variables, whether people carried matches and whether these same people had heart disease. According to Table 1 above, a chi-square test could be used and it would very likely show a positive and statistically significant

relationship. That is, we would find that people who carried matches were significantly more likely to have heart disease. But what does this mean? It cannot mean that carrying matches *causes* heart disease. The explanation is that people who carry matches are much more likely to be smokers, and it is smoking that leads to heart disease. Hence, in this example, the simple bivariate model is inadequate and a more complex, multivariable model is required, that includes smoking status.

The need for multivariable modeling arises when data are nonexperimental. In such cases, one cannot rely on randomization to net out the influence of other confounding variables [9]. The idea behind multivariable modeling is to include confounders in your model to remove their influence, thus approximating an experiment. This reduces *bias* (i.e., inaccurate estimates of the true relationship between the predictor variable(s) and the outcome) and leads to more reliable estimates of the effects of the predictor variable(s) of interest. There are many kinds of multivariable models, but two that we will discuss here figure quite prominently in clinical outcomes research. These are logistic regression analysis and linear regression analysis.

Logistic regression.   Logistic regression is used when multiple variables are included to predict a binary outcome, such as mortality or the occurrence of an adverse reaction [10]. The logistic regression model is designed to predict the likelihood of the outcome of interest. Predictor explanatory variables are selected in order to provide a model that predicts the outcome of interest the most accurately.

Logistic regression relates the natural logarithm of the odds ratio to a linear combination of the predictor variables. The odds ratio is defined as the probability of the outcome, π, divided by 1 minus the probability. To understand the logistic regression model requires some familiarity with probability, the odds ratio, and the natural logarithm of the odds ratio. Begin with the concept of probability. Suppose the ten year probability of an aortic dissection is 0.6, i.e., 60%. Then, the probability of not having an aortic dissection in ten years = 1 − 0.6 = 0.4 or 40%. The *odds* of a dissection are defined as the ratio of the probability of having a dissection divided by the probability of not having a dissection. In this example, the odds of an aortic dissection are 0.6/0.4 = 1.5. In other words, the odds are 1.5 to 1. If the probability of a dissection is 0.5, i.e.,

50-50% chance, then the odds of dissection is 1 to 1. The transformation from probability to odds is a *monotonic transformation*, meaning the odds increase as the probability increases or vice versa. Probability ranges from 0 to 1, while odds range from 0 to positive infinity.

But in logistic regression analysis, we take the natural logarithm of the odds ratio as our dependent variable and relate this to a linear combination of explanatory variables. Thus, logistic regression estimates the following relationship:

$$\ln(\pi/1 - \pi) = \alpha + \beta X + u, \qquad (6)$$

where α is a constant term, $\boldsymbol{\beta}$ a vector of parameters to be estimated, X a vector of risk factors, and u is the error term for those risk factors that we are not able to control for. The error term is assumed to follow a logistic distribution, which is similar to the bell-shaped normal distribution, but with heavier tails at either extreme. Because all the risk factors are assumed to be non-random, the dependent variable in equation (6) also follows a logistic distribution and the logistic regression model can be applied to equation (6) [9]. Why use the natural logarithm of the odds ratio instead of the odds ratio itself? One main reason is that the natural logarithm of the odds ratio varies from plus or minus infinity while the odds ratio is truncated at 0. It is more convenient to deal with a model that does not have this truncation issue.

Because of the way the logistic regression model is estimated, it does not directly provide the predicted odds ratio or probability of the event. To obtain the odds ratio (OR), one must exponentiate both sides of equation (6): $OR = e^{(\alpha + \boldsymbol{\beta X})}$, where e is Euler's number, equal to approximately 2.718. That is, one must take Euler's number and raise it to the power equal to the estimated equation, $\alpha + \boldsymbol{\beta X}$, in order to obtain the OR value. The predicted probability of the event, Prob, is given by the formula:

$$Prob = e^{(\alpha + \beta X)}/(1 + e^{(\alpha + \beta X)}). \qquad (7)$$

Because the exponential function is always positive, the predicted probability will always lie between 0 and 1. Standard statistical packages like SAS [11], Stata [12], and SPSS [13] may be employed to estimate logistic regression models. The effects of individual predictor variables are usually expressed in terms of odds ratios, but these effects may be transformed to provide marginal effects on the probability of the outcome, or dependent variable. The logistic regression model is similar to the probit regression model, in which the error term follows a normal rather than a logistic distribution (recall from above that the logistic distribution is like the normal but with heavier tails at either extreme). The advantage of the logistic regression is that it has a simpler analytical form and reports odds ratios that can be used to show the effects of risk factors.

Linear regression. Linear regression is used when the outcome variable of interest is continuous rather than binary [9]. For instance, if one wished to identify factors affecting aortic aneurysm size, linear regression would be used rather than logistic regression because aneurysm size is a continuous measure. Linear regression may be used to model one or more explanatory variables. When one explanatory variable is used, it is referred to as simple linear regression, and when more than one variable is used, as multiple linear regression.

As its name suggests, linear regression posits that the dependent or outcome variable is related to the explanatory variables in a linear fashion. This is the simplest, most commonly employed regression model when the outcome variable is continuous. A linear regression model may be written as:

$$Y = \alpha + \beta X + u \qquad (8)$$

where, as in the logistic regression, α is a constant term, $\boldsymbol{\beta}$ is a vector of parameters to be estimated, X is a vector of predictor variables and u is the error term for all the risk factors that we are not able to control for. Assuming there are i observations, the estimated regression line obtains a predicted value for each observation. The parameters α and $\boldsymbol{\beta}$ are estimated so as to obtain the regression line that is the "closest fit" to the dependent variable Y. This line will be the closest in the sense that the sum of the squared deviations between each value for Y and its predicted value from the regression line are minimized.

Survival analysis. In survival analysis, the outcome variables are duration measures, meaning what is assessed is the time until the event of interest occurs. Examples include time from diagnosis to surgery and the length of survival after aortic valve replacement. The duration, or the survival time, may be measured in days, weeks, months, or years. Survival

times are usually positive and censored. Censoring occurs when the process is ongoing and the information about the survival time is incomplete. Survival functions and hazard functions are key concepts in survival analysis. To illustrate these functions, let $T \geq 0$ denote the duration, and t a particular value of T. The *survival function* is defined as $S(t) = P(T > t)$; that is, the probability of surviving beyond time t. The *hazard function, $\lambda(t)$*, is defined as the instantaneous probability of failure at a particular point in time, t, given that the respondent has survived up to time t.

In survival analysis, we wish to know whether and to what extent patients' demographic characteristics, such as age, gender, or treatment factors, can affect survival times. Several methods are available to analyze the relationship between predictor variables and survival time [14]. *Parametric* methods assume that the underlying distribution of the survival times follows certain probability distributions, such as exponential or Weibull. If the distribution of the survival time T is exponential, the hazard rate becomes constant $\lambda(t) = \lambda$. In other words, the probability of failure in the next time period does not depend on how much time has been spent in the initial state, i.e., the process that drives survival time is independent of the duration. By contrast, if T has a Weibull distribution, hazard rates depend on the duration t, i.e., the hazard may increase or decrease with t. Another popular regression model for the analysis of survival data is the Cox proportional hazards regression model [15]. The Cox regression model has been widely used in survival data analysis, and allows one to adjust for covariates of interest. Cox regression analysis assumes the hazard ratio comparing two observations is constant over time.

The Kaplan-Meier method is a *nonparametric* estimator of the survival function [16]. A nonparametric estimator is one in which a prior functional form or relationship, such as a Weibull distribution, is not assumed in advance, but is determined by information derived from the data itself. This method has been widely used to estimate survival probabilities as a function of time. It estimates the probability of surviving beyond a given time t. The estimate is the product of a series of conditional probabilities:

$$S(t) = P(T > t) = p_1 * p_2 * p_3 * ... * p_t \qquad (9)$$

where p_1 indicates the proportion of patients surviving at least one year, p_2 indicates the proportion of

patients surviving the second year given that they have survived the first, and so on. This method may be used to test the differences between estimated survival rates among two or more groups of respondents, such as treated versus control groups, males versus females, etc. The log-rank test can be used to compare two survival curves and to determine whether the differences in survival between two groups or treatments are statistically significant [17,18]. The null hypothesis is that there is no difference between the survival curves. The test statistic for the log-rank test is chi-squared distributed. If the p-value of the test statistic is less than 0.05, then the two survival curves differ significantly at 95% confidence level.

Challenges in Statistical Modeling

The goal of statistical modeling is to provide insight into factors affecting an outcome of interest. Given the imperfect nature of clinical data, however, there are numerous potential problems that must be considered and addressed in order to ensure that one obtains reliable estimates. This section considers a number of the most commonly occurring issues.

Measurement error.   As its name suggests, the term measurement error refers to variables being measured inaccurately [19]. This may occur with the outcome as well as the predictor variables. Some medical record information may be recorded inaccurately. If patient survey data are used, patient recall may be imperfect. The results of imaging studies may be interpreted inaccurately. Measurement error thus introduces an element of random noise, which may bias or otherwise obfuscate the true relationships between the predictor variables and the outcome of interest.

The effect of measurement error differs for predictor and outcome variables. If the outcome variable is measured with error, there will be no bias in the estimated effect of the predictor variable on the outcome. A noisy outcome variable means there is a loss of precision. As a result, it will be more difficult to demonstrate a statistically significant relationship. Thus, while the estimated effect of the predictor variable will be unchanged with measurement error, the significance level will be reduced. Because measurement error in the outcome variable is random, it will be uncorrelated with the predictor variables. And

since it is uncorrelated with these variables, such measurement error can have no effect on the estimated relationship between the predictor variables and the outcome. We say that these estimates will be unbiased. So measurement error in the outcome variable results only in a loss in precision. And because a larger sample size improves the precision of one's estimates, collecting more data can help overcome the loss of precision associated with measurement error in the outcome variable.

When measurement error occurs in the predictor variable, matters are more serious because the estimated effect will be biased toward zero. That is, measurement error will cause one to estimate a smaller (in absolute value) relationship between the predictor variable and the outcome. For example, if patients self-report their antihypertensive medication usage quite inaccurately, one might estimate that there is no relationship between taking antihypertensive drugs and aneurysm growth rates. But this simply reflects that the predictor variable really isn't capturing antihypertensive drug use, since it is measured so inaccurately.

The best fix to measurement error is prevention; that is, ensuring that all variables are measured as accurately as possible. In reality, however, some measurement error is always present. One correction is to eliminate extreme values, on the theory that a disproportionate share of values very far from the mean (say three or more standard deviations) are likely to be erroneous and often regarded as outliers [20]. In this case, it is important to eliminate extreme values in a *symmetric* fashion. That is, if you eliminate values that are more than three standard deviations above the mean, you should also eliminate those that are more than three standard deviations below. Failure to do so will lead to bias. Consider for example, aneurysm growth rates. If some values are negative, the researcher may be tempted to remove them as that is clear evidence of measurement error (e.g., aneurysms do not naturally shrink). But just because those errors are more readily apparent does not mean that there are no errors at the high end of the distribution. And failure to trim extreme values at the high end as well as the low end of the distribution will lead to upward bias in estimated aneurysm growth rates in this example.

Because the problem with measurement error in the outcome variable is a lack of precision, increasing the sample size, if possible, will improve precision and may resolve the problem. Another approach here is the use of *instrumental variables* [19]. That is, finding a variable—called an instrument—that is measured with little error and that is correlated with the outcome variable of interest. The outcome variable is then regressed on this instrumental variable—only variation in the outcome that is not noise will correlate with this instrument. This approach has been used fruitfully in studies that have estimated aortic aneurysm growth rates. Aortic aneurysm growth rates are measured with some error due to the use of different imaging modalities, interobserver variation, and technical limitations in imaging studies [21,22]. But the *time interval between imaging studies* (the instrumental variable in this example) is measured quite accurately and should be well correlated with *actual changes in aneurysm size*. That is, patients whose imaging studies were taken at longer intervals should demonstrate greater true aneurysm growth.

Overfitting.   Overfitting occurs when too many explanatory variables are included in a model [23]. Overfitting is a problem because it can lead to unstable and imprecise estimates. On the one hand, the researcher wishes to include as many clinically relevant predictor variables as possible, both to identify all relevant factors and to avoid the problem of omitted variables bias to be discussed below. But the number of explanatory variables that may be included is limited by the sample size. To see this, suppose that one has a sample of 500 patients who underwent elective replacement of their thoracic aortic aneurysms and that 10 of these patients died. The researcher wishes to estimate a model that includes 10 risk factors for mortality in elective TAA repair. But with this database, including all of these variables in a logistic regression model would clearly lead to overfitting. With just 10 deaths out of 500 patients, there is simply too little variation to parse out effects to as many as 10 variables. In fact, in logistic regression, there is a rule of thumb known as the "10 to 1 rule" that provides some guidance as to how many variables may be included. This rule says that the maximum number of variables that may be included is equal to the number of observations on the less frequent outcome in the binary variable, divided by 10. So if you have 100 outcomes, 40 of which are death and 60 survival, a logistic regression of factors predicting mortality should include at most 4 (i.e., 40/10) explanatory variables. For linear regression models, one can generally include more explanatory variables for a given sample size because

the dependent variable has more variation, as it is a continuous rather than a binary measure.

Omitted variables bias. Omitted variables bias arises when an important explanatory variable that is correlated with both the dependent variable and the explanatory variable of interest is omitted [9]. Recall the example of carrying matches and heart disease. If a logistic regression model were estimated relating carrying matches to the likelihood of developing heart disease, one would almost surely obtain a positive estimate. That is, carrying matches is positively associated with the probability of heart disease. But surely this is not a meaningful relationship, as matches cannot *cause* heart disease. The problem is that an important variable has been omitted from the model; namely, smoking. Smokers are more likely to carry matches and more likely to have heart disease. Thus, smoking behavior should be added to the model. And when this is done, the resulting model would likely show that it is smoking behavior that leads to heart disease and that carrying matches now has no meaningful relationship to heart disease.

The thing to remember when building a model to predict cardiac clinical outcomes is that one should strive to be thorough by including as many clinically relevant predictor variables as possible to avoid the possibility of omitted variables bias as illustrated above. Sometimes we are unable to avoid omitted variables, as when those variables are not observable. Proxy variables may sometimes prove useful when there is an omitted variables problem. Suppose, for example that we wish to include systolic blood pressure as a risk factor, but this measure is unavailable. But suppose we do know whether the subject is hypertensive or not. We could then construct a binary variable measuring whether the subject is hypertensive as a proxy measure for the omitted variable. When we are unsure whether one variable should be included into the regression or omitted because of its relevance to the outcome variable that we are studying, it is often reasonable to include it into the regression. An irrelevant variable will have little effect on the estimated coefficients on the other risk factors. However, one must remain sensitive to the problem of overfitting, which occurs when too many explanatory variables are included, as discussed above.

Multicollinearity. As we add variables to make the model more complete, we may encounter a problem known as multicollinearity. Multicollinearity arises when two or more explanatory variables are highly correlated with one another [9]. When this happens, the explanatory variables will be estimated imprecisely. We may find, for example, that two variables are statistically insignificant. Yet when we drop either one of them from the model and reestimate, the remaining variable becomes highly significant.

How can we detect multicollinearity and what can we do about it? As a first step, one should look at simple correlations among explanatory variables. If the correlations are very high— say 80% or so—there is a good chance that multicollinearity will be present. Standard statistical packages like SAS [11], Stata [12], and SPSS [13] include formal tests for multicollinearity such as the Variance Inflation Factor (VIF) [24].

If multicollinearity is present, there are several options. First, we may simply drop one of the highly correlated explanatory variables. To return to the matches and heart disease example, suppose we found that carrying matches and smoking were extremely highly correlated so that, when we included both variables in a regression to predict heart disease, neither one was statistically significant. In this case, the solution is clear. One should simply drop the "carrying matches" variable because there is no clinical or other rationale as to why carrying matches should cause heart disease. It simply does not belong in the model on conceptual grounds and should be excluded.

But what if the situation is less clear? Suppose we find that hypertension (HTN) and hypercholestemia are both significant predictors of coronary artery disease (CAD) and that they are both highly correlated with one another. Since each could cause CAD, what are we to do? In this case we might consider collapsing both variables into one index variable. For example, we could define a variable that is equal to 2 if a patient has both HTN and hypercholestemia, equal to 1 if the patient has one of these conditions, and equal to 0 if the patient has neither condition. This avoids the problem of multicollinearity without dropping one of the variables.

A third possibility is to obtain more data. In general, the larger your database, the more highly correlated explanatory variables may be without multicollinearity becoming a problem. The reason is that more data means more variation in the dependent variable. And regression models fit this variation to each of the explanatory variables. So the more variation there is to be explained by the independent variables, the more

precisely they can be estimated, mitigating the problem of multicollinearity.

Reverse causation. Regression models presume that the risk factors or explanatory variables are predetermined in the sense that they affect the outcome variable but not the reverse. This must be true if one hopes to establish a causal relationship between the predictor variables and the outcomes. If this is not true, then it will be impossible to infer any causality.

Reverse causation occurs when the predictor variable is also affected by the outcome. To illustrate the problem of reverse causation, consider two variables that may be used to predict aortic aneurysm size. The first variable is age. Age may be regarded as a predetermined or exogenous variable here because, while it may affect aneurysm size, aneurysm size cannot affect age. The second variable is a binary variable indicating whether the subject exercises strenuously using weights. Now strenuous exercise may cause aneurysms to grow. But causation goes in the other direction as well, because people with larger aneurysms may refrain from strenuous exercise. So in this example, the researcher hypothesizes that strenuous exercise increases aneurysm size. But because of the negative reverse causation, this relationship will be underestimated, i.e., be less positive than the true effect of strenuous exercise. Indeed, if the negative reverse causation is strong enough, one may even estimate a negative association between strenuous exercise and aneurysm size. Because exercise cannot shrink aneurysms, this would be clear evidence of a reverse causation problem.

While statistical methods to correct for reverse causation are well known, [25,26] they have been rarely employed in the medical literature [27]. The data requirements, both in terms of sample size and breadth of variables needed to implement these techniques, are often lacking in clinical databases. And when these techniques are implemented, one can never be sure if reverse causation has been adequately corrected. The best approach is to avoid reverse causation altogether by excluding such variables. When this is not possible, the researcher should be aware of the potential for bias and inability to infer causality as discussed above.

Selection effects. Selection effects arise when the sample being studied is not representative of the population of interest [28]. This may occur for a variety

Table 2. To Optimize Clinical Studies from a Statistical Standpoint

- Measure accurately (both predictor and outcome variables)
- Make "n" as large as possible (analyze more charts, defer analysis until more patients recruited)
- Choose predictor variables carefully based on clinical judgment
- Do not overfit variables into regression analysis (so, use a parsimonious model)
- Avoid using predictor variables that are too highly correlated with each other (thus, avoiding multicollinearity)
- Be certain that predictor variables are not affected by outcome variables (reverse causation)

A link to a statistical program illustrating the issues discussed here using practice databases will be available to researchers in a subsequent issue in this Journal.

of reasons. For example, one may wish to study the natural history of disease progression such as aortic aneurysm growth. But patients with larger, rapidly growing aneurysms are differentially selected out for surgical correction. The resulting sample available to the researcher will include disproportionately high numbers of observations from patients with smaller, more stable aneurysms. As a result of this selection effect, estimated aneurysm growth rates will be smaller than the true natural history of disease progression. As with reverse causation, statistical corrections for selection effects are well known, but they are quite data intensive [28]. At a minimum, however, the researcher should be aware of the potential for selection effects and have some intuition about their implications for interpreting results. In the aneurysm growth rate example just discussed, selection effects mean that the true natural history of the disease was not estimated. Instead, aneurysm growth *given the availability of surgical correction* was estimated. This is still useful information, but it is different from the natural history.

Summary

Statistical analysis of clinical data is challenging because such data inevitably have limitations. Unlike clinical trial data, treatment and control groups may differ, data capture may be less complete, and variables are likely to be measured with less accuracy in many cases. On the other hand, such data

provide evidence of real-world treatments and out-comes and may allow one to obtain larger sample sizes. Moreover, ethical constraints often preclude certain potentially statistically convenient experimental designs. For example, one would not randomly assign patients to surgical aortic aneurysm repair versus medical management when the former was clearly indicated.

As discussed above, there are numerous statistical issues to be considered in working with clinical data. But when given proper attention, most if not all of these concerns can be adequately addressed. Proper use of clinical data for research purposes should begin with measuring risk factors and outcomes as accurately as possible and by obtaining as large a sample as feasible. And predictor variables should be chosen carefully and be included based on strong clinical or other theoretical considerations. Models should be parsimonious, in order to avoid overfitting. These variables should not be too highly correlated with one another, or else problems of multicollinearity will occur. And predictor variables should be predetermined and not a function of the outcome variable, so that one may draw causal inferences. Table 2 provides a summary checklist of issues to consider when preparing and analyzing a clinical database.

Applying these statistical checks is as much an art as it is a science. There is no absolute cutoff for what constitutes correlation among variables that that is "too high," for example. That will depend on the specific database, the size of the sample—and the researcher's own best judgment. As with clinical practice, experience and familiarity with statistical modeling improves the reliability and quality of one's results. And there are some excellent, user-friendly texts on statistical modeling tailored to the clinician for further reading on these topics [29–32]. We hope that the issues discussed in this article provide some assistance to clinical investigators as they work through these modeling issues in their own research.

Conflict of Interest

The authors have no conflict of interest relevant to this publication.

Comment on this Article or Ask a Question

References

1. National Center for Injury Prevention and Control. WISQARS leading causes of death reports, 1999-2007. 2013; http://webappa.cdc.gov/sasweb/ncipc/leadcaus10.html. Accessed November 15, 2013.
2. Elefteriades JA, Farkas EA. Thoracic aortic aneurysm clinically pertinent controversies and uncertainties. J Am Coll Cardiol. 2010; 55:841–857. 10.1016/j.jacc.2009.08.084
3. Davies RR, Kaple RK, Mandapati D, Gallo A, Botta DM Jr, Elefteriades JA, et al. Natural history of ascending aortic aneurysms in the setting of an unreplaced bicuspid aortic valve. Ann Thorac Surg. 2007;83:1338–1344. 10.1016/j.athoracsur.2006.10.074
4. Kuzmik GA, Feldman M, Tranquilli M, Rizzo JA, Johnson M, Elefteriades JA. Concurrent intracranial and thoracic aortic aneurysms. J Am Coll Cardiol. 2010;105:417–420. 10.1016/j.amjcard.2009.09.049
5. Hornick M, Moomiaie R, Mojibian H, Ziganshin B, Almuwaqqat Z, Lee ES, et al. 'Bovine' aortic arch - a marker for thoracic aortic disease. Cardiology. 2012;123:116–124. 10.1159/000342071
6. Albornoz G, Coady MA, Roberts M, Davies RR, Tranquilli M, Rizzo JA, et al. Familial thoracic aortic aneurysms and dissections–incidence, modes of inheritance, and phenotypic patterns. Ann Thorac Surg. 2006;82:1400–1405. 10.1016/j.athoracsur.2006.04.098
7. Angrist JD, Pischke JS. Mostly harmless econometrics: an empiricist's companion. Princeton, NJ: Princeton University Press. 2008.
8. Bertrand M, Duflo E, Mullainathan S. How much should we trust differences-in-differences estimates? Q J Econ. 2004;119:249–275. 10.1162/003355304772839588
9. Wooldridge JM. Introductory econometrics: a modern approach, 5th ed. Stamford, CT: Cengage Learning. 2012.
10. Gujarati DN, Porter DC. Basic econometrics, 5th ed. New York: McGraw Hill Higher Education. 2009.
11. SAS homepage: http://www.sas.com/en_us/home.html.
12. Stata homepage: www.stata.com.
13. SPSS homepage: http://www-01.ibm.com/software/analytics/spss/.
14. Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data, 2nd ed. New York: John Wiley & Sons. 2002.
15. Cox DR. Regression models and life-tables. J R Stat Soc B. 1972;34:187–220.
16. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc. 1958;53:457–481. 10.1080/01621459.1958.10501452
17. Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother Rep. 1966; 50:163–170.
18. Klein JP, Moeschberger ML. Survival analysis. Techniques for censored and truncated data, 2nd ed. New York: Springer Publishers. 2003.
19. Wooldridge JM. Econometric analysis of cross section and panel data. Cambridge, MA: MIT Press. 2010.
20. Rousseeuw P, Leroy A. Robust Regression and Outlier Detection, 3rd ed. Hoboken, NJ: John Wiley & Sons. 1996.
21. Rizzo JA, Coady MA, Elefteriades JA. Procedures for estimating growth rates in thoracic aortic aneurysms. J of Clin Epidemiol. 1998;51: 747–754. 10.1016/S0895-4356(98)00050-X
22. Coady MA, Rizzo JA, Hammond GL, Kopf GS, Elefteriades JA. Surgical intervention criteria for thoracic aortic aneurysms: A study of growth rates and complications. Ann Thorac Surg. 1999;67:1922–1926. 10.1016/S0003-4975(99)00431-2
23. Hawkins D. The problem of overfitting. J Chem Inf Model. 2004;44:1–12. 10.1021/ci0342472

24. Kennedy P. A guide to econometrics, 6th ed. Hoboken, NJ: Wiley-Blackwell. 2008.

25. Stock J, Trebbi F. Retrospectives: who invented instrumental variable regression? J Econ Perspect. 2003;17:177–194. 10.1257/089533003769204416

26. Heckman J. Instrumental variables: a study of implicit behavioral assumptions used in making program evaluations. J Hum Resour. 1997;32:441–462. 10.2307/146178

27. McClellan M, McNeil BJ, Newhouse JP. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. JAMA. 1994;272:859–866. 10.1001/jama.272.11.859

28. Heckman J. Sample selection bias as a specification error. Econometrica. 1979;47:153–161. 10.2307/1912352

29. Katz MH. Multivariable analysis: a practical guide for clinicians and public health researchers, 3rd ed. Cambridge, UK: Cambridge University Press. 2011.

30. Riffenburgh R. Statistics in medicine, 3rd ed. Amsterdam: Elsevier. 2012.

31. Kirkwood BA, Sterne JA. Essential medical statistics, 2nd ed. Oxford: Blackwell Science Ltd. 2003.

32. Motulsky H. Intuitive biostatistics: a non-mathematical guide to statistical thinking, 2nd ed. New York: Oxford University Press. 2010.