Generating New Knowledge in Cardiac Interventions

Eugene H. Blackstone, MD

KEYWORDS
- Data
- Analysis
- Clinical research
- Cardiac
- Cardiac interventions

KEY POINTS
- Understand how information generated from observations made during patient care is transformed into data suitable for analysis.
- Appreciate at a high level what constitutes appropriate analyses of those data.
- Effectively evaluate inferences drawn from those analyses.
- Apply new knowledge to better care for individual patients.

INTRODUCTION

Cardiac interventions are among the most quantitatively studied therapies in medicine.¹ These studies reveal a complex, multifactorial, and multidimensional interplay among patient characteristics, variability of the heart disease, effect of the disease on the patient, conduct of the intervention, and response of the patient to it. The introduction of medical report cards made it evident that multiple factors influencing results of therapy must be taken into account to make fair comparisons of outcomes. Thus, it is important for all involved with cardiac interventions to (1) understand how information generated from observations made during patient care is transformed into data suitable for analysis, (2) appreciate at a high level what constitutes appropriate analyses of those data, (3) effectively evaluate inferences drawn from those analyses, and (4) apply new knowledge to better care for individual patients.

This article should be read by (1) cardiac anesthesiologists, to improve their comprehension of the medical literature and to hone their skills in its critical appraisal; (2) trainees and junior faculty interested in becoming clinical investigators, who need instruction on how to pursue successful clinical research, (3) mature physician-investigators and their collaborating statisticians, mathematicians, and computer scientists, who will benefit from some of the philosophic ideas; and (4) data managers of larger clinical research groups, who need to fully appreciate their pivotal role in successful research: the knowledge-generating team (Box 1).

Large portions of this article are contained in section I of chapter 6 of Kirklin/Barratt-Boyes, Cardiac Surgery, 4th edition.

Clinical Investigations, Heart and Vascular Institute, Cleveland Clinic, 9500 Euclid Avenue, JJ40, Cleveland, OH 44195, USA
E-mail address: blackse@ccf.org

Anesthesiology Clin 31 (2013) 217–248
http://dx.doi.org/10.1016/j.anclin.2012.12.006
1932-2275/13/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
# Knowledge-generating team

## Structure
Regardless of whether the same individuals are involved, clinical research generally includes 2 fundamentally different activities: (1) continuous registry and database activity and (2) individual clinical studies activity. The registry activity involves gathering and entering data for a prescribed set of core data elements for every case. The individual clinical studies activity can be categorized roughly into 2 classes that require different skill sets: (1) clinical trials (either intramurally funded or extramurally sponsored by government or industry) and (2) studies of clinical experience (cohort studies).

## Roles

### Clinician-Investigator
The clinical investigator (with collaboration of key individuals in data management, statistics, and study coordination) must develop the clinical question (aims, objectives), define the study group of interest, identify variables and end points (outcomes) of interest, review the literature, and develop all elements of a study protocol. They must adjudicate data quality, gather values for variables in addition to the core data elements, help interpret the analyses performed, position findings in a clinical context, present the findings to colleagues, and write articles.

### Data Manager
There is no more key support person than the data manager. They are at the interface between data gathering and data analysis. Assembly of data for meaningful analysis is often complex, requiring information to be retrieved from a variety of electronic sources. Data managers usually need formal training in computer science and specifically in database construction and management. They must master an effective data query language. Their most valuable skill, perhaps inborn rather than developed, is attention to the smallest detail of the data. Physicians are usually not of the temperament for this kind of work, and statisticians by training are big-picture–oriented; if physicians see the forest, data managers must see the trees. Thus, data management is not simply skill in formulating databases, writing and executing query logic, and documenting these in detail (although these are important); rather, it is skill in examining the data, finding errors in them, finding inconsistencies and deviation from the norm that should be verified, verifying what appear as outliers, and assessing quality of data for every variable.

The physician-investigator and data manager organize the variables for analysis. If time-related or longitudinal outcomes are being assessed, the data manager must become expert in forming intervals and assessing time-related data (time zero, events, intervals), 2 of the most demanding and essential tasks for such analyses.

For larger clinical quality and research organizations, a statistical programmer must convert data from database format into analysis data sets that make sense to the statistician.

### Data Gatherers
Persons skilled in data gathering for data entry fall into a hierarchy. For gathering some variables, expert medical knowledge is essential. Other data elements can be extracted by individuals with little formal training other than in medical terminology. Essential ingredients are accuracy and integrity. Accuracy may be inborn and is indispensable; it can be assessed prospectively by testing and maintained by quality management and education.

### Education/Quality
If large quantities of data are maintained, 1 or more individuals must assess the quality of the data and from these findings educate the data gatherers. Such individuals must have expert medical knowledge. In large organizations, this role includes maintaining clinical documentation of the database, keeping current with procedural trends, and pruning variables that no longer are of value or are of questionable quality.
The potential obstacle for all is language. For the anesthesiologist, the languages of statistics, mathematics, data management, and computer programming pose a daunting obstacle of symbols, numbers, and algorithms. For collaborating statisticians, mathematicians, and computer scientists, the Greek and Latin language of medicine is equally daunting. This article attempts to surmount the language barrier by translating ideas, philosophy, and unfamiliar concepts into words.

Because this article is intended for a mixed audience, it focuses on the most common points of intersection between cardiac intervention and quantitative science, with the goal of establishing sufficient common ground for effective and efficient collaboration. It is not a substitute for statistical texts or academic courses, nor a substitute for the physician-investigator to establish a collaborative relationship with biostatisticians, nor is it intended to equip anesthesiologists with sufficient statistical expertise to conduct highly sophisticated data analyses themselves. It serves as a brief overview to the more detailed discussion contained in the first section of chapter 6 of the Kirklin/Barratt-Boyes text, Cardiac Surgery.²

The organizational basis for this article is the Newtonian inductive method of discovery.³ It begins with information about a microcosm of medicine, proceeds to translation of information into data and analysis of those data, and ends with new knowledge about a small aspect of nature. This organizational basis emphasizes the phrase, “Let the data speak for themselves.”⁴
**Information**

In health care, information is a collection of material, documentation of work flow, and recorded observations. Information may be recorded in paper-based medical records or in electronic (computer) format.

**Data**

Data consist of organized values for variables, usually expressed symbolically (eg, numerically) by means of a controlled vocabulary.\(^5\) Characterization of data includes descriptive statistics that summarize parts or all of the data and express their variability from patient to patient.

**Analysis**

Analysis is a process, often prolonged and repeated (iterative), that uses a large repertoire of methods by which data are explored, important findings are revealed and unimportant ones suppressed, and relations are clarified and quantified.

**Knowledge**

Knowledge is the synthesis of information, data, and analyses arrived at by inductive reasoning. However, generation of new knowledge does not occur in a vacuum; an important step is assimilating new knowledge within the body of existing knowledge.

New knowledge may take the form of clinical inferences, which are simple summarizing statements that synthesize information, data, and analyses, drawn with varying degrees of confidence that they are true. It may take the form of speculations, which are statements suggested by the data or by reasoning, often about mechanisms, without direct supportive data. Ideally, it takes the form of new hypotheses, which are testable statements suggested by reasoning or inferences from the information, data, and analyses.

New knowledge can be applied to many processes in health care, including (1) generating new concepts, (2) making individual patient care decisions, (3) obtaining informed consent from patients, (4) improving outcomes of interventions, (5) assessing the quality and appropriateness of care, and (6) making regulatory decisions.

**DRIVING FORCES OF NEW KNOWLEDGE**

Many forces drive the generation of new knowledge in cardiac interventions, including the economics of health care, need for innovation, clinical research, procedure success and failure, and awareness of medical error.

**Economics**

The economics of health care are driving changes in practice toward what is hoped to be less expensive, more efficient, yet higher quality care. Interesting methods for testing the validity of these claims have become available in the form of cluster randomized trials.\(^6,7\) In such trials, patients are not randomized, but physicians are. This situation leads to inefficient studies, which nevertheless can be effective with proper design and a large enough pool of physicians.\(^7,8\) It is a study design in which the unit of randomization (physician) is not the unit of analysis (individual patient outcome).\(^9\)

**Innovation**

Innovation is often at odds with cost reduction and is perceived as being at odds with traditional research. However, in all areas of science, injection of innovation is the...
enthalpy that prevents entropy, stimulating yet more research and development and more innovation. Without it, cardiac surgery and interventional cardiology would be unable to adapt to changes in managing ischemic heart disease, potential reversal of the atherosclerotic process, percutaneous approaches to valvar and congenital heart disease, and other changes directed toward less invasive, safer, more effective, and more appropriate therapy.

What is controversial is (1) when and if it is appropriate to subject innovation to formal clinical trial and (2) the ethics of innovation in interventions, for which standardization is difficult.\textsuperscript{10–13}

Reducing the Unknown

New knowledge in cardiac interventions has been driven from its inception by a genuine quest to fill voids of the unknown, whether by clinical research or laboratory research. Clinical research has historically followed 1 of 2 broad designs: randomized clinical trials and nonrandomized studies of cohorts of patients (clinical practice). However, increasing emphasis is being placed on translational research, that is, bringing basic research findings to the bedside. John Kirklin, a pioneer of heart surgery, called this the “excitement at the interface of disciplines.” Part and parcel of the concept of the incremental risk factor is that it is an essential link in a feedback loop that starts with procedure failure, and proceeds to identifying risk factors, drawing inferences about specific gaps in knowledge that need to be addressed by basic science, generating fundamental knowledge, and bringing these full circle to the clinical arena through testing and assessing the value of the new knowledge generated for improving medical care.\textsuperscript{14}

Intervention Success and Failure

Results of intervention in heart disease, particularly surgical failure, have driven much of the new knowledge generated by clinical research. In the late 1970s and early 1980s, a useful concept arose about surgical failures applicable to any failure of an intervention. That is, in the absence of natural disaster or sabotage, there are 2 principal causes of failure of interventions to provide a desired outcome for an individual patient: (1) lack of scientific progress and (2) human error.

The usefulness of this concept is that it leads to the programmatic strategies of research on the one hand and development on the other. Thus, lack of scientific progress is gradually reduced by generating new knowledge (research), and human error is reduced in frequency and consequences by putting available knowledge into practice (development), a process as vital in medicine as it is in the transportation and manufacturing sectors.\textsuperscript{15,16}

Error

Increased awareness of medical error drives the generation of new knowledge just as it drives increasing regulatory pressure and medicolegal litigation.\textsuperscript{17} Human error also places it into the context of cognitive sciences, human factors, and safety research.\textsuperscript{18,19} This interface of disciplines is essential for facilitating substantial reduction in injury from medical errors.

Philosophy

Clinical research as emphasized in this article consists largely of patient-oriented investigations motivated by a serious quest for new knowledge to improve clinical results, that is, to increase early and long-term survival; reduce complications;
enhance quality of life; extend appropriate interventions to more patients, such as high-risk subsets; and devise and evaluate new beneficial procedures.

This inferential activity, aimed at improving clinical results, is in contrast to pure description of experiences. Its motivation also contrasts with those aspects of outcomes assessment motivated by regulation or punishment, institutional promotion or protection, quality assessment by outlier identification, and negative aspects of cost justification or containment. These coexisting motivations stimulated Kirklin and me to identify, articulate, and contrast philosophies that underlie serious clinical research. It is these philosophies that should inform the approach to analysis of clinical experiences for generating new knowledge.

**Deduction Versus Induction**

*Let the data speak for themselves.*

Arguably, Sir Isaac Newton’s greatest contribution to science was a novel intellectual tool: a method for investigating the nature of natural phenomena. His method had 2 strictly ordered aspects, which for the first time were systematically expressed: a first, and extensive, phase of data analysis, whereby observations of some small portion of a natural phenomenon are examined and dissected, followed by a second, less emphasized, phase of synthesis, whereby possible causes are inferred and a small portion of nature revealed from the observations and analyses. This was the beginning of the inductive method in science: valuing first and foremost the observations made about a phenomenon, then letting the data speak for themselves in suggesting possible natural mechanisms.

This represented the antithesis of the deductive method of investigation, which had been so successful in the development of mathematics and logic. The deductive method begins with what is believed to be the nature of the universe (referred to by Newton as “hypothesis”), from which logical predictions are deduced and tested against observations. If the observations deviate from logic, the data are suspect, not the principles behind the deductions. The data do not speak for themselves.

Newton realized that it was impossible at any time or place to have complete knowledge of the universe. Therefore, a new methodology was necessary to examine portions of nature, with less emphasis on synthesizing the whole. The idea was heralded as liberating in nearly all fields of science.

**Determinism Versus Empiricism**

Determinism is the philosophy that everything (events, acts, diseases, decisions) is an inevitable consequence of causal antecedents: “Whatever will be will be.” If disease and patients’ response to disease and to disease treatment were clearly deterministic and inferences deductive, there would be no need to analyze clinical data to discover their general patterns. Great strides are being made in linking causal mechanisms to predictable clinical response. Yet, many areas of cardiovascular medicine remain nondeterministic and incompletely understood. In particular, the relation between a specific patient’s response to complex therapy such as a cardiac operation and known mechanisms of disease seems to be predictable only in a probabilistic sense. For these patients, therapy is based on empiric recognition of general patterns of disease progression and observed response to therapy.

Generating new knowledge from clinical experiences consists, then, of inductive inference about the nature of disease and its treatment from analyses of ongoing, empiric observations of clinical experience that take into account variability, uncertainty, and relationships among surrogate variables for causal mechanisms.
Collectivism Versus Individualism

Are you a lumper or splitter: woods or trees? When generating new knowledge about the nature of heart disease and its treatment, it is important both to examine groups of patients (the woods) and to investigate individual therapeutic failures (the trees). Both views give valuable insights into nature. Statistical methods emphasizing probabilities and general inferences tend to apply to the former, and those emphasizing optimum discrimination for identifying individual patients at risk tend to apply to the latter.

Continuity Versus Discontinuity in Nature

When we turn our focus from a specific patient experiencing an intervention failure to groups of patients, data analysis becomes mandatory to discover relationships between outcome and items that differ in value from patient to patient (called variables). A challenge immediately arises: many of the variables related to outcome are measured either on an ordered clinical scale (ordinal variables), such as New York Heart Association (NYHA) functional class, or on a more or less unlimited scale (continuous variables), such as age. Perhaps as an adaptive function of our brains, humans have a tendency to dichotomize: normal versus abnormal, for example. What the investigator must embrace is a key concept in the history of ideas: continuity in nature. The idea has emerged in mathematics, science, philosophy, history, and theology. The common practice of stratifying age and other more or less continuous variables into 2 or just a few discrete categories is lamentable, because stratifying loses the power of continuity (some statisticians call this “borrowing power”). Focus on small, presumed homogenous, groups of patients also loses the power inherent in a wide spectrum of heterogeneous, but related, cases: any trend observed over an ever-narrower framework looks more and more like no trend at all. Modern methods of machine learning that use classification methods may seem to stumble at this point, but repetition of analyses over thousands of sampled data sets combined with averaging achieves a close approximation to continuity in nature.

Single Versus Multiple Dimensionality

Univariable (1 variable at a time) statistics are attractive, because they are simple to understand. However, most clinical problems are multifactorial. At the same time, clinical data contain enormous redundancies that need to be taken into account (eg, height, weight, body surface area, and body mass index are highly correlated and relate to the conceptual variable “body size”). Multivariable analysis permits multiple factors to be examined simultaneously, takes into account redundancy of information among variables (covariance), and identifies a parsimonious set of variables, which, in cardiovascular disease, have been called risk factors. These are not cause-effect relations, but associations with underlying causal mechanisms. The relationships that are found may be spurious, fortuitous, hard to interpret, and even confusing because of the degree of correlation among variables. For example, women may be at a higher risk of mortality after certain cardiac procedures, but female gender may not be a risk factor, because other factors, such as body mass index, may be the more general variable related to risk, whether in women or men. Even so, it is simultaneously true that (1) being female is not per se a risk factor, but (2) women are at higher risk because on average they are smaller than men. This means that a close collaboration must exist between statistical experts and investigators, particularly in organizing variables for analysis.
Linearity Versus Nonlinearity

Risk factor methodology introduces another complexity besides increased dimensionality. The probability space is bounded by 0 at the bottom and 1 at the top, a floor and ceiling constraint. An important mathematical relationship between a continuous scale of risk and the probability of an adverse event occurring is the logistic equation.\textsuperscript{29,30} It is a symmetric S-shaped curve, which expresses the relationship between a scale of risk, called logit units, and a corresponding scale of absolute probability of experiencing an event (Fig. 1). This nonlinear relationship makes medical sense. Imagine a risk factor with a logit unit coefficient of 1.0. If all other things position a patient far to the left on the logit scale, a 1-logit-unit increase in risk results in a trivial increase in the probability of experiencing an event. As other factors move a patient closer to the center of the scale (0 logit units, corresponding to a 50% probability of an event), a 1-logit-unit increase in risk makes a huge difference. This finding is consistent with the medical perception that some patients experiencing the same disease, trauma, or complication respond differently. Some are medically robust, because they are far to the left (low-risk region) on the logit curve before the event occurred. Others are medically fragile, because their age or comorbid conditions place them close to the center of the logit curve. For the latter, a 1-logit-unit increase in risk can be the straw that breaks the camel’s back. It is this kind of relation that makes it difficult to show, for example, the benefit of bilateral internal thoracic artery grafting in young adults followed for even a couple of decades, but easy in patients who have other risk factors.\textsuperscript{31} Other types of analysis such as survival analysis have a similar S-shaped relation of outcome and risk, although the relationship might not be symmetric.

Raw Data Versus Models of Data

Because logistic regression and other statistical models generate an equation based on raw data (or an algorithm as from machine learning\textsuperscript{21}), it can be solved for a given

\[ P = \frac{1}{1 + \exp(-z)} \]

Fig. 1. Fundamental logistic relation of a scale of risk (logit units) to absolute probability of an event. Logistic relation, shown when risk factors are translated into logit units,\textsuperscript{28} is depicted along horizontal axis and probability of the outcome event along vertical axis. Logistic equation is inserted, where \( \exp \) is the natural exponential function.
set of values for risk factors. Whenever possible and appropriate, the results of clinical data analyses should be expressed in a form that can be solved after plugging in values for an individual patient’s risk factors to estimate absolute risk and its confidence limits. Equations are compact and portable, so that with the ubiquitous computer, they can be used to advise individual patients.\(^{32-34}\)

It can be argued that equations do not represent raw data. But in most cases, are we really interested in raw data? Archaeologists are interested in the past, but the objective of most clinical investigation is not to predict the past, but to draw inferences based on observations of the past that can be used in treating future patients. Thus, equations derived from raw data about the past are more useful than raw, undigested data.

**Nihilism Versus Predictability**

One of the important advantages of generating equations is that they can be used to predict future results for either groups of patients or individual patients. We recognize that when speaking of individual patients, we are referring to a prediction concerning the probability of events for that patient; we generally cannot predict exactly who will experience an event or when an event will occur.

Of course, the nihilist may say, “You can’t predict anything.” There are an increasing number of models for predicting institutional results of cardiac interventions, and many of these are narrowly focused on more or less homogenous subsets of patients. For many individuals with complex heart disease and comorbidities, these predictions may fall short of true risk because they do not account for all the complexities that a patient may bring.\(^{35}\)

**Blunt Instruments Versus Fine Dissecting Instruments**

A related use of predictive equations is in comparing alternative therapies: comparative effectiveness research. Some argue that the only believable comparisons are those based on randomized trials, and that clinical experience is irrelevant and misleading.\(^{36}\) However, many randomized trials are homogenous and focused and are analyzed by blunt instruments, such as an overall effect. On the other hand, real-world clinical experience involves patient selection that is difficult to quantify, may be a single-institution experience with limited generalizability except to other institutions of the same variety, is not formalized unless there is prospective gathering of clinical information into registries, and is less disciplined. Nevertheless, analyses of clinical experiences can yield a fine dissecting instrument in the form of equations that are useful for comparing alternative treatments and advising patients.

**Parsimony Versus Complexity**

Although clinical data analysis methods and results may seem complex at times, as in the large number of risk factors that must be assessed for comparing treatment strategies in ischemic heart disease, an important philosophy behind such analysis is parsimony (simplicity). Parsimony is needed because clinical data contain inherent redundancy, and 1 purpose of multivariable analysis is to identify that redundancy and simplify the dimensionality of the problem. Parsimony also aids assimilation of new knowledge by extracting the essence of the data. Thus, clinical inferences are often more simple than the multivariable analyses.

Simplicity is a virtue based on philosophic, not scientific, grounds. The concept was introduced by William of Ocken in the early fourteenth century as a concept of beauty: beauty of ideas and theories.\(^{37}\) Nevertheless, it is pervasive in science.
However, there are dangers associated with parsimony and beauty. The human brain seems to assimilate information in the form of models, not data. Thus, new ideas, innovations, breakthroughs, and new interpretations of the same data often hinge on discarding past paradigms. There are other dangers in striving for simplicity. Important relations may be missed because the threshold for detecting them is too high. Complex clinical questions may be reduced to simple but inadequate questions that we know how to answer.

New Knowledge Versus Selling Shoes

The philosophies described so far focus on the challenge of generating new knowledge from clinical experiences. However, clinical data have other uses.

Clinical data may be used as a form of advertising, just like selling shoes. Innovation stems less from purposefulness than from esthetically motivated curiosity, frustration with the status quo, sheer genius, fortuitous timing, favorable circumstances, and keen intuition. With innovation comes the need to promote, to sell the idea. However, promotional records of achievement should not be confused with serious study of safety, clinical effectiveness, and long-range appropriateness of interventions.

Of growing importance is the use of clinical information for regulation or to gain institutional competitive advantage (3 stars). Using clinical outcomes data to rank institutions or individual doctors has become popular in the United States. Many perceive clinical report cards as a means for punishment or regulation. What is troubling is that their use is based on a questionable quality-control model of outlier identification. Because doctors are people and not machines, this approach generates counterproductive ethical side effects, including defensiveness and hiding the truth. It hinders candid, nonculpable, serious examination of medical processes for the express purpose of improving patient care.

Critics of clinical report cards charge that to improve their rankings, some institutions refuse to operate on sicker patients. However, risk-adjusted mortality may remain high even for low-risk cases.

With the intense focus on institutional performance, another undesirable side effect of data analysis, decried years ago, has crept back in: undue emphasis on hospital mortality and morbidity. Studies of hospital events have the advantage of readily available data for extraction, but early events may be characterized incompletely. After repair of many congenital and acquired heart diseases, early risk of surgery extends beyond the hospital stay. This has led to reflection on the effect of time frame on studies of clinical experiences. Use of intermediate-term data is likely to characterize the early events well, but requires cross-sectional patient follow-up. Long-term follow-up is essential to establish appropriateness of therapy, but it is expensive and runs the risk of being criticized as being of historical interest only.

Clinical information is also used for profit or corporate advantage. At present, the philosophies of scientific investigation and business are incompletely reconciled. One thrives on open dissemination of information, the other on proprietary information offering a competitive advantage. In an era of dwindling public resources for research and increasing commercial funding, we may be seeing the beginning of the end of open scientific inquiry.

Past Versus Future

Is there a future for quantitative analysis of the results of therapy, as there was in the developmental phase of cardiac surgery and interventional cardiology? If treatment of heart disease requires complex procedures, and if most are palliative in the life history of chronic disease, there is a need to understand more fully the nature of the disease,
its treatment, and its optimal management. This requires approaches to data that are inescapably philosophic.

**CLINICAL RESEARCH**

In response to the American Medical Association’s Resolution 309 (I-98), a Clinical Research Summit and an ongoing Institute of Medicine Clinical Research Roundtable have sought to define and reenergize clinical research. The most important aspects of the definition of clinical research are that (1) it is but 1 component of medical and health research aimed at producing new knowledge; (2) the knowledge produced should be valuable for understanding the nature of disease, its treatment, and prevention; and (3) it embraces a wide spectrum of types of research. Here, we highlight 2 examples on that spectrum: clinical trials with randomly assigned treatment and clinical studies with nonrandomly assigned treatment, both of which are an integral part of clinical effectiveness research.

**Clinical Trials with Randomly Assigned Treatment**

Clinical trials in which cardiac interventions and medical therapy have been randomly assigned have made major contributions to our knowledge of treatment and outcomes of heart disease. Randomization of treatment assignment has 3 valuable and unique characteristics:

- It eliminates selection factors (bias) in treatment assignment (although this can be defeated at least partially by enrollment bias).
- It distributes patient characteristics equally between groups, whether they are measured or not, known or unknown.
- It meets assumptions of statistical tests used to compare end points.

Randomized clinical trials are also characterized by concurrent treatment, high quality and complete compilation of data gathered according to explicit definitions, and proper follow-up evaluation of patients. These operational by-products may have contributed nearly as much new knowledge as the random assignment of treatment.

It has become ritualistic for some to dismiss out of hand all information, inferences, and comparisons relating to outcome events derived from experiences in which treatment was not randomly assigned. If this attitude were valid, then much of the information used to manage patients with cardiac disease would need to be dismissed and ignored. However, moral justification (equipoise) may not be present for a randomized comparison of procedures and protocols that clinical experience strongly suggests, for at least some physicians, lead to important difference in outcomes. When Benson and Hartz investigated differences between randomized trials and observational comparisons over a broad range of medical and surgical interventions, they found “little evidence that estimates of treatment effects in observational studies reported after 1984 are consistently larger than or qualitatively different from those obtained in randomized controlled studies.” (However, see the rebuttal by Pocock and Elbourne.) These findings were confirmed by Concato and colleagues. Nevertheless, many acknowledge a hierarchy of clinical research study designs, and the randomized trial generates the most secure information about treatment differences.

Trials in which treatment is randomly assigned are testing a hypothesis, and hypothesis testing, in general, requires a yes or no answer unperturbed by uncontrollable factors. Thus, ideally, the study is of short duration, with all participants blinded and with a treatment that can be well standardized. However, in many clinical situations involving patients with heart disease, the time-relatedness of freedom from an
unfavorable outcome event is important and can jeopardize interpretation of the trial.\textsuperscript{63} This is because (1) individual patients assign different values to different durations of time-related freedoms (long-term benefit may be more important than short-term risk and vice versa), (2) differing severities of disease (and corresponding differences in natural history) affect different time frames, and (3) the longer the trial, the more likely there will be crossovers (such as from medical to interventional therapy).\textsuperscript{64} Also, the greater the number of risk factors associated with the condition for which treatment is being evaluated, the greater the potential heterogeneity (number of subsets) of patients with that condition and the greater the likelihood that a yes-no answer applies only to certain subsets of patients. In such situations, a randomized trial may have the disadvantage of including only a limited number of subsets or it may apply to no subset, because the average patient for whom the answer is derived may not exist except as a computation. Trials have addressed this problem by basing the randomization on subsets\textsuperscript{65} or by later analyzing subsets by stratification (but see concerns raised by Guillemin\textsuperscript{66}) or by multivariable analysis.\textsuperscript{67}

These considerations, in addition to ethical concerns,\textsuperscript{12,68,69} have fueled the debate on whether surgery and invasive intervention in general are an appropriate arena for randomized trials of innovation, devices, and operations.\textsuperscript{59,70–72} Some argue strongly that randomization should be required at the outset of introducing a new therapy.\textsuperscript{69} In 3 related articles arising from the Balliol Colloquium held at the University of Oxford between 2007 and 2009, clinicians and anesthesiologists sought to clarify the issues surrounding surgical clinical trials.\textsuperscript{73–75} They recognized important stages in developing a surgical technique, starting with innovation and progressing through development, exploration, assessment, and long-term outcomes. They then explored options for evaluative studies and barriers to each, including sham operations and nonoperative treatment alternatives.\textsuperscript{74} They ended with an IDEAL model for surgical development (idea, development, exploration, assessment, long-term study), and the role of feasibility randomized trials in exploration, definitive trials in assessment, and registries in long-term surveillance.\textsuperscript{75}

Moses\textsuperscript{76} and others\textsuperscript{62,70,77} present the case for a balance between randomized clinical trials and observational clinical studies. However, observational studies are beset with problems of selection bias and skill variance; thus, not to be overlooked are the development and rapid introduction of powerful new methods for drawing causal inferences from nonrandomized trials.\textsuperscript{78,79}

\textit{Clinical Studies with Nonrandomly Assigned Treatment}

Clinical studies with nonrandomly assigned treatment produce little new knowledge when improperly performed and interpreted. Because this is often the case, many investigators have a strong bias against these studies. However, when properly performed and interpreted, and particularly when they are multi-institutional or externally validated, clinical studies of real-world experience can produce secure knowledge.

The fundamental objection to using observational clinical data for comparing treatments is that many uncontrolled variables affect outcome.\textsuperscript{80} Thus, attributing outcome differences to just 1 factor (alternative treatments) stretches credibility. Even a cursory glance at the characteristics of patients treated 1 way versus another usually reveals that they are, on average, different groups. This should be expected, because treatment has been selected by experts who believe that they know what is best for a given patient. The accusation that apples and oranges are being compared is justified!\textsuperscript{81} Multivariable adjustment for differences in outcome is valuable but not guaranteed to be effective in eliminating selection bias as the genesis of a difference in outcome (a form of confounding).\textsuperscript{79,82,83}
Balancing Scores

Apples-to-apples nonrandomized comparisons of outcome can be achieved, within certain limitations, by use of balancing scores, of which the propensity score is the simplest.\(^7^8\) Balancing scores are a class of multivariable statistical methods that identify patients with similar chances of receiving 1 or the other treatment. Patients with similar balancing scores are well balanced with respect to nearly all patient, disease, and comorbidity characteristics taken into account in forming the balancing score. This balancing of characteristics permits the most reliable nonrandomized comparisons of treatment outcomes available. Developers of balancing score methods claim that the difference in outcome between patients who have similar balancing scores but receive different treatments provides an unbiased estimate of the effect attributable to the comparison variable of interest.\(^7^8\) That is technical jargon for saying that the method can identify the apples from among the mixed fruit of clinical practice variance, transforming an apples-to-oranges outcomes comparison into an apples-to-apples comparison.\(^8^4–^8^7\)

Randomly assigning patients to alternative treatments in clinical trials balances both patient characteristics (at least in the long run) and number of subjects in each treatment arm. In a nonrandomized setting, neither patient characteristics nor number of patients is balanced for each treatment. A balancing score achieves local balance of patient characteristics at the expense of unbalancing \(n\). Tables 1 and 2 show local balance of patient characteristics achieved by using a specific balancing score known as the propensity score. Table 1 shows that patients on long-term aspirin therapy have dissimilar characteristics from those not on this therapy. Unadjusted comparison of outcomes in these 2 groups is invalid (an apples-to-oranges comparison).\(^8^1\) There-fore, multivariable logistic regression analysis was performed to identify factors predictive of treatment received (chronic aspirin therapy vs not).\(^8^8\) The resulting logistic equation was solved for each patient’s probability of being on long-term aspirin therapy. This probability is 1 expression of what is known as a propensity score (in this case, the propensity to be on chronic aspirin therapy). Patients were then sorted according to the balancing (propensity) score and divided into 5 equal-size groups, called quintiles, from low score to high.\(^7^8\) Patients in each quintile had similar balancing scores (see Table 2).

### Table 1

<table>
<thead>
<tr>
<th>Patient Characteristic</th>
<th>ASA</th>
<th>No ASA</th>
<th>(P)</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>2455</td>
<td>4072</td>
<td></td>
</tr>
<tr>
<td>Men (%)</td>
<td>49</td>
<td>56</td>
<td>.001</td>
</tr>
<tr>
<td>Age (mean ± SD y)</td>
<td>62 ± 11</td>
<td>56 ± 12</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Smoker (%)</td>
<td>10</td>
<td>13</td>
<td>.001</td>
</tr>
<tr>
<td>Resting heart rate (beats/min)</td>
<td>74 ± 13</td>
<td>78 ± 14</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Ejection fraction (%)</td>
<td>50 ± 9</td>
<td>53 ± 7</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

**Abbreviations:** ASA, long-term aspirin use; SD, standard deviation.

* Table shows that patient characteristics differ importantly, making direct comparisons of outcome invalid. As shown in original article, many other patient characteristics differed between the 2 groups.

* Data from Gum PA, Thamilarasan M, Watanabe J, et al. Aspirin use and all-cause mortality among patients being evaluated for known or suspected coronary artery disease: A propensity analysis. JAMA 2001;286:1187–94.
Simply by virtue of having similar balancing scores, patients within each quintile were found to have similar characteristics (except for age in quintile I). As might be expected, patient characteristics differed importantly from 1 quintile to the next. For example, most patients in quintile I were women; most in quintile V were men. Except for unbalanced n, these quintiles look like 5 individual randomized trials with differing inclusion and exclusion criteria, which is what balancing scores are intended to achieve. Thus, the propensity score balanced essentially all patient characteristics within localized subsets of patients, in contrast to randomized clinical trials, which balance both patient characteristics and n globally within the trial.

To achieve this balance, a widely dissimilar number of patients received long-term aspirin therapy from quintile to quintile. Quintile I contained only a few patients who received long-term aspirin therapy, whereas quintile V had few not receiving aspirin.

**Propensity score**

The most widely used balancing score is the propensity score. It provides for each patient an estimate of the propensity toward (probability of) belonging to 1 group versus another (group membership). The following sections describe (1) designing the nonrandomized study, (2) constructing a propensity model, (3) calculating a propensity score for each patient using the propensity model, and (4) using the propensity score in various ways to achieve a balanced comparison.

**Designing the nonrandomized study**

The essential approach to a comparison of treatment outcomes in a nonrandomized setting is to design the comparison as if it were a randomized clinical trial and to interpret the resulting analyses as if they emanated from such a trial. This essential approach is emphasized in Rubin’s 2007 article, “The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials.”

As noted by Rubin, “I mean all contemplating, collecting, organizing, and analyzing data that takes place before seeing any outcome data.” He emphasizes by this

---

**Table 2**

<table>
<thead>
<tr>
<th>Patient Characteristic</th>
<th>Quintile I ASA</th>
<th>No ASA</th>
<th>Quintile II ASA</th>
<th>No ASA</th>
<th>Quintile III ASA</th>
<th>No ASA</th>
<th>Quintile IV ASA</th>
<th>No ASA</th>
<th>Quintile V ASA</th>
<th>No ASA</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>113</td>
<td>1092</td>
<td>194</td>
<td>1111</td>
<td>384</td>
<td>922</td>
<td>719</td>
<td>586</td>
<td>1045</td>
<td>261</td>
</tr>
<tr>
<td>Men (%)</td>
<td>22</td>
<td>22</td>
<td>57</td>
<td>63</td>
<td>74</td>
<td>71</td>
<td>78</td>
<td>78</td>
<td>88</td>
<td>87</td>
</tr>
<tr>
<td>Age (y)</td>
<td>55</td>
<td>49</td>
<td>56</td>
<td>55</td>
<td>61</td>
<td>61</td>
<td>62</td>
<td>64</td>
<td>63</td>
<td>65</td>
</tr>
<tr>
<td>Smoker (%)</td>
<td>15</td>
<td>13</td>
<td>15</td>
<td>15</td>
<td>12</td>
<td>11</td>
<td>11</td>
<td>13</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>Resting heart rate</td>
<td>84</td>
<td>83</td>
<td>79</td>
<td>79</td>
<td>76</td>
<td>76</td>
<td>76</td>
<td>76</td>
<td>71</td>
<td>73</td>
</tr>
<tr>
<td>Ejection fraction (%)</td>
<td>53</td>
<td>54</td>
<td>54</td>
<td>54</td>
<td>53</td>
<td>53</td>
<td>49</td>
<td>49</td>
<td>49</td>
<td>48</td>
</tr>
</tbody>
</table>

*Table shows that balancing patient characteristics by the propensity score comes at the expense of unbalancing number of patients within comparable quintiles.*

Data from Gum PA, Thamilarasan M, Watanabe J, et al. Aspirin use and all-cause mortality among patients being evaluated for known or suspected coronary artery disease: a propensity analysis. JAMA 2001;286:1187–94.
statement his thesis that a nonrandomized set of observations should be conceptualized as “a broken randomized experiment...with a lost rule for patient allocation, and specifically for the propensity score, which the analysis will attempt to construct.” For example, the investigator should ask, “Could each patient in all comparison groups be treated by all therapies considered?” If not, this constitutes specific inclusion and exclusion criteria. If this were a randomized trial, when would randomization take place? Variables must be used to construct a propensity score that would be known only at the time randomization would have occurred, not after that; this means that variables chosen in the propensity score analysis are not those that could be affected by the treatment.

**Constructing a propensity model** For a 2-group comparison, typically, multivariable logistic regression is used to identify factors predictive of group membership. In most respects, this is what clinical investigators have done for years: find correlates of an event. In this case, it is not risk factors for an outcome event, but rather correlates of membership in 1 or the other comparison group of interest.

I recommend initially formulating a parsimonious multivariable explanatory model that identifies common denominators of group membership. Once this traditional modeling is completed, a further step is taken to generate the propensity model, which augments the traditional model by other factors, even if not statistically significant. Thus, the propensity model is not parsimonious. The goal is to balance patient characteristics by whatever means possible, incorporating all information recorded that may relate to either systematic bias or simply bad luck, no matter the statistical significance. It is important to use as many continuous variables as possible to represent these patient characteristics because it produces a fine, as opposed to coarse, set of values when the propensity score is calculated.

When taken to the extreme, forming the propensity model can cause problems, because medical data tend to have many variables that measure the same thing (redundancy). The solution is to pick 1 variable from among a closely correlated cluster of variables as a representative of the cluster. An example is to select 1 variable representing body size from among height, weight, body surface area, and body mass index.

**Calculating a propensity score** Once the propensity modeling is completed, a propensity score is calculated for each patient. A logistic regression analysis, as used for the propensity model, produces a coefficient or numeric weight for each variable. The coefficient maps the units of measurement of the variable into units of risk. Specifically, a given patient’s value for a variable is transformed into risk units by multiplying it by the coefficient. If the coefficient is 1.13 and the variable is “male” with a value of 1 (for “yes”), the result is 1.13 risk units. If the coefficient is 0.023 for the variable “age” and a patient is 61.3 years old, 0.023 times 61.3 is 1.41 risk units.

One continues through the list of model variables, multiplying the coefficient by the specific value for each variable. When finished, the resulting products are summed. To this sum is added the intercept of the model, and the result is the propensity score. Technically, the intercept of the model, which is constant for all patients, does not have to be added; however, in addition to using the propensity score in logit risk units as described here, it may be used as a probability, for which the intercept is necessary.

**Using the propensity score for comparisons** Once the propensity model is constructed and a propensity score is calculated for each patient, 3 common types of comparison are used: matching, stratification, and multivariable adjustment. The propensity score can be used as the sole criterion for matching pairs of patients (Table 3). Although several matching strategies have been used by statisticians.
for many years, new optimal matching algorithms have arisen within computer science and operations research. These algorithms have been motivated by the need to optimally match volume of intranet and Internet traffic to computer network configurations. In addition, (Rubin, personal communication, 2008) has suggested matching with replacement versus the usual “greedy” matching, which removes matched patients from further consideration. Matching can be bootstrapped, creating multiple matched comparison groups, over which outcome can be averaged.94

Exact matches are rarely found. Instead, a patient is selected from the smaller of the 2 groups being compared with a propensity score nearest to that of a patient in the larger group. If multiple patients are close in propensity scores, optimal selection among these candidates can be used.84 Problems of matching on multiple variables disappear by compressing all patient characteristics into a single score (compare Table 3 with unmatched data in Table 1).86

Matching works well. The comparison data sets have all the appearances of a randomized study. The average effect of the comparison variable of interest is assessed as the difference in outcome between the groups of matched pairs. However, unlike a randomized study, the method does not balance unmeasured variables well, and this may be fatal to the inference.

Once patients are matched, it is important to diagnostically test the quality of matching. This test can be accomplished visually by graphs of standardized differences, defined as the difference in mean value between groups divided by the pooled standard deviation. This quantity is similar to the test statistic in a $t$-test (Fig. 2).84 Differences that were substantial should virtually disappear. If they do not, it is possible that interaction terms (multiplicative factors rather than additive factors) may be required.

A graph of propensity scores for the groups is instructive (Fig. 3). The scores for 2 treatments may nearly overlap, as they would for a randomized trial. On the other hand, there may be little overlap, as in Fig. 4, and the comparison focuses on the center part of the spectrum of propensity scores, where there is substantial overlap (the region of virtual equipoise).

Outcome can be compared within broad groupings of patients, called strata or subclasses, according to propensity score. After patients are sorted by propensity score, they are divided into equal-sized groups. For example, they may be split into

<table>
<thead>
<tr>
<th>Patient Characteristic</th>
<th>ASA</th>
<th>No ASA</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>1351</td>
<td>1351</td>
</tr>
<tr>
<td>Men (%)</td>
<td>49</td>
<td>51</td>
</tr>
<tr>
<td>Age (y)</td>
<td>60</td>
<td>61</td>
</tr>
<tr>
<td>Smoker (%)</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>Resting heart rate (beats/min)</td>
<td>77</td>
<td>76</td>
</tr>
<tr>
<td>Ejection fraction (%)</td>
<td>51</td>
<td>51</td>
</tr>
</tbody>
</table>

*Table shows ability of the propensity score to produce what appears to be a randomized study balancing both patient characteristics and n.

Data from Gum PA, Thamilarasan M, Watanabe J, et al. Aspirin use and all-cause mortality among patients being evaluated for known or suspected coronary artery disease: a propensity analysis. JAMA 2001;286:1187–94.
Fig. 2. Covariable balance plot before and after propensity score matching on selected covariables. Symbols depict percent standardized differences for covariables between patients in less invasive and conventional groups. BMI, body mass index; BUN, blood urea nitrogen; COPD, chronic obstructive pulmonary disease; Dysfunct., dysfunction; HTN, hypertension; LV, left ventricular; Regurg., regurgitation; TV, tricuspid valve. (Data from Johnston DR, Atik FA, Rajeswaran J, et al. Outcomes of less invasive J-incision approach to aortic valve surgery. J Thorac Cardiovasc Surg 2012;144:852–8.e3.)

Fig. 3. Distribution of propensity scores for conventional and less invasive approaches for aortic valve replacement. (Data from Johnston DR, Atik FA, Rajeswaran J, et al. Outcomes of less invasive J-incision approach to aortic valve surgery. J Thorac Cardiovasc Surg 2012;144:852–8.e3.)
5 groups, or quintiles (see Table 2; Table 4), but fewer or more groups may be used, depending on the size of the study. Comparison of outcome for the comparison variable of interest is made within each stratum. If a consistent difference in outcome is not observed across strata, intensive investigation is required. Usually, something is discovered about the characteristics of the disease, the patients, or their clinical condition that results in different outcomes across the spectrum of disease. For example, in their study of ischemic mitral regurgitation, Gillinov and colleagues discovered that the difference in survival between those undergoing valve repair versus replacement progressively narrowed as complexity of the pattern of regurgitation increased and condition of the patient worsened (Fig. 5). Apparent anomalies such as this give important insight into the nature of the disease and its treatment.

The propensity score for each patient can be included in a multivariable analysis of outcome. Such an analysis includes both the comparison variable of interest and the propensity score. The propensity score adjusts the apparent influence of the comparison variable of interest for patient selection differences not accounted for by other variables in the analysis.

Occasionally, the propensity score remains statistically significant in such a multivariable model. This constitutes evidence that adjustment for selection factors by multivariable analysis alone is ineffective, something that cannot be ignored. It may mean...
that not all variables important for bias reduction have been incorporated into the model, such as when a simple set of variables is used. It may mean that an important modulating or synergistic effect of the comparison variable occurs across propensity scores, as noted in Fig. 5 (eg, the mechanism of disease may be different within quintiles). It may mean that important interactions of the variable of interest with other variables have not been accounted for, leading to a systematic difference identified by the propensity score. The collaborating statistician must investigate and resolve these possibilities. Understanding aside, this statistically significant propensity score has performed its intended function of adjusting the variable representing the group difference.

In some settings in which the number of events is small, the propensity score can be used as the sole means of adjusting for the variable representing the groups being compared.92

**Oranges**
The propensity score may reveal that a large number of patients in 1 group do not have scores close to patients in the other.31 Thus, some patients may not be matched. If stratification is used, quintiles of patients may have hardly any matches at 1 or the other, or both, ends of the propensity spectrum, and these remaining may not be well matched.

The knee-jerk reaction is to infer that these unmatched patients represent apples and oranges unsuited for direct comparison.96 However, the most common reason for lack of matches is that a strong surrogate for the comparison group variable has

---

### Table 4
Balance in patient and selection characteristics achieved by unbalancing number of cases in each propensity-ranked group in 3 separate studies

<table>
<thead>
<tr>
<th>Study</th>
<th>Factor Present (n)</th>
<th>Factor Absent (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Long-Term Aspirin Use</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quintile 1</td>
<td>113</td>
<td>1192</td>
</tr>
<tr>
<td>Quintile 2</td>
<td>194</td>
<td>1111</td>
</tr>
<tr>
<td>Quintile 3</td>
<td>384</td>
<td>922</td>
</tr>
<tr>
<td>Quintile 4</td>
<td>719</td>
<td>586</td>
</tr>
<tr>
<td>Quintile 5</td>
<td>1045</td>
<td>261</td>
</tr>
<tr>
<td>Natural Selection: Preoperative AF in Degenerative MV Disease</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quintile 1</td>
<td>2</td>
<td>225</td>
</tr>
<tr>
<td>Quintile 2</td>
<td>13</td>
<td>214</td>
</tr>
<tr>
<td>Quintile 3</td>
<td>32</td>
<td>195</td>
</tr>
<tr>
<td>Quintile 4</td>
<td>78</td>
<td>149</td>
</tr>
<tr>
<td>Quintile 5</td>
<td>162</td>
<td>66</td>
</tr>
<tr>
<td>OPCAB vs On-Pump</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quintile 1</td>
<td>40</td>
<td>702</td>
</tr>
<tr>
<td>Quintile 2</td>
<td>71</td>
<td>671</td>
</tr>
<tr>
<td>Quintile 3</td>
<td>61</td>
<td>682</td>
</tr>
<tr>
<td>Quintile 4</td>
<td>90</td>
<td>652</td>
</tr>
<tr>
<td>Quintile 5</td>
<td>219</td>
<td>524</td>
</tr>
</tbody>
</table>

*Abbreviations*: AF, atrial fibrillation; MV, mitral valve; OPCAB, off-pump coronary artery bypass grafting.
been included inadvertently in the propensity score. This variable must be removed and the propensity model revised. If this is not the case, the analysis may have identified truly unmatchable cases (mixed fruit). In some settings, they represent a different end of the spectrum of disease, for which different therapies have been applied systematically.\(^95,97\) Often, the first clue to this anomaly is finding that the influence of the comparison variable of interest is inconsistent across quintiles.\(^95\) This finding emphasizes the nature of comparisons with balancing score methodology: the comparisons relate only to the subset of patients who are apples-to-apples. Comparing these apples with the remaining oranges with respect to outcomes is not valid. The oranges result from systematic selection of patients for one versus the other treatment.

Thus, when apples and oranges (and other mixed fruit) are revealed by a propensity analysis, investigation should be intensified rather than the oranges being set aside.\(^81\) After the investigations are complete, comparisons among the well-matched patients can proceed with known boundaries within which valid comparisons are possible.

**Limitations**

Some investigators claim that balancing score methods are valid only for large studies, citing Rubin.\(^98\) It is true that large numbers facilitate certain uses of these scores, such as...
as stratification. However, experience suggests that there is considerable latitude in matching that still reduces bias; the method seems to work even for modest-sized data sets.

Another limitation is having few variables available for propensity modeling. The propensity score is seriously degraded when important variables influencing selection have not been collected. A corollary to this is that unmeasured variables cannot be reliably balanced. If these variables are influential on outcome, a spurious inference may be made.

The propensity score may not eliminate all selection bias. This may be attributed to limitations of the modeling itself imposed by the linear combination of factors in the regression analysis that generates the balancing score. If the comparison data sets are comparable in size, it may not be possible to match every patient in the smaller of the 2 data sets, because closely comparable patients have been used up, unless bootstrap sampling with replacement has been used.

Perhaps the most important limitation is inextricable confounding. Suppose on-pump coronary artery bypass grafting is to be compared with off-pump operations. A study is designed to compare the results of institution A, which performs only off-pump bypass, with those of institution B, which performs only on-pump bypass. Even after careful application of propensity score methods, it remains impossible to distinguish between an institutional and a treatment difference, because they are inextricably intertwined (confounded); that is, the values for institution and treatment are 100% correlated.

**Extension**

At times, a comparison of more than 2 groups may be wanted. Under this circumstance, multiple propensity models are formulated, such as by fully conditional multiple logistic regression.

Most applications of balancing scores have been concerned with dichotomous (yes-no) comparison group variables. However, balancing scores can be extended to a multiple-state ordered variable (ordinal) or even a continuous variable. An example of the latter is use of correlates of heart valve prosthesis size as a balancing score to isolate the possible causal influence of valve size on outcome.

Logistic regression is not the only way to formulate propensity scores. A nonparametric machine learning technique (random forests) can be used and has been found by Lee and colleagues to better balance groups, with reduced bias. Our group has formulated a generalized theorem as an extension of the work of Imai and van Dyk for propensity scores and devised a data-adaptive, random-forest, nearest-neighbor algorithm that simultaneously matches patients and estimates the treatment effect from thousands of bootstrap samples and simultaneously refines the characteristics of true oranges (noncomparable patients).

**TECHNIQUE FOR SUCCESSFUL CLINICAL RESEARCH**

Marbán and Braunwald, in reflecting on training the clinician-investigator, provide guiding principles for successful clinical research. Among these:

- Choose the right project
- Embrace the unknown
- Use state-of-the-art approaches
- Do not become the slave of a single technique
- Never underestimate the power of the written or the spoken word
Because of increasingly limited resources for conducting serious clinical research, a deliberate plan is needed to successfully carry a study through from inception to publication. The following sections outline such a plan for study of a clinical question for which clinical experience (a patient cohort) provides the data. This plan appears as a linear work flow (Fig. 6); in reality, most research efforts do not proceed linearly but rather iteratively, with each step more refined and usually more focused, right up to the last revision of the article.

**Research Proposal**

Every serious clinical study needs a formal proposal that clarifies the question. A common mistake is to ask questions that are unfocused, uninteresting, unimportant, or overworked. Marbán and Braunwald write, “Ask a bold question... about which you can feel passionate.” Brainstorming with collaborators is essential.

The next step is to define the inclusion and exclusion criteria for the study group. A common mistake is to define this group too narrowly, such that cases are overlooked or homogeneity precludes discovering trends. Inclusive dates should be considered carefully. Readers are suspicious if the dates are not whole years or at least half-years. Similarly, suspicion arises when a study considers a neat number of patients, such as “the first 1000 aortic valve procedures.”

In defining the study group, particular care should be taken to include a denominator or comparison group to put events into context. A retrospective study is a study of only numerators, such as only patients who have required renal dialysis after an intervention. If the denominator is included, it is a prospective or cohort study. If the cohort is placed in the context of alternative approach to the disease, it is a comparative effectiveness study.

End points (results, outcomes) must be clearly defined in a reproducible fashion. Generally, every event should be accompanied by its date of occurrence. A common failing is that repeated end points (eg, thromboembolism or assessments of functional status) are recorded only the first or most recent time that they occur. Techniques are available to analyze repeated end points.

Careful attention must be paid to the variables that are studied. They should be pertinent to the study question. A common failing is to collect values for too many variables such that quality suffers. This error is understandable: the clinical investigator reasons that because the patient’s records must be reviewed, several other variables should also be abstracted at that time. Or, realizing the complexity of the clinical setting, the investigator feels compelled to collect information on all possible ramifications of the study, even if it is peripheral to the focus of the study. This is termed the “Christmas tree effect”, meaning adding ornament on ornament until they dominate what once was a fine tree. There needs to be a balance between so sparse a set of variables that little can be done by way of risk factor identification or balancing characteristics of the group, and so rich a set of variables that the study flounders or insufficient care is given to the quality and completeness of relevant variables.

Study feasibility must then be assessed. A common failing is forgetting that if an outcome event is the end point, the effective sample size is the number of events observed. A study may have 1000 patients, but if only 10 events are observed, one cannot find multiple risk factors for those events.

Investigators should plan data analysis at the beginning of a study. Often, the setup for the analysis data set is specific to the methods of analysis. This needs to be communicated to the data managers (see Box 1).
Clinical Research Proposal

Specify study group

Translate into query language

Query

**Study group identified**

Narrow group?

Are the data sufficient?

No

STOP

Yes

Proposed Variables and Values

Identify data sources

Exist electronically

New

Extract information from each source

Assemble patient documents

Create and test database

Extract values for variables

Verify values

Join electronic data sets

Variable and Value Manipulation

Create intervals from date: time variables

Create (numeric) indicator variables for:

- Dichotomous values
- List of values
- Events
- Missing values

Combine values

Continue to screen and scrub values for variables

Impute missing values

Organize variables into medically meaningful categories

Clinical Analysis Data Set(s)

**Fig. 6.** Linearized work flow for a clinical research study: Transforming information to data suited for analysis.
A necessary step is review of the literature. Sifting through articles is often painful, but it should result in identifying a few key articles that are pertinent to the study. The search is too often confined to recent literature, and this may result in duplication.

A realistic timeframe with deliverables should be established with collaborators. A common failing is not providing sufficient time for data verification and other aspects of data management that are the heart of a high-quality study. Analysis of data may consume one-tenth the time of high-quality data preparation.

The completed formal research proposal is likely to be updated throughout the course of a study, facilitated by online tracking of each study, with periodic updates of the protocol as necessary.

**Database Development and Verification**

The next step for successful research is careful attention to the data themselves. If electronically available data are to be used, every variable must be defined both medically and at the database content level. If data are to be collected de novo, then an appropriate database must be developed (see Fig. 6). Every variable must be in a format of 1 value per variable. These variables must follow a controlled vocabulary for analysis, not free text.

**Data Collection**

A core set of variables should be collected for each patient. In cardiac surgical and interventional cardiology settings, these data elements are stipulated by regulatory agencies (eg, the state of New York) or societies (eg, Society of Thoracic Surgeons National Database, the American College of Cardiology National Cardiovascular Data Registry). They include demographics (it is essential to record patients’ date of birth rather than age because age can be calculated from date of birth to any chosen time zero), cardiac procedure and clinical symptoms and status at time of procedure, cardiac medical history (particularly previous cardiac procedures), disease cause, coexisting cardiac defects, coexisting noncardiac morbidity (such as diabetes), laboratory measurements known to be consistently associated with clinical outcomes, findings of diagnostic testing, findings during the procedure, support techniques during the procedure, and factors related to experience (such as date of procedure).

Investigators also need variables specific to a particular study. These variables should be identified and reproducibly defined. The danger is specifying too many variables; however, a thoughtfully compiled list adds depth to a study. Further, experienced investigators realize that in the midst of a study, it occasionally becomes evident that some variables require refinement, others collecting de novo, others rechecking, and others redefining. When this occurs, the variables must be refined, collected, rechecked, or redefined uniformly for every patient in the study.

Clinical studies are only as accurate and complete as the data available in patients’ records. Therefore, physician-investigators seriously interested in scientific progress must ensure that their preintervention, intervention, and postintervention records are clear, organized, precise, and extensive, so that information gathering can be complete and meaningful. Records should emphasize description; although records may contain the conclusions of the moment, descriptions of basic observations become useful in later analyses.

**Verification**

The first step in data verification is to enter values for each data element (variable) for 5 to 10 patients only. These values reveal problems of definition, incomplete pick lists, missed variables, difficult-to-find variables that may not be worth the effort to locate,
poor-quality variables, inconsistent recording, and questionable quality of observations. Once these issues are addressed, general data abstraction may proceed.

When all values for variables are in a computer database, formal verification commences. This verification can take 3 general forms: (1) value-by-value checking of recorded data against primary source documents, (2) random quality checking, and (3) automatic reasonableness checking. If a routine activity of recording core data-elements is used, it is wise to verify each element initially to identify those that are rarely in error (these can be spot checked by a random process) and those that are more often in error. The latter are usually a small fraction of the whole and are often values requiring interpretation. These values may require element-by-element verification.

When it is thought that data are correct (this is an iterative process with verification), they are checked for reasonableness of ranges, including discovery of inconsistencies among correlated values. For example, the database may indicate that a patient had a quadrangular resection of the mitral valve, but someone had failed to record that the posterior leaflet was prolapsing and had ruptured chordae, or the database records that a patient is 60 cm tall and weighs 180 kg; this is likely a problem of confused units of measurement.

**Data Conversion for Analysis**

An often underappreciated, unanticipated, and time-consuming effort is the conversion of data elements residing in a database to a format suitable for data analysis. Even if the day comes when all medical information is recorded as values for variables in a computer-based patient record, this step is unavoidable. Statistical procedures require data to be arranged in columns and rows, with each column representing values for a single variable (often in numeric format), and each row either a separate patient or multiple records on a single patient (as in many-to-1 repeated-measures longitudinal data analysis). This conversion process may involve redundancy, such as the necessity to again document all variables and provide a data key to the possible values for each.

This process nearly always involves creating additional variables from a single variable, such as a separate variable for each mutually exclusive cause of cardiomyopathy. These polytomous variables (lists) are then converted to a series of dichotomous variables (best expressed as 0 for absence and 1 for presence of the listed value).

Some categorical variables are ordinal, such as NYHA functional classes. These variables may need to be reformulated as an ordered number sequence (eg, 1–4). Variables recorded with units (eg, weight in kilograms, weight in pounds) must be converted to a common metric.

Calculated variables are also formed. These variables include body surface area and body mass index from height and weight, z values from measured cardiac dimensions, ejection fraction from systolic and diastolic ventricular volumes, intervals between date and time variables for which event indicator variables are created, and many other calculations. Because data conversion, creation of derived variables, and formation of calculated variables is time consuming and error prone, groups that conduct a large number of studies often store trusted, well-verified computer code to perform these operations on a repetitive basis.

Often information is gathered from multiple databases with queries, concatenations, and joining functions. These otherwise arduous functions can, under some circumstances, be automated. Alternatively, a data warehouse composed of multiple disparate electronic data sources can be implemented and maintained.
Managing sporadic missing data is important. If too many data are missing, the variable may be unsuitable for use in analyses. Otherwise, missing value imputation is necessary so that entire patients are not removed from analyses, the default option in many analysis programs.108

Data Analysis

The data analysis process should lead first to understanding of the raw data, often called exploratory data analysis.109 This understanding is gleaned from simple descriptive statistics, correlations among variables, simple life tables for time-related events, cumulative distribution graphs of continuously distributed variables, and cluster analyses, whereby variables with shared information content are identified.

The analytical process then attempts to extract meaning from the data by various methods akin to pattern recognition.21 Answers are sought for the following questions: which variables relate to outcome and which do not? What inference can be made about whether an association is or is not attributable to chance alone? Might there be a causal relationship? For what might a variable associated with outcome be a surrogate?

What will be discovered is that answering such questions in the most clinically relevant way often outstrips available statistical, biomathematical, and algorithmic methodology. Instead, a question is answered with available techniques, but not the question. One of the purposes of this article is to stimulate collaboration between physician-investigators and data analysis experts so that data are analyzed thoroughly and with appropriate methodology.

Interpreting Analyses

It is one thing for a statistician to provide a statistical inference; it is another for the physician-investigator to draw meaningful interpretations that affect patient care (clinical inferences).

The most successful way to embark on this interpretive phase of clinical research is to write down the truest 2 or 3 sentences that capture the essence of the findings (and no more).110 This important exercise produces an ultramini abstract for an article (whether or not it is required by a journal) and provides the roadmap for writing the article.111

Communicating the Findings

A common error of the physician-investigator is to summarize the data instead of drawing meaningful clinical inferences from the data and analyses by asking the following questions: (1) What new knowledge has been gleaned from the clinical investigation? (2) How can this new knowledge be incorporated into better patient care? (3) What do the data suggest in terms of basic research that needs to be stimulated? (4) How can I best communicate information to my colleagues? (5) How can I best present this information to the cardioligic world at large?

Meaningful new knowledge may not be generated because the statistical inferences from data analyses are accepted as the final result. Results need to be studied carefully. Often, this leads to additional analyses that increasingly illuminate the message that the data are trying to convey. Graphic depictions are of particular importance in transforming numbers into insight. Depictions must lead beyond statistical inference to clinical inference. What have the data revealed about how to better care for patients? This question is the one best linked to the original purpose of the study. If the study has suggested ways to improve patient care, the next step is to put what has been learned into practice.
Most studies generate more new questions than they answer. Some of these new questions require additional clinical research. Others require the physician-investigator to stimulate colleagues in the basic sciences to investigate fundamental mechanisms of the disease process.

Because most physician-investigators are part of a group, an important facet of generating new knowledge is discussing with colleagues the results, statistical and clinical inferences, and implications of a study. Multiple points of view nearly always clarify rather than obscure their interpretation.

Clinical research is not a proprietary activity. Yet, too often research does not result in an article. One reason may be that an abstract was not accepted for a meeting, perhaps because the data were not thoroughly digested before its submission. Although abstract deadlines may be important mechanisms for wrapping up studies, they too often stifle a serious and contemplative approach to generating new knowledge. A second reason articles do not get written is that the physician-investigator views the task as overwhelming. Possibly they have not developed an orderly strategy for writing. A third barrier to writing is time demands on the physician-investigator. Usually, this situation results from not making writing a priority in their professional life. This is a decision that should be made early in one’s medical career. If dissemination of new knowledge is a desire, then writing must be made a high priority part of one’s life style.

REFERENCES