Key Issues in Outcomes Research

Some Old and Some New Statistical Tools for Outcomes Research

Sharon-Lise T. Normand, PhD

Outcomes research “seeks to understand the end results of particular health care practices and interventions”1 to inform the development of clinical practice guidelines, evaluate the quality of medical care, and foster effective interventions to improve the quality of care.2 Although randomized trial designs have been used to assess quality of care and to identify effective interventions in the real world,3,4 the empirical basis of outcomes research largely rests on data collected in the observational setting (eg, in the routine setting of everyday practice).

With more emphasis placed on increasing the value of health care in terms of lives saved and morbidity avoided, outcomes researchers are making unprecedented demands of observational databases. This is evidenced by the increasing number of and participation in national registries. These include the National Cardiac Data Registry of the American College of Cardiology; the Implantable Cardioverter Defibrillator Registry launched jointly by the American College of Cardiology and the Heart Rhythm Society; the National Cardiac Surgery Database of the Society of Thoracic Surgeons; and the Interagency Registry of Mechanically Assisted Circulatory Support Devices funded by the National Heart, Lung, and Blood Institute, the Centers for Medicare and Medicaid Services, and others. Empirical analyses of these databases require statistical tools that can handle the complexity of the data: observational, sometimes hierarchical, often with multiple outcomes, and always with some missing data.

The purpose of the present article is to review key statistical methods important to outcomes research and to introduce newer methodology. The article describes 4 methodological issues commonly present when observational data are analyzed; summarizes the primary assumptions associated with strategies to handle these common problems; demonstrates methods to assess the plausibility of the assumptions associated with each strategy; and illustrates these concepts using examples from cardiovascular outcomes research. Although the intent of the present report is not to provide a comprehensive summary of statistical approaches to data analysis, it is intended to provide a clear understanding of the assumptions associated with some common methodological tools and of the strategies used to assess their plausibility. If these 2 goals are achieved, then both the rigor of and the scientific findings from outcomes research will be strengthened substantially.

Four Common Statistical Problems

An observational study is an empirical investigation in which the objective is to understand causal effects. Specifically, an observational investigation “concerns treatments, interventions, or policies, and the effects they cause.”5 Because subjects are not randomized to treatments, several potential sources of bias exist that threaten the validity of findings. The problems induced by lack of randomization are not new, nor are the analytical strategies used to strengthen conclusions from observational studies. Nonetheless, more appropriate use and consistent reporting of these analytical strategies should be adopted in cardiology outcomes research. A series of 3 papers in the “Education and Debate” section of the British Medical Journal describe several practical questions to be asked by researchers when reading results from observational studies.6–8

Another often-ignored problem involves the structure of the data. Data are frequently clustered or “hierarchical” in nature, and this structure induces specific relationships among the data. Virtually all data have some hierarchical structure (eg, patients are clustered in hospitals), and ignoring the data structure will often lead to erroneous conclusions. Missing data are yet another challenge to researchers, occurring in both randomized and observational studies. Despite availability of statistical tools to handle missing data, unsupported methods continued to be used. Finally, an emerging issue relates to the increasing use of multiple outcomes and multiple informants in outcomes research.9 Investigators collect and assess multiple outcomes to comprehensively assess a treatment or policy effect but continue to use ad hoc pooling strategies to reach their conclusions.

Absence of Randomization

Although sometimes not stated explicitly, the most common goal in outcomes research involves the establishment of causation. For example, do drug-eluting stents (DES) cause excess mortality compared with bare-metal stents (BMS)? Does early catheterization in unstable angina patients lead to better in-hospital outcomes? Does invasive cardiac manage-
ment increase survival after acute myocardial infarction (AMI)? Causal inference focuses on what would happen to a specific individual under different treatment options. In contrast, predictive inference focuses on the comparison of outcomes between groups of individuals who have received different treatments. Causal inference can be thought of as a special case of predictive inference in which subjects who could have received either treatment are identified and used to infer treatment effects (eg, what would have happened to a patient’s survival had the patient received a different treatment than the one observed?).

Specific features of a randomized clinical trial10 permit researchers to conclude whether a treatment or intervention is efficacious. First, the experimenter determines the assignment of treatments to patients using a known mechanism. This mechanism is the randomization allocation probability determined by the experimenter and implemented with standard software. Treatment allocation may correspond to equal allocation between treatment arms for all trial patients or equal allocation between treatment arms within important patient groups, such as diabetic and nondiabetic patients. Allocation probabilities can be fixed, or they can be adaptive procedures11 that change as the study progresses. The outcome of the random assignment (eg, subject is assigned to treatment A) has several key properties, 2 of which include (1) that it is predictive of treatment taken, so that if we were to model the probability of treatment received as a function of treatment assignment, the odds ratio of the treatment assignment variable would be large, and (2) that treatment assignment is not related to outcome if we take into account the treatment received. This implies that it is the treatment received that causes a change in patient outcome and not the treatment assigned. A variable with these properties is also referred to as an “instrumental variable.”12,13

A second characteristic of a randomized trial is a surprisingly simple fact: Every subject who meets the study inclusion criteria has a chance of receiving the treatment. This implies that the probability that a trial participant receives the study treatment is always greater than zero. This is due both to study inclusion/exclusion criteria that are developed to define the target population and to experimental control over who receives the study treatment. This seemingly trivial point is often ignored in observational studies.

The third feature is that, in theory, no unmeasured or measured variables (denoted confounders) are present that relate to both treatment assignment and outcome. Statistically, this implies that the “potential” outcome and treatment assignment are independent given the patient covariates. This means that because participants have been allocated randomly to treatment groups, and each participant had a chance of receiving treatment, the only difference between the treatment groups is treatment assignment. The standard estimate of the treatment effect is the intention-to-treat estimate, in which the average outcome of those assigned to treatment is subtracted from the average outcome of those assigned to the comparison treatment. The intention-to-treat estimate is only valid under the assumption of full treatment compliance and no missing data,14 and very few studies meet these criteria. Randomized studies have additional important features, such as blinding, which will not be discussed here.

**No Unmeasured Confounders**

Observational studies generally fail to meet most of the assumptions required to support a causal conclusion (Table 1). If no unmeasured variables exist that confound the relationship between treatment assignment and the outcome, the assignment mechanism is said to be “ignorable.” The basis for this term is that the investigator can “ignore” the treatment assignment as long as the observed confounders are used to adjust outcome comparisons. For causal inferences in the ignorable setting, approaches to data analysis fall into 3 general categories. The most common is a regression model in which the confounders and treatment received are regressed on the outcome, and adjusted outcomes are estimated. Numerous statistical packages are available that researchers can use. A second approach is through matching or stratifying of subjects, by which categories formed by the set of confounders are created that contain both treated and comparison subjects. Outcome differences within each category are computed and then combined to form an overall estimated effect. The third approach is a combination of the first 2 in which the set of confounders is reduced to a single balancing score, often into a propensity score,15 and outcomes are examined within groups defined by the score.

Each approach is associated with specific assumptions. In addition to the usual distributional and independence assumptions associated with regression modeling, 2 additional assumptions need assessment: (1) determination of sufficient overlap of the measured confounders to permit sensible estimation of treatment effects and (2) determination of similar “distributions” of confounders so that conclusions do not depend on the distributional assumptions made by the regression model. In fact, regression modeling can perform poorly when the variances of the confounders between treatment groups are unequal, which is very often the case in observational studies. Matching or stratifying on the confounders is sensible but becomes difficult when the number of confounders is large. With 20 confounders, each assuming 2 categories, there will be approximately 1 million matching categories. The combination of approaches offers a practical alternative and facilitates assessment of the plausibility of statistical assumptions.

**Illustrative Example: Long-Term Clinical Outcomes After DES and BMS in Massachusetts**

Mauri and colleagues16 compared all-cause mortality, revascularization, and myocardial infarction rates between 11 516 patients undergoing DES implantation and 6210 patients undergoing BMS implantation between April 1, 2003, and September 30, 2004. Patients were not randomized to stent type and differed on several observed confounders. To compare average differences between the groups in each confounder on a common scale, percent standardized differences in mean values (difference in mean values divided by the pooled SDs for the DES and BMS groups) for a number of important confounders were computed (Figure 1). Large differences were observed for commercial health insurance,
acute coronary syndrome (ACS), status of the procedure, and ejection fraction.

To assess comparability of the distributions of the observed confounders, plots of the relative frequency of values for each confounder (a density estimate) for the BMS and DES groups were constructed. Figure 2 presents density estimates for 3 (of a total of 65) confounders and for the log-odds of the estimated propensity scores. ACS for more than 1 day (yes versus no) was transformed by subtracting the mean value of the entire cohort and dividing by its SD so that positive values indicated a higher than average chance of having ACS for more than 1 day and negative values indicated the opposite. The ACS distributions had different means but the same shapes (Figure 2A); the days on market distributions had different means and different skews for the 2 groups (Figure 2B), but the age distributions appear comparable (Figure 2C). When all confounders were considered simultaneously, the propensity score distributions (Figure 2D) had different means and different skews for the DES and BMS groups. Figure 2D also demonstrates a lack of overlap in the left tail between the 2 groups, which suggests that there may have been no DES patients “comparable” to BMS patients for these propensity scores. Figure 3 suggests greater comparability of the distributions when a matched sample of 3752 BMS and DES pairs created from the estimated propensity scores was used. For this subset of patients, the assumptions for comparability on measured confounders were met.

**Unmeasured Confounders**

If insufficient measured confounders are present to adequately capture the selection of treatments and outcomes, then the investigator cannot ignore the treatment-assignment mechanism. In this case, the treatment assignment is said to be “nonignorable.” Determination of whether the treatment assignment is ignorable is based on the clinical problem and the richness of the measured variables. One method available to researchers when treatment assignment is nonignorable is through the use of instrumental variables. The estimated treatment effect is loosely calculated as the difference in mean outcomes between treatment groups divided by the difference in treatment assignment predicted by the instrument between the 2 groups. An estimate that uses an instrument that is weakly associated with treatment, eg, one that does not predict treatment assignment well, may give misleading results. Several assumptions exist that need to be satisfied when an instrumental variables analysis is used (Table 2), as well as different approaches to estimating the treatment effect.17 Some popular software packages include instrumental variables analysis such as Stata ivreg and ivprobit (Stata Corp, College Station, Tex) or SAS Proc SYSLIN (SAS Institute Inc, Cary, NC).

**Illustrative Example: Early Versus Late Catheterization in Unstable Angina or ST-Elevation Myocardial Infarction Patients**

Ryan et al18 compared the effects of early versus late use of catheterization on death, reinfarction, stroke, cardiogenic
shock, and congestive heart failure using an observational cohort of patients treated at 310 US hospitals. The authors used weekday (7:01 AM Sunday through 4:59 PM Friday) versus weekend (5 PM Friday through 7 AM Sunday) presentation as an instrumental variable for early catheterization. They observed 45,548 patients who presented on a weekday and 10,804 who presented on a weekend. The median times to catheterization and to percutaneous coronary intervention were 23.4 hours and 22.6 hours for the weekday group and 46.3 and 44.5 hours for the weekend group, respectively. This observation supports the assumption that weekday is predictive of who receives early catheterization. Table 2 provides justification for each of the instrumental variable assumptions in this particular example.

Illustrative Example: Effects of Invasive Cardiac Management on Survival After AMI

The report by Stukel and colleagues provides an excellent application of propensity scores and instrumental variable analysis to determine the effects of invasive cardiac management on survival after AMI. The authors implemented a propensity-based matched analysis by first modeling receipt of cardiac catheterization as a function of patient, hospital, and area-level covariates, selecting pairs of patients with similar probabilities of undergoing catheterization in which 1 member of the pair received the procedure and the other member did not, and then estimated the relative risk of mortality in these matched pairs using Cox regression. The authors found a 50% relative reduction in mortality risk for catheterized patients using Cox regression. They next used regional cardiac catheterization rates as instrumental variables to conduct an instrumental variable analysis. The mean cardiac catheterization rates varied from 42.8% to 65% across regions, with corresponding 4-year mortality rates of 43.1% to 38.9% (authors’ Table 4). A crude instrumental variable estimate is thus \( \frac{43.1 - 38.9}{65 - 42.8} \times 100\% = 55\% \) absolute mortality reduction and is interpreted as, “If we increased the cardiac catheterization rates by 22%, we would observe an 18.9% reduction in 4-year mortality.” Using the Stata function ivreg, the authors calculated an estimate of a 16% relative or 9.7% absolute (authors’ Table 5) mortality reduction at 4 years.

Figure 1. Percent standardized differences between DES and BMS patients for selected prestent characteristics stratified by type of characteristic. Open circles (○) denote mean standardized difference for 17,726 patients; filled squares (■) denote mean standardized differences for 3,752 matched pairs. HMO indicates health maintenance organization; NYHA, New York Heart Association; CCS, Canadian Classification System; EF, ejection fraction; LCX, left circumflex; LM, left main; and RCA, right coronary artery.

Pre-Stent Differences

Demographics

<table>
<thead>
<tr>
<th>Demographic</th>
<th>Larger for BMS</th>
<th>Larger for DES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yrs Over 65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td></td>
<td></td>
</tr>
<tr>
<td>African American</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other or Missing Race</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Commercial Insurance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HMO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Insurance</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Presentation

<table>
<thead>
<tr>
<th>Presentation</th>
<th>Larger for BMS</th>
<th>Larger for DES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emergent or</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Salvage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Urgent</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACS 6-24 hrs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACS &gt; 24 hrs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACS &lt; 6 hrs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ve Stress Test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-ve Stress Test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NYHA I or II</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NYHA III or IV</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCS 1 or 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCS 3 or 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EF Measured</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EF &lt; 30%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Lesion Characteristics

<table>
<thead>
<tr>
<th>Lesion Characteristic</th>
<th>Larger for BMS</th>
<th>Larger for DES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Vessel</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multivessel</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LM Disease</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 Treated Lesions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 Treated Lesions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 Treated Lesions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LCX Treated</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LM Treated</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RCA Treated</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High Risk Lesion</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Medications

<table>
<thead>
<tr>
<th>Medication</th>
<th>Larger for BMS</th>
<th>Larger for DES</th>
</tr>
</thead>
<tbody>
<tr>
<td>IIB/IIIA Contraindicated</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IIB/IIIA Given</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clopidogrel Contraindicated</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clopidogrel Given</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thrombolytics Contraindicated</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thrombolytics Given</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Percent standardized differences between DES and BMS patients for selected prestent characteristics stratified by type of characteristic. Open circles (○) denote mean standardized difference for 17,726 patients; filled squares (■) denote mean standardized differences for 3,752 matched pairs. HMO indicates health maintenance organization; NYHA, New York Heart Association; CCS, Canadian Classification System; EF, ejection fraction; LCX, left circumflex; LM, left main; and RCA, right coronary artery.
How do we interpret these results? Both estimates indicate a benefit of invasive cardiac management, but the size of the benefit differs (50% versus 16% relative reduction). Three explanations are possible: Both estimates are wrong, only 1 estimate is wrong, or they are estimating different treatment effects. If both estimators estimate the average treatment effect, then if no residual confounding was present and the treatment benefit was constant across patient groups, both estimates should agree. The authors used a linear regression model for their instrumental variable analysis, which assumes the treatment effect is additive. This implies that the instrumental variable estimate of the average treatment effect is measured by the absolute, not the relative, difference. The 4-year absolute mortality benefits are not comparable: 19.1% for the propensity-based matched pairs (authors’ Table 1, difference in mortality rates reported in the last row between matched pairs: 55.4% – 36.3%) and 9.7% for the instrumental variable analysis (authors’ Table 5). If a constant treatment effect is found across different risk groups of patients, such a difference implies that the propensity score estimate is biased. Is there evidence of a nonconstant treatment effect? It appears so: The authors reported predicted absolute 1-year mortality benefit ranging from 3.3% in the lowest propensity score decile to 0.8% in the highest decile. The author’s Table 4 also suggests a nonconstant treatment effect across regions. Because the propensity-score–matched estimate and the instrumental variable estimate use different subsets of the original sample, there is no guarantee that these subsets are the same. Thus, there is no reason to expect the estimates to be the same; therefore, both could be correct but targeted to 2 different subpopulations.

Sensitivity to Unmeasured Confounders
Sensitivity analyses are another underused and powerful tool. Rosenbaum proposed an elegant approach to quantify how study conclusions would be changed by unmeasured confounding. The key idea involves the creation of a confounding measure that quantifies the degree of unmeasured confounding and the use of plausible values of the measure to calculate new P values to quantify how study conclusions would change. The confounding measure is the ratio of odds that 2 patients with identical observed confounders receive the treatment. When this OR is 1, the study is free of unmeasured confounders; when it is larger than 1 (say 2), then 2 patients who appear similar on the measured confounders could differ in their odds of receiving treatment by as much as 2. Rosenbaum developed several formulas for bounds in P values for common test statistics such as the Wilcoxon signed rank statistic and McNemar test statistic. If the study findings remain statistically significant for several plausible values of the OR, the investigator may conclude the study is insensitive to hidden confounders.

Three comments are in order. First, there will always be a value of the OR at which the P value changes from statistically significant to not statistically significant; at this value, the investigator would conclude that unobserved confounders could explain the observed association between the treatment and the outcome. Second, the sensitivity analysis requires that the investigator a priori specify plausible values for the

Figure 2. Distribution of selected covariates for DES and BMS patients (17 726 patients). X axis is the value of the covariate, and y axis is a density estimate that reflects the proportion of the DES (dashed lines) and BMS (solid lines) groups having the particular values of the covariate.
confounding measure (the OR). The ORs selected should be problem-specific and will depend on the type and number of measured confounders already included in the model. For example, if only age and sex are used to adjust for treatment selection and risk of reinfarction, then large values of the OR are plausible. On the other hand, if in addition to demographic variables, the investigator includes signs and symptoms on presentation, such as cardiovascular history, preprocedure variables, and contraindications to medications, then large values of the OR are not as plausible. Third, the fact that the study results are insensitive does not mean that no unmeasured confounders exist.

Illustrative Example: Validation of Catheterization Guidelines for AMI Patients

An example of sensitivity analyses can be found in a study validating coronary angiography guidelines for AMI patients. The authors used propensity-score matching with 105 clinical variables obtained in ~20,000 Medicare beneficiaries. Their goal was to estimate the benefit of coronary angiography for patients, which was classified as clinically necessary or clinically appropriate or for which uncertainty of the clinical benefit was present. They found an absolute 3-year survival benefit of 17.6% (95% confidence interval 15.1% to 20.1%) in patients undergoing clinically necessary angiography and a smaller benefit (8.8%; 95% confidence interval 6.8% to 10.7%) in patients for whom the benefit was uncertain. The variation in survival benefits suggests a nonconstant treatment effect.

The authors determined that to eliminate the survival benefit in patients for whom the procedure was judged necessary, an unmeasured confounder not related to the 105 observed confounders already included in the model would have to increase the odds of angiography by more than 2. The authors also compared 2-day mortality for the matched pairs assuming any clinically meaningful difference would indicate the presence of residual confounding. They observed a small benefit of 1.5% (95% confidence interval 1.0% to 2.0%) in patients for whom the procedure was necessary. Finally, the authors found a benefit regardless of the hospital’s capability to perform coronary angiography.

Clustered Data

Data are clustered when some units are nested completely within other units. Models to deal with these types of data go by many names: hierarchical models, multilevel models, random-effects models, mixed models, random coefficient models, subject-specific models, and empirical Bayes models. Common outcomes research examples include patients nested within hospitals, patients nested within health plans, and patients nested with surgeons. A common example of nesting involves longitudinal data for which the measurement occasion is nested within the subject, as would be the case when one measures health status on 4 occasions after a cardiac event: baseline, 30 days, 6 months, and 1 year. Clustered data are not unique to health outcomes research; for example, educational researchers often deal with data collected from students who are nested within classrooms or within teachers. Finally, the levels of clustering can be >2, as is the case if longitudinal measures are taken for patients.
Table 2. Assumptions Needed When Conducting an Instrumental Variables Analysis

<table>
<thead>
<tr>
<th>Assumption</th>
<th>Example: Instrument—Weekend vs Weekday Hospital Presentation; Treatment—Early Catheterization; Outcome—Death, Myocardial Infarction, Stroke, Shock, or Congestive Heart Failure</th>
</tr>
</thead>
<tbody>
<tr>
<td>The instrument is not related to unmeasured confounders; cannot be tested empirically</td>
<td>Investigators used detailed data collected as part of a registry that used common definitions. Means of observed covariates and use of in-hospital medications (β-blockers, etc) appeared similar between the weekend-presenting patients and weekday-presenting patients.</td>
</tr>
<tr>
<td>(1) The relationship between the instrument and treatment received is nonzero (a powerful predictor of treatment received) and (2) the instrument is related to outcome only through treatment</td>
<td>(1) Median time to catheterization was 23 h for weekday patients and 46 h for weekend patients. Regressed weekday presentation, patient characteristics, and hospital characteristics on early catheterization; weekday presentation was statistically predictive of early catheterization. (2) No statistically significant relationship between weekday presentation and in-hospital outcome after adjustment for patient characteristics, hospital characteristics, and early catheterization.</td>
</tr>
<tr>
<td>There are no patients who would receive the treatment if the instrument was false but who would not receive the treatment if the instrument was positive.</td>
<td>Need to assume that there are no patients who would receive early catheterization if they presented on a weekend but who would not receive early catheterization if they presented on a weekday. This appears reasonable.</td>
</tr>
<tr>
<td>For patients whose treatment received would not have been changed by the instrument, there is no effect of the instrument on the outcome.</td>
<td>Investigators would need to argue that for patients who would never undergo early catheterization, weekend presentation has no effect on in-hospital outcomes. Similarly, for patients who would always undergo early catheterization, weekend presentation has no effect on in-hospital outcomes. This assumption appears plausible.</td>
</tr>
<tr>
<td>Treatment assignment for 1 patient does not affect the outcome of another patient.</td>
<td>This assumption would be violated if the likelihood for a patient to undergo early catheterization is related to the survival of another patient. This assumption is reasonable for 2 patients presenting at 2 different hospitals. This assumption could be violated for 2 patients treated within the same hospital if the mortality experience of patients treated in the hospital influenced the likelihood of early angiography for other patients within the hospital (which appears unlikely).</td>
</tr>
</tbody>
</table>

Applied to study by Ryan et al. Confounders are variables related to the probability of treatment assignment and to the outcomes of interest.

treated within hospitals. Here, 3 clusters are present: occasion nested in patient nested in hospital.

Clustered data complicate an analysis primarily because there is no reason to believe that observations within a cluster are statistically independent. Consider the classic example of assessing the yield of different insecticides, in which each insecticide is sprayed on a different tree. All leaves on the tree are exposed to the same insecticide, so that we would expect that the health of 2 leaves sampled from the same tree would be more alike than the health of 2 leaves sampled from 2 different trees. What do leaves and insecticides have to do with cardiac outcomes research? With many outcomes studies, interventions are assessed for patients who are treated within practice settings. When assessing healthcare quality provided within ambulatory care settings, all patients within the practice are exposed to the same level of quality, so that we would expect that the likelihood of getting guideline care treatments for 2 patients sampled from the same practice would be more alike than the likelihood for 2 patients sampled from 2 different practices.

Several naïve approaches have been used to analyze clustered data; these involve ignoring the clustering entirely and estimating a usual regression model, with the inclusion of a dummy variable for each cluster (eg, a dummy variable for each practice setting) in the regression model, or performing an ecological regression in which cluster characteristics and mean patient characteristics are regressed on the mean practice outcome. Not properly accounting for the structure of the data can often lead to incorrect conclusions. For example, ignoring the hierarchical structure altogether can lead to the wrong standard errors associated with the covariate effects, and the within- and between-cluster effects would be confounded. This implies, for example, that we cannot determine whether healthy patients are treated better within each practice or whether good practices have healthy patients. Inclusion of dummy variables results in a loss of all between-practice information, because it is not possible to include practice covariates simultaneously with the dummy practice variables. Finally, the ecological regression results in ecological bias (eg, researchers incorrectly interpret “practices with younger patients have better outcomes” as “younger patients have better outcomes”).

With clustered data, the types and numbers of parameters are increased. Interest could center on covariate effects for which the main parameters of interest are the regression coefficients that describe the association of patient or practice-setting characteristics with the outcome. Investigators may be interested in the “random effects,” such as the cluster-specific risk-adjusted outcomes. Finally, although not common, interest may focus on estimation of the variance components, such as the between-practice setting variance after adjustment for patient and practice-level covariates. The appropriate statistical model will depend on the primary goal of the study. A marginal approach is adopted when the main goal involves estimation of covariate effects. With a marginal approach, the investigator specifies covariates that are thought to be associated with the outcome and selects a covariance pattern to account for correlation of the observa-
tions. For example, the investigator may assume that the correlations between outcomes of patients within the same cluster are equal. The between-cluster variation is simply a nuisance that must be accommodated to make statistically valid conclusions about the effects of risk factors. Marginal models can be estimated with Proc GENMOD in SAS and xtgee in Stata. On the other hand, a hierarchical model is adopted when the main goal involves the estimation of variance components or cluster effects, perhaps in addition to covariate effects. With this approach, the investigator specifies covariates thought to be associated with the outcome but also specifies random effects for the outcome. The random effects are assumed to vary from patient to patient according to a probability distribution. The introduction of the random effects induces a particular covariance structure for the observations. Proc MIXED and Proc NLMIXED in SAS and the Stata functions xtmixed, glamm, and xtreg can be used to estimate hierarchical models.

Interpretation of the covariate effects differs between the 2 approaches. The regression parameters in a marginal model describe the association of patient-level covariates with changes in the population mean outcome, whereas the parameters in a hierarchical model describe the association of patient-level covariates with changes in a patient’s outcome. Although the difference is subtle, it does exist. Coefficients obtained from a marginal model are smaller in absolute value than those from a hierarchical model, and as the between-cluster variance increases (for example, larger practice differences), the discrepancy between regression coefficients increases.

Random effects are interpreted as unobserved variables that are shared by (and therefore induce some dependence among) the outcomes for patients within the same cluster. For example, in health quality studies, it is often reasonable to assume that a shared latent variable (eg, underlying quality of care) influences patients’ outcomes, even after adjustment for observed confounders. A similar assumption is present when one combines studies in a meta-analysis or when one combines measurement occasions within an individual in a longitudinal study. Both hospital-specific risk-adjusted mortality rates used in hospital profiling and subject-specific change rates used in growth-curve analyses are random effects. Hierarchical models, not marginal models, permit estimation of the random effects. This is accomplished with the use of information from patients within a particular cluster but also by borrowing information from other clusters. The resulting estimates are often superior to those that rely only on a cluster’s own data.

Finally, the variance components of hierarchical models may themselves be of primary interest. Investigators may wish to decompose the variance attributed to each cluster, such as that due to hospital and that due to patient, or transform it into the intraclass correlation coefficient, which is interpreted as the fraction of variance due to the cluster (eg, due to the hospital). The overall goal here involves describing how much of the variation is due to the specific clusters (eg, due to random error, due to doctor, and due to hospital). See Larsen and Merlo for practical guidance on interpreting variance components in logistic hierarchical models.

**Illustrative Examples**

Roe and colleagues examined the association of treatment by inpatient cardiac specialty service versus noncardiac service specialty on mortality for patients with non-ST-segment elevation ACS. They used data from patients admitted to hospitals participating in the CRUSADE (Can Rapid Risk Stratification of Unstable Angina Patients Suppress Adverse Outcomes With Early Implementation of the ACC/AHA Guidelines) study. Because their focus was on the regression coefficient associated with the inpatient cardiac specialty service indicator variable after adjustment for patient characteristics, they estimated a marginal model to account for clustering of patients within hospitals. On the other hand, Krumholz and colleagues focused on estimating hospital-specific risk-adjusted mortality rates for AMI patients. As in the study by Roe et al, patients were clustered in hospitals; however, the primary focus in the Krumholz study was on the random effects that represented all-cause 30-day mortality adjusted for patient characteristics. Hofer and colleagues focused on the variance components when modeling, among other outcomes, level of glycemic control for diabetic patients clustered within practices. Specifically, the authors estimated the different sources of variation and concluded that <4% of the total variance was attributable to between-practice variation.

**Missing Data**

Despite the best data collection efforts, not all planned observations are made. The reasons for missing data, covariates and outcomes alike, are as numerous as they are varied. Subjects may have missed a visit for a practical or administrative reason; data may not have been collected on a particular day because of equipment failure; or a subset of data may have been lost. Alternatively, subjects may be missing data because side effects associated with the treatment prohibited them from continued study participation, or they may not report their health status outcome because they are too sick.

According to the taxonomy introduced by Rubin, the underlying process to generate missing data falls into 1 of 2 classes: ignorable or nonignorable, similar to the distinction made relative to treatment assignment. Missing data are ignorable if all the characteristics that are associated with the probability of missingness are observable and statistically adjustable. The mechanism is ignorable because sufficient information is available for us to “ignore” the missing-data mechanism. In the simplest case, if the probability of data being missing is not associated with any observed or unobserved variable, whether covariate or outcome, then the mechanism is missing completely at random (MCAR). Data MCAR would arise if some data were not collected on a particular day because of equipment failure, for example. If the probability that a variable has missing data depends only on collected information, then the mechanism is missing at random (MAR). It may be that older patients are more likely...
Horton and Kleinman provide a summary of procedures using GENMOD, and Proc NLMIXED in SAS. When MAR or MCAR: Proc MIANALYZE, Proc MI, Proc software packages provide methods to handle missing data. Researchers can adjust for the detailed information or impute bolstered through collection of detailed information so that the number of observations with complete data so that the power to detect differences is smaller. When data are MAR, it is permissible to exclude the missing observations provided that a regression model controls for all the variables that affect the probability of missingness. The assumption of MAR can be bolstered through collection of detailed information so that researchers can adjust for the detailed information or impute the missing data using the collected information. Several software packages provide methods to handle missing data when MAR or MCAR: Proc MIANALYZE, Proc MI, Proc GENMOD, and Proc NLMIXED in SAS.

If missingness is NMAR, then the investigator needs to model the missingness mechanism, accept the bias that may exist, or implement an MAR analysis. Three strategies are available to deal with missing data: selection models, pattern-mixture models, and shared-parameter models. With selection models, the researcher models the probability that an observation is missing on the basis of the observed and unobserved data. The parameters in the missing-data model can be estimated, but the resulting estimates are very sensitive with regard to how the missing-data model is specified. In pattern-mixture models, the researcher permits a different outcome model for each pattern of missing values. The assumption is that the observed data are a mixture of the different outcome models, with weights equal to the probability of each missing-data pattern. For example, in a longitudinal study in which patient reporting of angina pain is planned on 3 occasions, there may be 2 patterns of missing data: patients missing only the third outcome and patients missing angina reports at times 2 and 3, and for those missing angina reports at time 3 only. In the shared-parameter model, the researcher assumes the existence of a subject-specific latent variable that affects both the outcome measurement and the missingness probability. In the pain example, the latent variable would represent the true underlying level of pain, and the researcher would need to assume that the missingness process depends on the true level of angina perceived by a patient and that this true level may vary over time.

Some points are worth emphasizing. First, an investigator can never use data to prove that the missing-data mechanism is ignorable. It is possible to differentiate between MCAR and MAR in many situations. If meaningful differences exist between those with and without missing data for some variables, this provides evidence against MCAR. An investigator rarely has sufficient data to prove that the missing-data mechanism is NMAR, because the nonignorability assumption asserts something that is not available to the researcher. Second, simple methods for handling missing data are often associated with implausible assumptions, a fact that is especially ironic given common complaints by researchers that...
the more valid statistical procedures are “too complicated.” Use of the last observation carried forward, for example, requires the researcher to believe that the patient’s measurements remain at the same level during the period they are unobserved. Third, researchers should not ignore the problem of missing data if they have a high response rate, because even a small fraction of missing data (eg, 5%), can lead to incorrect conclusions depending on the missingness mechanism.

**Illustrative Examples: Ignorable Missing Data**

Researchers analyzing survey data typically impute data, often because large survey data sets normally will be made public; they then release the multiply imputed data sets to ensure that users treat missing data in a statistically valid manner. For example, the Consumer Assessments of Healthcare Providers and Systems (CAHPS) uses multiple imputation to fill in missing values. On the other hand, researchers using registry data have also implemented imputation strategies. For example, Vaccarino and colleagues imputed data from the National Cardiovascular Network to determine whether sex differences existed in hospital mortality after CABG surgery. In an analysis using retrospectively collected clinical data for AMI patients, Krumholz and colleagues implemented the popular approach of including for each continuous predictor with missing values an additional indicator variable that identifies which patients have missing data on that variable and adding an extra missingness category for categorical predictors with missing data. For example, if blood urea nitrogen was measured within 24 hours of hospital admission, its value would be entered into the model for the patient, and the missingness indicator would have a value of 0; if this value was missing for the patient, then the indicator value for missingness would have a value of 1, and the blood urea nitrogen value would have a value of 0. This imputation strategy may yield biased estimates of regression coefficients for other predictors included in the model, because it is assumed the relationship between the other predictors and the outcome is the same for those with and without blood urea nitrogen values. Because the goal of the Krumholz study was to compare hospital-specific risk-adjusted mortality estimates based on administrative claims data and on clinical data, and not to focus estimation on any particular predictor, this method is acceptable.

**Nonignorable Missing Data**

Nonignorable missing-data models have not yet become commonplace in cardiology outcomes research literature and have been applied only to a few clinical problems. Magder illustrates a sensitivity analysis that assumes different mechanisms of nonignorable missing outcome data.

**Multiple Outcomes or Multiple Informants**

The inclusion of multiple cardiac outcomes is an increasingly frequent strategy in both clinical trials and observational studies. The desire to include more than 1 outcome arises because a single outcome may not adequately describe the disease, there may be a lack of consensus on the most important clinical outcomes, there may be a desire to demonstrate clinical effectiveness for several outcomes, or researchers wish to increase their statistical power to detect a treatment effect by increasing the number of events. All-cause mortality, myocardial infarction, and repeat revascularization are often combined to assess the success of revascularization strategies, for example.

Equally increasing is the use of multiple data sources or “informants” to report on the same underlying outcome. The use of multiple informants has a long history in psychiatric research, in which multiple informants (eg, teacher, child, and parent) are used to assess the effectiveness of behavioral health interventions. Multiple informants are used in family history studies, in which several relatives may be questioned about the status of the proband and other family members. Multiple-informant reports also arise in the process of gathering information about healthcare service utilization and quality when information is obtained from both users of medical services and their providers. The use of multiple informants or multiple outcomes is neither new nor unique to cardiology outcomes research, but it is increasing (for more background and related literature, see www.biostat.harvard.edu/multiform).

Two types of multiple outcomes exist. Commensurate outcomes are multiple outcomes that measure the same underlying construct using the same scale, whereas noncommensurate outcomes neither measure the same underlying construct nor are measured on the same scale. Examples of commensurate outcomes include conformance on process-of-care indicators for an AMI patient (eg, receipt of aspirin or of β-blockers), multiple items that characterize patient limitations in the Seattle Angina Questionnaire, and multiple outcomes that define restenosis. Noncommensurate outcomes include length of stay and readmission, because they are measured on different scales.

The most common strategies used with multiple outcomes or informants are to pool the outcomes or to analyze each outcome separately. When pooling outcomes, the resulting measure is called a composite end point. Composites are most often created by implementing an “or rule,” referred to as a compensatory scoring algorithm, or an “and rule,” referred to as a conjunctive scoring algorithm. The extensive literature describing desirable attributes of composite end points (see, for example, Moyer), such as requiring that associations between the treatment and each outcome be in the same direction, will not be reviewed here. Another strategy used by researchers involves analyzing each outcome separately and then drawing some conclusions about the intervention.

Similar pooling strategies have been used in handling multiple-informant reports in which a condition is classified as present when reported by at least 1 informant. For example, if either the patient or the family member indicates the presence of angina, then angina is endorsed. With this approach, emphasis is placed on minimizing false-negatives, which yields prevalence estimates that may be too high.

Although a compensatory or conjunctive pooling strategy is simple to describe and implement, it has several limitations. The optimal method for combining outcomes depends
on the type of measurement error present. Pooling does not permit differences in treatment effects across outcomes, and it becomes difficult if the scale of measurement is different among the outcomes. Often, continuously scaled outcomes are converted to binary outcomes for pooling with other outcomes. Finally, the pooling algorithms are typically not defined when outcomes or informants are missing. Separate analysis of each source is also problematic. Separate analyses yield findings that are associated with each outcome without a formal statistical method to generalize the conclusions, and they may be based on different subsets of patients if some patients are missing 1 outcome and others are missing another.

Several statistical approaches are now available to researchers to analyze multiple outcomes or informants. With multiple informants, researchers can jointly model the associations between informant and outcome using a marginal modeling approach available in standard software packages. In the multiple-informant setting, a single outcome exists (eg, patient was hospitalized), but multiple informants provide information about the same predictors (eg, patient and family member reports of angina pain). Joint modeling means that the investigator specifies a separate model for the single outcome for each informant; the equations are linked through the variance terms, and so regression coefficients and other model parameters are estimated simultaneously. Joint modeling permits statistical tests to determine whether a difference exists in associations between the patient and family member reports of angina on risk of hospitalization, as well as tests to determine whether the associations between other patient factors in the model and hospital admission risk differ by informant. If no statistically significant informant effects are found, a pooled model can be estimated that will have smaller standard errors, so that the power for finding an effect is increased.

Several strategies also exist for joint modeling of multiple outcomes. As with clustered data, a marginal model can be used in which the investigator specifies covariates believed to be associated with the outcomes and selects a covariance pattern to account for correlation of the observations. Alternatively, the investigator could assume the multiple outcomes are observed manifestations of a latent variable. For example, conformance with multiple AMI quality indicators may be thought of as manifestations of the underlying quality of care provided at a particular hospital. The advantages of joint modeling of binary commensurate outcomes relate to those observed when performing “seemingly unrelated regressions”:43 The standard errors of covariate effects are smaller than those obtained when the outcomes are modeled separately, and if investigators anticipate that different covariates may be associated with different outcomes, serious reductions in the standard error of the nonshared covariates can be accomplished by use of a joint modeling approach. These findings hold even when one considers noncommensurate outcomes, eg, in joint modeling of binary target-lesion revascularization and proportion diameter stenosis in patients undergoing revascularization.

Illustrative Examples

Numerous examples are available in the cardiology outcomes research literature involving the use of composite end points. Occurrence of a major adverse cardiac event is a common composite outcome that has been used to assess revascularization strategies in real-world settings. A conjunctive scoring algorithm used to create the “all-or-none” composite measure for assessment of healthcare quality has also been proposed.45 The outcomes research literature is also abundant with separate analyses of multiple outcomes. Joint modeling of multiple outcomes is new to clinical research, and therefore, few clinical examples are available. In a health services research application, Landrum and colleagues modeled multiple binary quality measures for AMI patients to assess hospital quality of care; Timbie and Normand simultaneously modeled survival and costs for AMI patients and compared their methods with standard approaches.

Fewer applications of multiple-informant analyses exist outside of the mental health literature. However, Lash and colleagues evaluated the relationship between several comorbidity indices and receipt of tamoxifen therapy in a cohort of breast cancer patients. The different comorbidity indices represented the multiple informants. O’Malley and colleagues evaluated the association of the culture of a community health center and the center’s participation in quality-improvement initiatives, where culture was assessed by several informants (executive director of the center, medical director, deputy director, and a center provider).

Conclusions

Statistical modeling involves making some compromises. Cardiology outcomes researchers will not always need to impute missing data, account for clustering, use propensity scores or find instruments, or undertake a multiple-informant analysis. The overall goal of the study and the structure of the data should guide outcomes researchers to an appropriate statistical model. Is the goal predictive or causal? Are multiple outcomes used to assess an overall effect, or is 1 outcome truly primary? Is the goal to estimate hospital-specific summaries, or is it to characterize associations between covariates and outcome?

Although no model will always be correct, some models are clearly wrong. The goal of cardiology outcomes research is to design good studies, whether observational or randomized. To that end, researchers need to understand and sometimes estimate the treatment-assignment process, to understand and sometimes estimate the missing-data process, and to sometimes explicitly model the structure of the data. Although the present report discusses the 4 statistical problems separately, they can and do occur together. For this reason, ad hoc methods are much more problematic. Cardiology researchers should capitalize on the statistical advances and lessons learned from other fields, such as psychiatry, in which many of the same methodological problems are present.

Acknowledgments

I am indebted to several collaborators on the development of statistical tools for outcomes research: A. James O’Malley, PhD,
Richard G. Frank, PhD, and Mary Beth Landrum, PhD (Harvard Medical School, Boston, Mass) for causal modeling; Constantine A. Gatsonis, PhD (Brown University, Providence, RI) and Alan Zaslavsky, PhD (Harvard Medical School) for hierarchical modeling; and Garrett Fitzmaurice, ScD (Harvard Medical School and Harvard School of Public Health, Boston, Mass), Nicholas Horton, ScD (Smith College, Northampton, Mass), Stuart Lipsitz, ScD (Brigham and Women’s Hospital, Boston, Mass), and Nan Laird, PhD (Harvard School of Public Health) for multiple-outcomes and multiple-informant analyses.

Sources of Funding
This work was supported in part by grants R01-MH054693 and R01-MH061434, both from the National Institute of Mental Health, and by funding from the Massachusetts Department of Public Health (620022A4PRE).

Disclosures
None.

References
33. Heckman JJ. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. Am Econ Rev. 1976;5:475–492.


**Key Words:** data interpretation, statistical clustering causality outcomes assessment
Some Old and Some New Statistical Tools for Outcomes Research
Sharon-Lise T. Normand

Circulation. 2008;118:872-884
doi: 10.1161/CIRCULATIONAHA.108.766907
Circulation is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2008 American Heart Association, Inc. All rights reserved.
Print ISSN: 0009-7322. Online ISSN: 1524-4539

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://circ.ahajournals.org/content/118/8/872

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Circulation can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to Circulation is online at:
http://circ.ahajournals.org//subscriptions/