Editor’s Perspective

Missing Data and Convenient Assumptions

Sharon-Lise T. Normand, PhD

This issue contains the first of several planned statistical primers on methodological problems commonly encountered by outcomes researchers. Although several resources exist for the interested outcomes researcher, they are often scattered throughout different literatures and, in particular, are not translated to the “outcomes” setting. Much of the empirical basis of outcomes research involves observational data—a setting in which the researcher has no control over what interventions are given and why they are given—yet desires to understand the causes of the interventions. Many clinicians are trained to understand the results of clinical trials in which randomization is used to quantify the efficacy of medical treatments. However, outcomes research not only involves assessing the results of randomized trials to make clinical decisions but also assessing findings from observational studies of medical interventions and policies, financial arrangements, organizational systems, and combinations thereof. A solid understanding of the statistical methods used in outcomes research to aid in drawing conclusions is therefore required.

The objectives of the Statistical Primer Series are several-fold. A key goal is to introduce statistical techniques of particular relevance to observational data analyses. Second, the series aims to provide an enhanced understanding of analytical methods specifically within the context of outcomes research. In this setting, representative examples of common problems will be used to illustrate methods. Third, because of the natural complexity of the empirical data that forms the basis of outcomes research, the series will be used to introduce new statistical techniques. Thus, an additional goal of the series is to expose outcomes researchers to modern statistical techniques faster, with the hope of increasing their diffusion into the clinical literature.

Several methodological areas have been selected for the forthcoming articles in the Statistical Primer Series. Because of the predominant role causality plays in outcomes research, Dr Sue Marcus and colleagues (United States) describe foundational principles of causal inference. The key task in casual inference is to establish causation—to determine what would happen to a specific patient under different treatment options. Accomplishing this task requires a clear understanding of what constitutes a causal risk factor and after identifying such a risk factor, an approach to establishing with reasonable assurance its effect on patient outcomes in the observational setting. Outcomes researchers face particularly challenging tasks, not only because of the lack of randomization but also because of the added complexity attributed to the nature of the interventions. For example, what is the optimal duration of dual antiplatelet therapy after coronary stenting? This seemingly simple question is surprisingly daunting. In practice, patients may terminate therapy due to the occurrence of excessive bleeding or to the occurrence of a serious adverse event. In such settings, the duration of treatment is not observed but rather censored due to a treatment-terminating event. Alternatively, the patient may have their therapy interrupted for surgery. Landmark analyses, in which patients having treatment-censoring events before a “landmark” time are excluded from analysis, have become fashionable to address this type of problem. Dr Ourani Dafni (Greece) carefully describes the assumptions accompanying a landmark analysis and suggests other statistical tools to identify treatment-censoring events and appropriately account for them in modeling.

The explosion of registries has tempted outcomes researchers to examine more than 1 outcome for each subject. The tendency to collect and report on more than 1 outcome is not unique to observational studies—examination of contemporary clinical trials reveals a similar trend. Trialists have attempted to overcome the multiple end points problem through the creation of composites. Outcomes researchers have simply reported adjusted results for each outcome. Although the issue of overall error rate should not be ignored for those who are frequentists, a vexing problem relates to the impact that missing data have on the separate analyses and the validity of overall conclusion. How do we arrive at a decision when information is based on different sets of subjects within the same study? Dr Armando Teixeira-Pinto (Portugal) provides solutions to this question through an examination of new and old approaches to the multiple outcomes problem.

The inaugural topic presented in this issue is missing data. Why another article about missing data? The answer is simple: Despite enormous developments in general statistical methodology for missing data, these developments have had little impact on the way outcomes researchers routinely handle missing data. Too often, researchers use methods with assumptions motivated by convenience rather than by substance.1 It is far easier to eliminate observations with missing data than to use methods that account for missing values and the uncertainty they introduce. For outcomes researchers, incomplete data comprise a particularly vexing problem because of the need to “risk adjust”—to adjust for observed

1. The opinions expressed in this article are not necessarily those of the editors or of the American Heart Association.

From the Department of Health Care Policy, Harvard Medical School, and Department of Biostatistics, Harvard School of Public Health, Boston, Mass.

Correspondence to Sharon-Lise T. Normand, PhD, Department of Health Care Policy, Harvard Medical School, 180 Longwood Ave, Boston, MA 02115. E-mail Sharon@hcp.med.harvard.edu

(Circ Cardiovasc Qual Outcomes. 2010;3:2-3)

© 2010 American Heart Association, Inc.

Circ Cardiovasc Qual Outcomes is available at http://circoutcomes.ahajournals.org

DOI: 10.1161/CIRCOUTCOMES.109.931543
differences among intervention groups conditioning on many variables to avoid selection bias problems. However, when missing values occur on more than 1 variable, the incomplete cases will comprise a substantial fraction of the entire data set.

Missing data in longitudinal studies pose additional complexity. With increased focus on assessing the course and quality of care using patient-reported outcomes, researchers will observe the proportion of cases with missing data increasing over time due to subject dropout, mortality, or intervening events. Standard software packages for longitudinal modeling use procedures that can range from eliminating all subjects with missing information to retaining all subjects. The first option involves the convenient assumption that the probability of missingness is entirely unrelated to the value of the missing variable and also unrelated to any observed or unobserved values of covariates and outcomes. The second option involves another assumption—that the probability of missingness is unrelated to the unmeasured value once accounting for all other observed information. Although the second assumption is more plausible than the first, its plausibility depends on the particular problem.

What should outcomes researchers do in the presence of missing data? In most cases, ignoring the problem through case deletion will be unacceptable because of the multivariate context—outcomes researchers collect many variables so that the amount of missing information could be large. Statistically valid options exist, but each includes assumptions about why data are missing, that is, the nature of the missing data mechanism. In this issue, He (United States) describes the problem of missing data and illustrates the use of valid tools to deal with missingness. The main tool described is multiple imputation, a technique in which each missing value is replaced by more than 1 simulated value. The simulated values are obtained through regression techniques based all the observed data, outcomes, and covariates alike. Specific assumptions are made when implementing multiple imputation—they are, however, explicitly stated. Software packages such as SAS and Stata contain procedures that (1) permit researchers to impute the data and (2) combine the results from the separate data sets to make a single inference. He provides convincing scientific rationale as to why outcomes researchers must think about the missing data mechanism, and he provides a realistic example that will guide outcomes researchers in implementation of imputation methods.

Outcomes researchers can and should change the way missing data are routinely handled. This can be achieved primarily through careful examination of the plausibility of assumptions made and, more specifically, justifying convenient assumptions.

Disclosures

None.

References


Keywords: registries • statistics • imputation • models
Missing Data and Convenient Assumptions
Sharon-Lise T. Normand

doi: 10.1161/CIRCOUTCOMES.109.931543
Circulation: Cardiovascular Quality and Outcomes is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2010 American Heart Association, Inc. All rights reserved.
Print ISSN: 1941-7705. Online ISSN: 1941-7713

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://circoutcomes.ahajournals.org/content/3/1/2

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Circulation: Cardiovascular Quality and Outcomes 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: Cardiovascular Quality and Outcomes is online at:
http://circoutcomes.ahajournals.org//subscriptions/