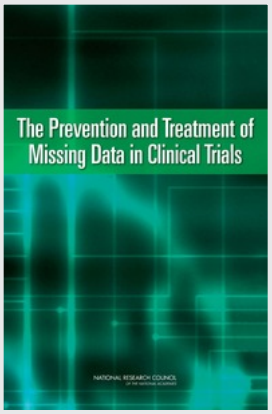




# The Prevention and Treatment of Missing Data in Clinical Trials (2010)

## DETAILS

162 pages | 6 x 9 | HARDBACK
ISBN 978-0-309-38686-9 | DOI 10.17226/12955

## CONTRIBUTORS

Panel on Handling Missing Data in Clinical Trials; Committee on National Statistics; Division of Behavioral and Social Sciences and Education; National Research Council

## SUGGESTED CITATION

National Research Council 2010. *The Prevention and Treatment of Missing Data in Clinical Trials*. Washington, DC: The National Academies Press. https://doi.org/10.17226/12955.

# The Prevention and Treatment of Missing Data in Clinical Trials

Panel on Handling Missing Data in Clinical Trials

Committee on National Statistics

Division of Behavioral and Social Sciences and Education

NATIONAL RESEARCH COUNCIL
*OF THE NATIONAL ACADEMIES*

THE NATIONAL ACADEMIES PRESS
Washington, D.C.
**www.nap.edu**

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, N.W. Washington, DC 20001

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

This study was supported by contract number HHSF223200810020I, TO #1 between the National Academy of Sciences and the U.S. Food and Drug Administration. Support for the work of the Committee on National Statistics is provided by a consortium of federal agencies through a grant from the National Science Foundation (award number SES-0453930). Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the view of the organizations or agencies that provided support for this project.

International Standard Book Number-13: 978-0-309-15814-5
International Standard Book Number-10: 0-309-15814-1

Additional copies of this report are available from the National Academies Press, 500 Fifth Street, N.W., Lockbox 285, Washington, DC 20055; (800) 624-6242 or (202) 334-3313 (in the Washington metropolitan area); Internet, http://www.nap.edu.



Printed in the United States of America

Suggested citation: National Research Council. (2010). *The Prevention and Treatment of Missing Data in Clinical Trials*. Panel on Handling Missing Data in Clinical Trials. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

# THE NATIONAL ACADEMIES

*Advisers to the Nation on Science, Engineering, and Medicine*

The **National Academy of Sciences** is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Ralph J. Cicerone is president of the National Academy of Sciences.

The **National Academy of Engineering** was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Charles M. Vest is president of the National Academy of Engineering.

The **Institute of Medicine** was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Harvey V. Fineberg is president of the Institute of Medicine.

The **National Research Council** was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Ralph J. Cicerone and Dr. Charles M. Vest are chair and vice chair, respectively, of the National Research Council.

**www.national-academies.org**

## PANEL ON HANDLING MISSING DATA IN CLINICAL TRIALS

RODERICK J.A. LITTLE (*Chair*), Department of Biostatistics, University of Michigan, Ann Arbor
RALPH D'AGOSTINO, Department of Mathematics and Statistics, Boston University
KAY DICKERSIN, Department of Epidemiology, Johns Hopkins University
SCOTT S. EMERSON, Department of Biostatistics, University of Washington, Seattle
JOHN T. FARRAR, Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine
CONSTANTINE FRANGAKIS, Department of Biostatistics, Johns Hopkins University
JOSEPH W. HOGAN, Center for Statistical Sciences, Program in Public Health, Brown University
GEERT MOLENBERGHS, International Institute for Biostatistics and Statistical Bioinformatics, Universiteit Hasselt and Katholieke Universiteit Leuven, Belgium
SUSAN A. MURPHY, Department of Statistics, University of Michigan, Ann Arbor
JAMES D. NEATON, School of Public Health, University of Minnesota
ANDREA ROTNITZKY, Departmento de Economia, Universidad Torcuato Di Tella, Buenos Aires, Argentina
DANIEL SCHARFSTEIN, Department of Biostatistics, Johns Hopkins University
WEICHUNG (JOE) SHIH, Department of Biostatistics, University of Medicine and Dentistry of New Jersey School of Public Health
JAY P. SIEGEL, Johnson & Johnson, Radnor, Pennsylvania
HAL STERN, Department of Statistics, University of California, Irvine

MICHAEL L. COHEN, *Study Director*
AGNES GASKIN, *Administrative Assistant*

# Acknowledgments

I would like to express appreciation to the following individuals who provided valuable assistance in producing this report. Particular thanks to Robert O'Neill and Tom Permutt at the U.S. Food and Drug Administration (FDA) for initiating the project, providing excellent presentations at the first meeting of the panel, and continuing support in providing timely information. We also thank Frances Gipson, FDA's technical representative, who assisted greatly in arranging the panel's first meeting at FDA and acquiring FDA documents throughout the study. The following FDA staff members presented invaluable information to the panel at its first meeting: Sharon Hertz, Henry Hsu, Robert O'Neill, Tom Permutt, Bruce Schneider, Norman Stockbridge, Robert Temple, Steve Winitsky, Lilly Yue, and Bram Zuckerman. At the panel's workshop on September 9, 2009, we benefited very much from the presentations of the following knowledgeable experts: Abdel Babiker, Don Berry, James Carpenter, Christy Chuang-Stein, Susan Ellenberg, Thomas Fleming, Dean Follmann, Joseph Ibrahim, John Lachin, Andrew Leon, Craig Mallinckrodt, Devan Mehrotra, Jerry Menikoff, David Ohlssen, and Edward Vonesh.

I am particularly indebted to the members of the Panel on Handling Missing Data in Clinical Trials. They worked extremely hard and were always open to other perspectives on the complicated questions posed by missing data in clinical trials. It was a real pleasure collaborating with all of them on this project.

I also thank the staff, especially our study director, Michael L. Cohen, who converted the musings of the panel into intelligible prose, arbitrated differences in opinion with good humor, and worked very hard on writing

and improving the report. I also thank Agnes Gaskin, who performed her usual exemplary service on all administrative matters. Eugenia Grohman provided extremely useful advice on presenting the material in this report, along with careful technical editing.

This report has been reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise, in accordance with procedures approved by the Report Review Committee of the National Research Council (NRC). The purpose of this independent review is to provide candid and critical comments that will assist the institution in making its published report as sound as possible and to ensure that the report meets institutional standards for objectivity, evidence, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process. We wish to thank the following individuals for their review of this report: Christy J. Chuang-Stein, Statistical Research and Consulting Center, Pfizer, Inc.; Shein-Chung Chow, Biostatistics and Bioinformatics, Duke University School of Medicine; Susan S. Ellenberg, Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania School of Medicine; Thomas Fleming, Department of Biostatistics, School of Public Health and Community Medicine, University of Washington; Yulei He, Department of Health Care Policy, Harvard Medical School; Robin Henderson, School of Mathematics and Statistics, University of Newcastle; Devan V. Mehrotra, Clinical Biostatistics, Merck Research Laboratories; Donald B. Rubin, Department of Statistics, Harvard University; and Steve Snapinn, Global Biostatistics and Epidemiology, Amgen, Inc.

Although the reviewers listed above have provided many constructive comments and suggestions, they were not asked to endorse the conclusions or recommendations nor did they see the final draft of the report before its release. The review of this report was overseen by Gilbert S. Omenn, Center for Computational Medicine and Biology, University of Michigan Medical School and Joel B. Greenhouse, Department of Statistics, Carnegie Mellon University. Appointed by the NRC's Report Review Committee, they were responsible for making certain that an independent examination of this report was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of this report rests entirely with the authoring panel and the institution.

Finally, the panel recognizes the many federal agencies that support the Committee on National Statistics directly and through a grant from the National Science Foundation. Without their support and their commitment to improving the national statistical system, the work that is the basis of this report would not have been possible.

Roderick J.A. Little, *Chair*
Panel on Handling Missing Data in Clinical Trials

# Contents

# Glossary

**Active Control:** In situations where the experimental therapy is to be an alternative to some existing standard of care, ethical or logistical constraints may dictate that the experimental therapy be tested against that "active" therapy that has previously shown evidence in an adequate and well-controlled clinical trial as an effective therapy. The ideal would be that patients would be randomized in a double blind fashion to either the experimental therapy or the active control, though the logistical difficulties of producing placebos for each treatment sometimes precludes a double blind study structure.

Contrasted with **Placebo control:** In situations where the experimental therapy is to be added to some existing standard of care, it is best to randomize subjects in a double-blind fashion to either the experimental therapy or a placebo control that is similar in appearance.

**Common Analysis Estimands:**

**Per Protocol:** In a per-protocol analysis, the analysis may be restricted to participants who had some minimum exposure to the study treatments, who met inclusion/exclusion criteria, and for whom there were no major protocol violations. The specific reasons for excluding randomized participants from a per-protocol analysis should be specified in advance of unblinding the data.

**Intention to Treat:** In an intention-to-treat analysis, all participants that satisfy the exclusion criteria are analyzed as belonging to the treatment arms to which they were randomized, regardless of whether they received or adhered to the allocated intervention for the full duration of the trial.

**As Treated:** In an as-treated analysis, the participants are grouped according to the treatment regimen that they received, which is not necessarily the treatment to which they were initially assigned.

**Complier-Averaged Causal Effect (CACE):** A parameter used to estimate the average effect of the treatment in the subpopulation of individuals that could remain on study or control treatments for the full length of the study.

**Dropout:** Treatment dropout is the result of a participant in a clinical trial discontinuing treatment; analysis dropout is the result of the failure to measure the outcome of interest for a trial participant.

**Enrichment:** Treatments are often only tolerated by or are only efficacious for a subset of the population. To avoid problems associated with treatment discontinuation, and to test a treatment on the subpopulation that can most benefit from it, it can be advantageous to determine whether a potential trial participant is a member of the subpopulation that can either tolerate or benefit from a treatment. This pretesting and selection of participants for trial participation prior to randomization into the treatment and control arms is called **enrichment**, and can include (1) selecting people with potentially responsive disease, (2) selecting people likely to have an event whose occurrence is the outcome of interest, (3) selecting people likely to adhere to the study protocol, and (4) selecting people who show an early response to the test drug.

**Last Observation Carried Forward (LOCF):** A single imputation technique that imputes the last measured outcome value for participants who either drop out of a clinical trial or for whom the final outcome measurement is missing. **Baseline Observation Carried Forward (BOCF):** A single imputation technique that imputes the baseline outcome value for participants who either drop out of a clinical trial or for whom the final outcome measurement is missing.

**Noninferiority vs. Superiority Trials:** A noninferiority clinical trial compares the experimental therapy to some active control with the aim of establishing that the experimental therapy is not unacceptably worse than an active control that showed evidence as an effective treatment in previously conducted adequate and well-controlled clinical trials. A noninferiority trial is often conducted in a setting in which (1) the experimental therapy, if approved, would be used in place of some existing treatment that was previously found to show evidence of effect, (2) it is not ethical or feasible to conduct a placebo controlled trial, (3) it would be clinically appropri-

ate to approve a new treatment that is only approximately equivalent to a current standard therapy with respect to some primary clinical outcome, and (4) the new experimental therapy might have other advantages such as a better adverse event profile, ease of administration, etc. Rather than rejecting a null hypothesis of equality between the experimental therapy and control treatment, a noninferiority clinical trial is designed to reject a null hypothesis that the experimental therapy is some specified amount ("the noninferiority margin") worse than the active control. Selection of the noninferiority margin must consider such issues as the magnitude of effect estimated for the active control in prior clinical trials, any bias that might be present in those previous trials relative to the effect of the active control in the population and setting used in the noninferiority trials, the proportion of effect that must be preserved for any approved treatment, etc.
A **Superiority** clinical trial is one in which an experimental therapy would be approved only if that therapy showed statistically credible evidence of superiority over a clinically relevant control therapy in an adequate and well-controlled clinical trial. The superiority trial is designed to reject a null hypothesis of equality between the experimental and control therapies.

**Randomized Withdrawal:** A clinical trial design in which all participants are initially provided the study treatment. Then, participants that have a positive response to the study treatment are randomly selected either to remain on the study treatment or to be switched to a placebo. Positive indications are when those that continue on study treatment are observed to have better outcomes than those who are switched to the placebo.

**Run-In Design:** Similar to an enrichment design, a run-in design is a design incorporating an initial period in which a subset of the participants are selected given indications as to their likelihood of compliance or the magnitude of their placebo effect. The key difference between a run-in design and an enrichment design is that the active treatment is not used to identify the subset of participants for study.

**Titration:** In opposition to a fixed dose protocol, titration is the adjustment of dosage to increase the treatment benefit and tolerability for participants during the course of a clinical trial.

**Washout:** (Placebo) washout is a period of time without active treatment that is scheduled before the beginning of use of study treatment, often used to eliminate any residual effects that might remain after a previous period on active treatment.

# Summary[1]

Randomized clinical trials are the primary tool for evaluating new medical interventions. More than $7 billion is spent every year in evaluating drugs, devices, and biologics. Randomization provides for a fair comparison between treatment and control groups, balancing out, on average, distributions of known and unknown factors among the participants. Unfortunately, a substantial percentage of the measurements of the outcome or outcomes of interest is often missing. This "missingness" reduces the benefit provided by the randomization and introduces potential biases in the comparison of the treatment groups.

In light of this problem, the Panel on the Handling of Missing Data in Clinical Trials was created at the request of the U.S. Food and Drug Administration (FDA) to prepare "a report with recommendations that would be useful for FDA's development of a guidance for clinical trials on appropriate study designs and follow-up methods to reduce missing data and appropriate statistical methods to address missing data for analysis of results."

The panel's work focused primarily on Phase III confirmatory clinical trials that are the basis for the approval of drugs and devices. For these trials, the bar of scientific rigor is set high; however, many of our recommendations are applicable to all randomized trials.

Missing data can arise for a variety of reasons, including the inability or unwillingness of participants to meet appointments for evaluation. And in some studies, some or all of data collection ceases when participants

[1]This Summary contains the highest priority recommendations. For a complete list of recommendations in the report, see Chapter 6.

discontinue study treatment. Existing guidelines for the design and conduct of clinical trials, and the analysis of the resulting data, provide only limited advice on how to handle missing data. Thus, approaches to the analysis of data with an appreciable amount of missing values tend to be ad hoc and variable.

The panel concludes that a more principled approach to design and analysis in the presence of missing data is both needed and possible. Such an approach needs to focus on two critical elements: (1) careful design and conduct to limit the amount and impact of missing data, and (2) analysis that makes full use of information on all randomized participants and is based on careful attention to the assumptions about the nature of the missing data underlying estimates of treatment effects. In addition to the highest priority recommendations here, in the body of the report the panel offers additional recommendations on the conduct of clinical trials and techniques for analysis of trial data.

Modern statistical analysis tools—such as maximum likelihood, multiple imputation, Bayesian methods, and methods based on generalized estimating equations—can reduce the potential bias arising from missing data by making principled use of auxiliary information available for nonrespondents. The panel encourages increased use of these methods. However, all of these methods ultimately rely on untestable assumptions concerning the factors leading to the missing values and how they relate to the study outcomes. Therefore, the assumptions underlying these methods need to be clearly communicated to medical experts so that they can assess their validity. Sensitivity analyses are also important to assess the degree to which the treatment effects rely on the assumptions used.

## TRIAL DESIGN

There is no "foolproof" way to analyze data subject to substantial amounts of missing data; that is, no method recovers the robustness and unbiasedness of estimates derived from randomized allocation of treatments. Hence, the panel's first set of recommendations emphasizes the role of design and trial conduct to limit the amount and impact of missing data.

A requisite for consideration of trial design is to clearly define the target population, and the outcomes that will form the basis for decisions about efficacy and safety. The treatment of missing data depends on how these outcomes are defined, and lack of clarity in their definition translates into a lack of clarity in how to deal with missing data issues. In addition, given the difficulties of adequately addressing missing data at the analysis stage, the design process needs to pay more attention to the potential hazards arising from substantial numbers of missing values.

**Recommendation 2: Investigators, sponsors, and regulators should design clinical trials consistent with the goal of maximizing the number of participants who are maintained on the protocol-specified intervention until the outcome data are collected.**

## DROPOUTS

A major source of missing data in clinical trials occurs when participants discontinue their assigned treatment. The two most common reasons for participants' dropping out are reactions to the treatment—it is ineffective, has unacceptable side effects, or is perceived as having worked—or moving to a different location where the treatment is not available. We call these "treatment dropouts," and distinguish them from analysis dropouts, which arise when the study outcomes are not measured, and are therefore unable to be included in the data analysis.

In some trials, protocols are written so that treatment dropout leads to analysis dropout because the sponsor may see no need to record study outcomes after a participant deviates from the study protocol. This approach can seriously undermine any inferences that can be drawn about effects. The panel concludes that it is important not only to obtain information about dropouts to the extent possible, but also to anticipate and plan for missing data in trial protocols.

**Recommendation 3: Trial sponsors should continue to collect information on key outcomes on participants who discontinue their protocol-specified intervention in the course of the study, except in those cases for which a compelling cost-benefit analysis argues otherwise, and this information should be recorded and used in the analysis.**

The techniques we suggest to limit the amount of missing data include

- choices of study sites, investigators, participants, study outcomes, time in study and times of measurement, and the nature and frequency of follow-up to limit the amount of missing data;
- the use of rescue therapies or alternative treatment regimens, to allow meaningful analysis of individuals who discontinue the assigned treatment;
- limiting participant burden in other ways, such as making follow-up visits easy in terms of travel and child care;
- providing frequent reminders of study visits;
- training of investigators on the importance of avoiding missing data;

- providing incentives to investigators and participants to limit dropouts; and
- monitoring of adherence and in other ways dealing with participants who cannot tolerate or do not adequately respond to treatment.

**Recommendation 6: Study sponsors should explicitly anticipate potential problems of missing data. In particular, the trial protocol should contain a section that addresses missing data issues, including the anticipated amount of missing data, and steps taken in trial design and trial conduct to monitor and limit the impact of missing data.**

## DATA ANALYSIS

Despite efforts to minimize missing data in the design and conduct of clinical trials, the statistical analysis often has to deal with a nontrivial amount of missing data. There is no single correct method for handling missing data. All methods require untestable assumptions because the missingness mechanism involves assumptions about the relationships among variables with missing values and results often vary depending on the assumptions made about these relationships. Crucially, the validity of these assumptions cannot generally be determined from the collected data. Consequently, the critical need is to understand the assumptions associated with any particular analysis, and those assumptions need to be expressed in as transparent a manner as possible so that researchers and practicing clinicians are able to assess their validity in any given setting.

**Recommendation 9: Statistical methods for handling missing data should be specified by clinical trial sponsors in study protocols, and their associated assumptions stated in a way that can be understood by clinicians.**

The panel believes that in nearly all cases, there are better alternatives to last observation carried forward and baseline observation carried forward imputation, which are based on more reasonable assumptions and hence result in more reliable inferences about treatment effects.

**Recommendation 10: Single imputation methods like last observation carried forward and baseline observation carried forward should not be used as the primary approach to the treatment of missing data unless the assumptions that underlie them are scientifically justified.**

Especially when the degree of missingness is appreciable and information about the characteristics of participants with missing data is limited,

the sensitivity of the inference to reasonable departures from the assumptions of the missing data method needs to be assessed. This additional uncertainty in the regulatory environment should motivate manufacturers of drugs, devices, and biologics to pay much greater attention to the use of techniques for reducing the frequency of missing data (see Chapters 2 and 3).

**Recommendation 15: Sensitivity analyses should be part of the primary reporting of findings from clinical trials. Examining sensitivity to the assumptions about the missing data mechanism should be a mandatory component of reporting.**

## NEW RESEARCH AND USE OF EXISTING RESEARCH

The FDA has a very large database of clinical trials that has not been tapped for its potential to inform the best practices for clinical trials. At the same time, there are a wide range of techniques that have been very thoroughly explored both theoretically and in practice over the past 20 years that are not being used in clinical trials. There seems to be a reticence on the part of analysts in both industry and the FDA to adopt those techniques. This reticence may be due in part to a lack of training and education.

**Recommendation 16: The U.S. Food and Drug Administration and the National Institutes of Health should make use of their extensive clinical trial database to carry out a program of research, both internal and external, to identify common rates and causes of missing data in different domains and how different models perform in different settings. The results of such research can be used to inform future study designs and protocols.**

**Recommendation 17: The U.S. Food and Drug Administration (FDA) and drug, device, and biologic companies that sponsor clinical trials should carry out continued training of their analysts to keep abreast of up-to-date techniques for missing data analysis. FDA should also encourage continued training of their clinical reviewers to make them broadly familiar with missing data terminology and missing data methods.**

**Recommendation 18: The treatment of missing data in clinical trials, being a crucial issue, should have a higher priority for sponsors of statistical research, such as the National Institutes of Health and the National Science Foundation. There remain several important areas in which progress is particularly needed, namely: (1) methods for sensitiv-**

**ity analysis and principled decision making based on the results from sensitivity analyses, (2) analysis of data where the missingness pattern is nonmonotone, (3) sample size calculations in the presence of missing data, (4) design of clinical trials, in particular plans for follow-up after treatment discontinuation (degree of sampling, how many attempts are made, etc.), and (5) doable robust methods, to more clearly understand their strengths and vulnerabilities in practical settings. The development of software that supports coherent missing data analyses is also a high priority.**

# 1

# Introduction and Background

Randomized clinical trials (RCTs) currently occupy a central role in assessing the effectiveness of proposed interventions to prevent and treat disease. Clinical trials are sponsored by the pharmaceutical and device industries, by government organizations such as the National Institutes of Health (NIH), by academic organizations, and by private organizations. The design and conduct of clinical trials and the analysis of the resulting data are carried out and/or overseen by the trial sponsor. However, for new drugs and devices, oversight, approval, and ultimate decision authority, in the form of regulation, is the purview of the U.S. Food and Drug Administration (FDA). Currently, more than $7 billion (Drennan, 2003) is spent annually on clinical trials by U.S. pharmaceutical and device companies to evaluate the safety and effectiveness of new drugs, devices, and biologics. (Given the date of this estimate, it is reasonable to assume that the current total is higher.) An NIH panel estimated that clinical trials represented one-third of NIH's expenditures for clinical research (see Nathan and Wilson, 2003).

At the request of FDA, the National Research Council convened the Panel on the Handling of Missing Data in Clinical Trials, under the Committee on National Statistics, to prepare "a report with recommendations that would be useful for FDA's development of a guidance for clinical trials on appropriate study designs and follow-up methods to reduce missing data and appropriate statistical methods to address missing data for analysis of results." The charge further specified:

> [t]he panel will use as its main information-gathering resource a workshop that will include participation from multiple stakeholders, including clinical trialists, statistical researchers, appropriate experts from the National Institutes of Health and the pharmaceutical industry, regulators from FDA, and participants in the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH).
>
> In both the workshop and report, the panel will strive to identify ways in which FDA guidance should be augmented to facilitate the cost-effective use of appropriate methods for missingness by the designers and implementers of clinical trials. Such guidance would usefully distinguish between types of clinical trials and missingness situations. For example, it could be useful to provide guidance on such questions as:
>
> (1) When missingness is likely to result in an appreciable bias such that sophisticated methods for reducing bias would be needed, and, conversely, under what circumstances simple methods such as case deletion could be an acceptable practice, and
>
> (2) How to use the leading techniques for variance estimation for each primary estimation method, along with suggestions for implementing these often complex techniques in software packages.

## RANDOMIZATION AND MISSING DATA

A key feature of a randomized clinical trial is comparison with a control group, with the assignment to either the control or the treatment group carried out using a random process. This eliminates intentional or unintentional bias from affecting the treatment assignment. Randomization also (probabilistically) balances the control and treatment groups for known and, more importantly, unknown factors that could be associated with the response or outcome of interest. By using randomization, the comparison between the treatment and control groups is made as fair as possible. Thus, randomization provides a basis for inference in the assessment of whether the observed average outcome for the treatment group is or is not sufficiently different than that for the control group to assert that the measured difference is or is not due to random variation. That is, randomization permits generalizations about outcomes.

Unfortunately, this key advantage, derived from the use of random selection for treatment and control groups, is jeopardized when some of the outcome measurements are missing. By missing data we mean when an outcome value that is meaningful for analysis was not collected. So, for example, a quality-of-life measure after death is not meaningful for analysis and should not be referred to as a missing outcome. Since whether or not data are missing can be related to the assigned treatment and to the

response, the absence of these data can bias the estimate of the treatment effect and weaken the resulting inference.

A common taxonomy for missing data, which is defined more rigorously in Chapter 4, distinguishes between missing data that are missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR):

- In the case of MCAR, the missing data are unrelated to the study variables: thus, the participants with completely observed data are in effect a random sample of all the participants assigned a particular intervention. With MCAR, the random assignment of treatments is assumed to be preserved, but that is usually an unrealistically strong assumption in practice.
- In the case of MAR, whether or not data are missing may depend on the values of the observed study variables. However, after conditioning on this information, whether or not data are missing does not depend on the values of the missing data.
- In the case of MNAR, whether or not data are missing depends on the values of the missing data.

If MAR or MNAR holds, then appropriate analysis methods must be used to reduce bias. It is important to note that increasing the number of participants is insufficient for reducing bias.

There are a number of choices for trial outcomes, trial designs, and trial implementation that can substantially increase or decrease the frequency of missing data. Some of the aspects of clinical trials that can affect the amount of missing data include whether data collection continues for participants who discontinue study treatment, the use of outcomes that are at risk of being undefined for some patients, the rate of attrition, and the use of composite outcomes.

***Missing Data Due to Discontinuation of Study Treatment*** It is common for some participants in a clinical trial to discontinue study treatment because of adverse events or lack of efficacy. (And, there may be more than one reason for any specific outcome to be missing, for example, a combination of an adverse effect and a lack of efficacy.) Some trial protocols stipulate that data collection stop or be abbreviated following discontinuation of study treatment. For example, in some trials, data collection is only continued for a short period (e.g., 14 days) following treatment discontinuation, assuming that adverse events after that point in time are unlikely to be attributable to the randomly assigned study treatment. Moreover, in some trials, participants are offered an alternative treatment that is not part of the study following discontinuation. As a result, subsequent data collection may be considered to be uninformative for comparing the randomly assigned treat-

ments. A positive outcome, such as symptom relief, recovery, or cure, may also lead to discontinuation of treatment.

Since treatment discontinuation often arises from changes in a participant's health status, the data that are not collected after treatment discontinuation are likely to be informative of the change in health status. Another relevant factor is that in nonblinded studies, such as device trials, a participant who knows if he/she does or does not have the study treatment might be more or less likely to either report adverse events or to report more or less efficacy. Such knowledge could be related to whether the participant discontinues treatment.

***Use of Outcomes That Are at Risk of Being Undefined for Some Patients*** Some clinical trials use outcomes that may not be ascertainable for all participants. Examples include: (1) a quality-of-life assessment that cannot be obtained due to the death of the participant, (2) a measurement (e.g., a 6-minute walk test) from a procedure that some participants cannot complete because of their health status, and (3) assessment of renal progression for participants, some of whom undergo kidney transplant during the course of the study. Since all of these situations involve health status, it is likely that whether or not data are missing is related to the changes in health status and hence are MNAR.

Although it is important to define clinical endpoints that are measurable for as many participants as possible in order to reduce the impact of missing data, in doing so one must also consider the impact on the relevance of the primary endpoint. So, for example, forming composite outcomes to include events such as "discontinuation of treatment" or "exposure to rescue treatment," while useful in reducing the frequency of missing data, may lessen the clinical relevance of the outcome of interest.

***Missing Data Because of Attrition in the Course of the Study*** The longer the planned length of a clinical trial, the greater the chance that participants will drop out of the trial due to their moving out of the area or otherwise experiencing changes in their lives that preclude or complicate further participation. If dropping out due to these situations is known to be unrelated to changes in health status, an MAR assumption for the missing values seems justified; however, if dropping out is related to health status (e.g., a move to live with and be cared for by a parent or offspring), then the MAR assumption is not justified, and the missing data are likely not MAR.

***Missing Data in Composite Outcomes*** Outcomes that are composites of a variety of variables, such as health indices, or combined measures that address the multidimensional nature of the benefit from an intervention may not be defined when any of the variables that are being combined are

missing (although there are composite outcome measures for which this is not necessarily the case).

***Missing Data Due to Death*** The treatment of death in the context of missing data is complicated. There are three kinds of approaches, which are linked to situations. One approach is to consider cause-specific death as a primary endpoint (e.g., death related to cardiovascular event). In this case, death for other reasons (e.g., not related to the clinical study) may properly be treated as a censoring event. For example, death due to an auto accident could be considered to be a censoring event. Particular care must be taken in this situation to ensure that censoring due to death by other reasons can be grouped together with general censoring patterns. It may very well be the case that censoring due to death for other reasons is dependent on the primary endpoint itself, in which case the censoring could be a missing not a random process.

A second approach is to fold death into another outcome to form a composite outcome: for example, time to AIDS-defining illness or death.

The third situation and the main complication for a clinical trial is when death is related to the outcome of interest, as with AIDS-related death in a study where CD4 is the primary outcome. In this case, the estimand must be carefully defined, possibly as CD4 among those who would remain alive on either treatment. This approach is related to principal stratification on a postrandomization event (see Frangakis and Rubin, 2002). Inverse probability weighting can also be used in this case. The key consideration here is that the estimand must represent a causal contrast. A nontrivial complication in interpreting the estimand is that it applies to a subgroup that cannot necessarily be identified; namely, those who would have survived in either treatment group.

Two general lines of attack have been employed to address the problem of missing values in clinical trials. The first is simply to design and carry out the clinical trial in a manner that limits the amount of missing data. As discussed in Chapters 2 and 3, there are a variety of techniques for doing this, and these techniques are not used as much as one would hope. One reason for this is that the designs for limiting missing data may involve tradeoffs against other considerations, such as generalizability or relevance of outcome measure. However, many of these techniques incur relatively minor costs. In any case, we believe that if the benefits of these methods were better appreciated and these methods were more widely implemented, the frequency of missing values could be substantially reduced in many clinical trials.

The second line of attack for the treatment of missing data is to apply analysis methods that exploit partial information in the observed data about the missing data to reduce the potential bias created by the missing

data. Many of the techniques currently used for this purpose are simplistic and rely on relatively extreme assumptions. Superior analysis techniques are often not applied, for several reasons. First, the expense of developing new interventions leads naturally to risk-averse behavior when drug or device developers are faced with the regulatory process. Second, FDA may at times prefer the use of older analysis methods that are better understood. Third, the need to prespecify analyses in the study protocol inhibits the use of complex analysis methods. Fourth, until recently, some of the newer techniques have lacked readily available and tested software. Finally, there seems to be a need for more training of biostatisticians both in industry and at FDA in the use of state-of-the-art missing data methods. The lack of experience with the new methods results in a lack of consensus about how and when these methods should be used in clinical trials.

Improvements in trial design, trial conduct, and the analysis of trial data in the presence of missing data are not adequately recognized in current U.S. and international guidelines for clinical trials. Although these official documents have provided some very useful guidance, overall they are too general, and they therefore fail to be sufficiently prescriptive. That is, they lack detailed suggestions as to when and how specific methods can be implemented. In this report, we provide some guiding principles and specific methods for handling missing data in clinical trials. Our goal is to improve the quality of estimates of treatment effects and their associated estimates of uncertainty in randomized clinical trials.[1]

## THREE KINDS OF TRIALS AS CASE STUDIES

In this report, we use three types of trials to illustrate how clinical trial design and other aspects of trial conduct can be modified to limit the impact of missing data on regulatory decisions; trials for chronic pain, trials for the treatment of HIV, and trials for mechanical circulatory devices for severe symptomatic heart failure. These examples are chosen both because they are important in their own right and because they share many characteristics with a wide variety of other types of clinical trials. In this section, we describe the usual analytic approaches and their deficiencies for these examples.

### Trials for Chronic Pain

Clinical trials are used to assess the ability of an intervention to provide symptomatic relief from conditions, such as osteoarthritis, that cause

[1]The *Journal of Biopharmaceutical Statistics* had a special issue in 2009 "Missing Data—Prevention and Analysis" (Vol. 19, No. 6) that we recommend to readers of this report.

chronic pain. These trials are typically conducted over 12 weeks, and they are subject to very high rates of treatment discontinuation. The reasons for treatment discontinuation usually differ between the treatment and the control groups. For example, in placebo-controlled trials, discontinuation in the placebo group often stems from inadequate efficacy (i.e., lack of pain relief), while discontinuation in the treatment group more often arises because of poor tolerability (of the medication being tested). Trial designs that involve fixed doses leave few treatment options for patients who experience inadequate efficacy or poor tolerability. Patients who stop study treatment usually switch to a proven (approved) effective therapy, and the trial sponsors typically stop collecting pain response data on those patients who discontinue study treatment.

In current practice, the data from these types of clinical trials have been analyzed by using (single-value) imputation to fill in for the missing outcome values. In particular, it has been common to use the last observation carried forward (LOCF) imputation technique to impute for missing values. LOCF implicitly assumes that a participant who had good pain control in the short term and then dropped out would have had good pain control in the long term. This assumption seems questionable in many settings. Another frequently used, although somewhat less traditional imputation technique is the baseline observation carried forward (BOCF) technique, which assumes that a participant's pain control is the same as that measured at the beginning of the trial. Since most patients in chronic pain studies, including those on placebos, improve substantially from the baseline over time, BOCF is likely to underestimate the effectiveness of any treatment. Furthermore, use of such imputation schemes, in conjunction with complete data techniques, can result in estimated standard errors for treatment effects that fail to properly reflect the uncertainty due to missing data.

### Trials for the Treatment of HIV

The goal of many HIV trials is to determine whether a new drug has safety and efficacy that is comparable with that of an approved drug used for initial antiretroviral treatment (ART). The studies involve samples of ART-naïve participants and use noninferiority designs (U.S. Food and Drug Administration, 2002). The focus for current purposes is on the primary efficacy outcome, which is the percentage of participants with sufficiently low viral load at the end of the reference period. (Other considerations, such as choice of control, noninferiority margin, and blinding, are therefore ignored.) Since combination treatment is the norm for HIV, the typical design in this setting is new drug A plus background treatment compared with current drug B plus the same background treatment, measured over a period of 24 or 48 weeks.

The primary outcome efficacy for these trials is typically based on plasma RNA measurements (see, e.g., U.S. Food and Drug Administration, 2002, Appendix B), in which the primary outcome is the success rate among all participants randomized into the trial. Treatment failures are defined to include (1) study participants who die or switch away from the study drug before 48 weeks, (2) study participants who do not attend the 48-week visit, and (3) study participants who remained on the study drug but who have an HIV RNA level equal to or greater than 50 copies/mL at 48 weeks. This definition can be viewed as a composite outcome in which failure is due to treatment discontinuation, to missing data, and to not meeting the "success" level.

Many such trials have moderate to large numbers of patients who either discontinue treatment before 48 weeks or who do not attend the 48-week visit. Participants are typically not followed after discontinuing treatment, and there are probably various reasons for discontinuation.

One problem with the current approach involves the use of an analysis of the percentage of participants with a viral load of less than 50 copies/mL at 48 weeks for all randomized participants according to their initially assigned treatment. Such an analysis is not possible because data collection is discontinued following failure. With this approach, reasons for "failure" (e.g., losses to follow-up and lack of efficacy due to virologic failure) are given equal weight, which may complicate the interpretation of the results. Furthermore, discontinuation of data collection after failure limits analyses that can be performed on separate components of the composite outcome. Moreover, it can result in failure to capture critical long-term effects of a discontinued study drug whose use may have increased the probability of resistance to alternative therapies.

### Trials for Mechanical Circulatory Devices for Severe Symptomatic Heart Failure

For patients with advanced heart failure, heart implantable left ventricular assist devices (LVADs) have been shown to be effective when used as a bridge to heart transplants.[2] Furthermore, because many patients are not eligible candidates for transplantation, the use of LVADs as destination therapy has been shown to be effective, and its use is increasing.

Over time, it has been possible to make the devices smaller and more durable. With the new devices, thrombogenic (tendency to clot) and infec-

[2]For two examples, see Rose et al. (2001), which established the superiority of an LVAD over medical therapy as destination therapy, and Slaughter et al. (2009), which compared a new LVAD device to the type used in the first trial and established the superiority of the new design.

tion risks are more easily managed, and morbidity and mortality have been reduced during the periods before, during, and immediately after the operation. Thus, there is interest in using these devices as destination therapy in patients with symptomatic heart failure who are severely impaired in spite of optimal medical therapy but who are less sick than patients studied in earlier destination therapy trials. As a case study, we consider a superiority design trial in which the goal is to determine whether an LVAD is superior to optimal medical management for prolonging patients' survival and optimizing their health status.

In many device trials, blinding of patients and investigators is not possible. A successful trial is one in which the LVAD would substantially improve functional status and not negatively impact survival.

In such a clinical trial, survival status will be ascertained for nearly all patients. However, functional status during follow-up may be missing because of early death (for example, as a consequence of the implantation procedure), failure to attend examinations, or inability to perform functional tests like a 6-minute walk. Also, as several LVADs are already approved for use in patients with more advanced disease, it is expected that some control patients will be implanted during the course of the study as their disease progresses. In addition, some patients will receive heart transplants and this, too, will complicate the interpretation of functional measures. Finally, some patients may have the LVAD removed for other reasons.

Some of the design and data analysis considerations for such a trial would also apply to a trial that compared two LVAD devices (e.g., an unapproved newer design and an approved one with an older design) as destination devices among patients ineligible for transplantation.

A key issue in the design is the definition of the major outcomes related to survival and health status. For destination trials among patients who are not eligible for transplantation, FDA has accepted a composite outcome of death or disabling stroke after 2 years as the outcome. In some trials, a second operation to replace the device is also included in the primary outcome.

Measures of functional status and a patient's health-related quality of life have been key secondary outcomes in trials that studied patients who are ineligible for transplantation. In a target population that is not as sick as those eligible for transplantation, a measure of health status might be considered as a coprimary outcome because that outcome will be an important consideration in device approval and use. In addition, because the superiority of LVADs to medical therapy has been established in patients ineligible for transplantation, when the above criteria are met by patients in a control group that receives optimal medical management, use of an LVAD will have to be permitted. Thus, those criteria may also have to be considered as a component of a composite outcome.

In the LVAD trials done to date, a major problem has been missing health status data. In addition, obtaining objective assessment of health status (e.g., health-related quality of life, or being able to do a 6-minute walk) is complicated by the fact that it is not possible to have blinding in such trials. In these trials, missing data occur because of death either as a result of the implantation procedure or underlying disease (see discussion above), failure to attend examinations, inability to perform functional tests (e.g., a 6-minute walk), or "questionnaire fatigue" for self-administered quality-of-life instruments. Also, many health status measures include a collection of responses to multiple items that comprise different domains, and item nonresponse is also a problem (i.e., some, but not all items, on a quality-of-life instrument are missing). Analyses have typically used methods that assumed the data were missing at random, but this assumption is clearly not appropriate given the reasons for the missing data. To determine whether the degree of missing data has had a sufficient impact on the analysis to substantially affect the study findings, a sensitivity analysis is required. We discuss how to carry out a sensitivity analysis in this context in Chapter 5.

## CLINICAL TRIALS IN A REGULATORY SETTING

This report focuses primarily on issues concerning the treatment of missing data in randomized controlled clinical trials that are intended to support regulatory applications for drugs, medical devices, and biologics. Several aspects of the regulatory setting have particular bearing on how missing data issues are handled. In particular:

- Regulators generally must render yes or no decisions rather than just describing the data and possible interpretations.
- Clinical trial sponsors, who must make substantial investment decisions in pursuit of regulatory approval, seek predictability regarding what findings would support a favorable decision; and regulators, eager to ensure common requirements across all sponsors and to enable quality development, also prefer to improve predictability where feasible.
- Regulators generally require a high level of confidence before making a conclusion of safety and efficacy, preferring in close or ambiguous cases to be "conservative," that is, to err on the side of withholding approval. This conservatism results, in part, from the fact that a regulatory approval may make further studies unlikely (due to lack of feasibility or funding).
- In most cases, clinical trials in the regulatory process are focused on determining the effects of a specific product. Effects that occur after

switching to rescue therapy in patients who did not tolerate or respond well to the study therapy are sometimes disregarded because those effects may well not be attributable to the study therapy in question.

In the regulatory environment, a strong premium is placed on specification of analytic methods prior to a trial. Such specification serves not only to help preserve the type I error (i.e., the error of asserting that a treatment is more effective than the control when it is not), but also to improve the predictability of the decision process. Pretrial specification of the planned primary analyses are of particular importance and therefore receive the greatest attention. However, secondary and sensitivity analyses can also play a key role in the decision process, and they certainly are more valuable than post-hoc, exploratory analyses. Therefore, there is also a need to specify the secondary analysis prior to a trial and to specify in advance the approach for analyzing the sensitivity of the primary analysis to divergences from the statistical models used to accommodate missing data in that analysis (and the sensitivity to other divergences, such as outlying values).

We believe that the need for a dichotomous decision and the tendency for conservatism should create particularly strong incentives on the part of sponsors to minimize the quantity and effects of missing data and to use statistical models, when analyzing the resulting data, that are based on assumptions that are plausible and, when possible, validated. As there are many potential approaches to handling missing data, pretrial specification of an approach to be used in the primary analysis is particularly important to help ensure predictability. However, because the assumptions underlying any one approach to handling missing data may be invalid, prospective definition of sensitivity analyses with different underlying assumptions will help assess the robustness of the conclusions and help support effective decision making.

Obtaining regulatory approval for a therapy involves generating information on many aspects of its effects, often including, but not limited to, short-term effects, long-term effects, effects at various fixed doses, effects in various clinical settings, and effects with various concomitant therapies. In some cases, attempts to address many of these aspects in the same trial may lead to problems with missing data, particularly in assessing long-term effects. Such problems may be avoidable by designing each trial specifically to address fewer aspects, though this would raise the development costs.

Although the above considerations regarding missing data may be particularly applicable to trials in the regulatory setting, many are also relevant to trials in other clinical trial settings. Therefore, we believe that most of the recommendations and discussion in this report are also applicable to trials outside the regulatory setting.

## DOMESTIC AND INTERNATIONAL GUIDELINES ON MISSING DATA IN CLINICAL TRIALS

There have been several recent documents that lay out a set of general principles and techniques for addressing the problems raised by missing data in clinical trials. These documents include

1. *Draft Guidance on Important Considerations for When Participation of Human Subjects in Research Is Discontinued*, from the Office for Human Research Protections in the U.S. Department of Health and Human Services (2008).
2. *Guidance for Sponsors, Clinical Investigators, and IRBs*: *Data Retention When Subjects Withdraw from FDA-Regulated Clinical Trials*, from the U.S. Food and Drug Administration (2008).
3. *Statistical Principles for Clinical Trials; Step 5: Note for Guidance on Statistical Principles for Clinical Trials*, from the European Medicines Evaluation Agency (EMEA) International Conference on Harmonisation (ICH) (1998) Topic E9.
4. *Guideline on Missing Data in Confirmatory Clinical Trials*, Committee for Medicinal Products for Human Use (CHMP) from the European Medicines Evaluation Agency (2009).

The first three documents are currently in use; the fourth *Guideline on Missing Data in Confirmatory Clinical Trials* had been issued only in draft form at the time of this writing.

In this section, we summarize the main points in these documents. They agree on several points:

- There is a need to anticipate the amount and nature of missing data during the design of the study and when making analysis plans. Careful planning will help specify a reasonable approach to handling missing data and will also help to specify a range of sensitivity analyses that could explore the impact of departures from the expected missing data pattern.
- It is important to collect complete data for all randomized participants, including those who discontinue treatment. This point motivates the important distinction between continuation of treatment and continuation of follow-up for major outcomes.
- The CONSORT (consolidated standards of reporting trials) guidelines for reporting the results of trials should be adhered to. Given that there will almost always be some missing data, a trial may still be regarded as providing valid results if the methods of dealing with missing values are sensible.

- The use of various single imputation methods is criticized, including the LOCF method.
- Given that no universally applicable methods of handling missing values can be recommended, an investigation should be made concerning the sensitivity of the results of analysis to the method of handling missing values, especially if the percentage of missing values is substantial.

The panel believes that the need for conservative methods receives too much emphasis in these guidelines. However, in general, this report can be seen as reinforcing and expanding on many of the suggestions and recommendations found in the four documents. We support and refine many of the basic principles proposed regarding the treatment of missing data in clinical trials, and we provide more detailed suggestions on specific techniques for avoiding missing data in the design and conduct of clinical trials and on appropriate analysis methods when there are missing data.

Recently, O'Neill (2009) stated that the issue of how to best handle missing data in clinical trials was a long-standing problem, especially in regulatory submissions for trials intended to support efficacy and safety and marketing approval, and he called for the development of a consensus as to the proper methods for use. He added that more information was needed on why subjects withdraw from their assigned therapies, and, when they do withdraw, on what amount of bias is introduced in the resulting estimates through the use of various methods. In addition, he pointed out that a key question is how to specify, in a trial protocol, the primary strategy for dealing with missing data when one has yet to observe the patterns of missing values. Finally, FDA's critical path initiative has identified the issue of missing data as a priority topic.

## REPORT SCOPE AND STRUCTURE

The panel believes it is important to provide a consensus on good practice in trial design, trial conduct, and the treatment of missing output values in the analysis of trial data. That is the goal of this report. More particularly, the focus of this report is the treatment of missing data in confirmatory randomized controlled trials of drugs, devices, and biologics, although, as noted above, we believe the material is also relevant for other types of clinical trials, including those carried out by academics and NIH-funded trials, and more generally for various biostatistical investigations. We note that no further mention is made in this report about methods for the treatment of missing data for biologics because they raise no issues that are not already raised in drug trials.

While the main context for our report is randomized trials, regulatory agencies such as FDA also evaluate evidence from trials where randomiza-

tion of interventions is considered impractical, as for example in some trials of devices or surgical procedures. These trials do not possess the balancing property of randomization with respect to the distribution of observed or unmeasured covariates, and hence are subject to potential bias if there are important differences in these distributions across intervention groups.

The threat to validity from missing data is similar for nonrandomized and randomized trials—in fact the threat is potentially greater given the inability to mask the treatments—so the principles of missing data analysis described in this report apply in a similar fashion to nonrandomized trials. These include the need to design and conduct trials to minimize the amount of missing data, the need to use principled missing data adjustments based on scientifically plausible assumptions, the need to conduct sensitivity analyses for potential deviations from the primary assumed mechanisms of missing data, and the need to collect covariate information that is predictive of missingness and the study outcomes. The need for good covariate information is, if anything, even greater for nonrandomized trials, since this information can also be used to reduce differences in intervention groups arising from the nonrandomized allocation of interventions.

This study included only four panel meetings, one of which was a workshop, and therefore cannot be comprehensive. The focus was on identifying principles that could be applied in a wide variety of settings. We recognize that there are a wide variety of types of clinical trials for a wide variety of health issues and that there will always be idiosyncratic situations that will require specialized techniques not directly covered here. Also, it is important to point out that we focus on the assessment of various forms of intervention efficacy: this report does not do any more than touch on the assessment of the safety of medical interventions.

The next two chapters provide details and recommendations on trial designs and trial conduct that are useful for reducing the frequency of missing data. Chapters 4 and 5 describe methods of analysis for data from clinical trials in which some of the values for the outcome or outcomes of interest are missing: Chapter 4 considers drawing inferences when there are missing data, and Chapter 5 considers sensitivity analyses. The final chapter presents the panel's recommendations.

# 2

# Trial Designs to Reduce the Frequency of Missing Data

> Good design may not eliminate the problem of missing data, but . . . it can reduce it, so that the modern analytic machinery can be used to extract statistical meaning from study data. Conversely, we note that when insufficient attention is paid to missing data at the design stage, it may lead to inferential problems that are impossible to resolve in the statistical analysis phase. (Lavori et al., 2008, p. 786)

The primary benefit from randomizing clinical trial participants into treatment and control groups comes from balancing the distributions of known and unknown characteristics among these groups prior to study treatments. But baseline comparability cannot be assured by randomization when data are missing. Although there are techniques that can be applied to ameliorate the impact of missing data (see Chapters 4 and 5), avoiding or minimizing missing data is always preferred.

Any approach to statistical analysis involving missing data will involve unprovable assumptions, particularly because there is always some uncertainty about the reasons why data are missing. Consequently, the appropriate assumptions and analytic treatment—and, therefore, the appropriate inference—may be unclear. For example, in a study of weight gain, Ware (2003, pp. 2,136-2,137) writes that: "It is unfortunate, however, that so much effort must be devoted to evaluating the implications of missing observations when a seemingly simple effort to obtain study weights according to the follow-up protocol would probably have been successful with most participants. Complete evaluation of enrolled patients, irrespective of their adherence to study therapy, deserves wider recognition as an

important part of good clinical-trials practice." Therefore, an important objective in the design and implementation of a clinical trial is to minimize missing outcome data.

As briefly discussed in Chapter 1, there are a variety of reasons for discontinuation of treatment, and for discontinuation of data collection, which we refer to as "analysis dropout," in clinical trials. The frequency of missing data depends on the health condition under study, the nature of the interventions under consideration, the length of the trial, and the burden of the health evaluations and how much they are facilitated. Common reasons for dropout include (1) inability to tolerate the intervention, (2) lack of efficacy for the intervention, and (3) difficulty or inability to attend clinical appointments and complete medical evaluations. As noted in Chapter 1, in some trials, treatment dropout leads to analysis dropout because data collection is discontinued. In many studies, this is the major reason for missing data. Other reasons include subjects who withdraw their consent, move out of the area, or who otherwise experience changes in their lives that preclude or complicate further participation.

This chapter primarily concerns the sources of missing outcome information and how the frequency of missing outcome values can be reduced. However, missing values of covariates and other auxiliary variables that are predictive of the outcome of interest should also be reduced,[1] and the techniques discussed here can be helpful for that purpose as well. There is clearly a need for more research on the specific reasons underlying missing data, a topic addressed in Chapter 3.

## TRIAL OUTCOMES AND ESTIMANDS

A clinical trial typically measures outcomes that quantify the impact of the interventions under study for a defined period of time. Inference focuses on summaries of these measures (such as the mean) for the target population of interest. These summary quantities are often called parameters, or *estimands*. For example, consider a trial in which the primary outcome measure is change in blood pressure between baseline and 6 weeks after the initiation of treatment. The estimand of interest might be the difference in the mean change in blood pressure over 6 weeks for the target and control populations. An estimate of this parameter is the difference in sample means for participants in the treatment group and participants in the control group. This estimate is unbiased if the assignment to treatment is random, and there are no missing data (Little and Rubin, 2002). The goal

[1]Some missing demographic data that are likely to have remained invariant across the study duration could be collected after a participant completes the trial.

is to attribute the difference between the treatment and the control to the causal effect of the intervention.

Estimation of the primary (causal) estimand, with an appropriate estimate of uncertainty, is the main goal of a clinical trial. For example, estimates of estimands based on the measurement of symptom relief are required for the regulatory evaluation of treatments for many disorders, including mental illnesses, inflammatory bowel disease, and chronic pain. Estimands in trials of interventions for cancer and heart failure are often based on survival or disease recurrence. Estimands for trials of interventions for HIV and hypertension may use surrogate outcomes, such as CD4 counts or blood pressure measures, respectively.

The choice of estimand involves both the outcome measure and the population of interest. For instance, the population of interest might be unrestricted, or it might be restricted to people who can tolerate an intervention for a given period. The outcome measure also requires specification of a period of action of the intervention after assignment. It may be measured at one time or over a period of time and can reflect either short-term or longer-term effects. In addition, the outcome measure could be an absolute measure of change from the baseline or a percentage change.

In order to avoid confusion and clearly assess the potential for bias from missing data in a randomized clinical trial, it is important to be clear about the choice of estimand, particularly about the outcome measure and target population of interest. The estimand should be decided before the protocol for a clinical trial is final, since alternative choices of estimand may have important implications for trial design and implementation, the inferences that are made, and the amount of missing data that might be expected. For instance, an important consideration is whether to collect outcome data after a subject discontinues assigned treatment or there are other deviations from the protocol. The answer to this question depends on the choice of trial estimand.

To make this discussion more concrete, in this section we discuss some estimands. We assume for simplicity that there are just two groups, a treatment group and a control group. We discuss different trial designs for each estimand, which provide varying degrees of confidence in the trial conclusions, given the likely frequency of missing outcome values. In this discussion, "outcome" refers to the primary outcome in the trial, which might be a measure of symptoms, a surrogate outcome, or a time to an event. The "duration of protocol adherence" refers to the time after randomization for which a subject received the study intervention according to the protocol. This period may be shorter than the full study duration for a variety of reasons, including lack of tolerance, inefficacy, and study burden.

Five possible estimands are described here in the context of a symptom relief trial. A similar range of potential estimands can be identified for trials

in which the outcome is a time to event or a surrogate measure of progress. Estimands (1), (2), (4), and (5) are frequently used. Estimand (3) is rarely used, but is included to help explain the impact of the various choices of estimand on the likelihood of missing data.

1. ***(Difference in) Outcome Improvement for All Randomized Participants*** This estimand compares the mean outcomes for the individuals randomized to the treatment and control arms, regardless of what treatment participants actually received. Often called the "intention-to-treat" estimand, it assesses the benefits of a treatment policy or strategy relative to a control.[2] Since the estimand relates to a treatment policy, the observed differences reflect the effect of the initially assigned treatment as well as subsequent treatments adopted as a result of intolerance or lack of efficacy.

A trial design that supports the use of this estimand is a parallel-group randomized trial in which outcome data are collected on all subjects, regardless of whether the study treatment is received. A trial design that does not support the use of this estimand is a parallel-group randomized trial in which outcome data are *not* collected on participants after they drop out or switch from the assigned treatment.

2. ***(Difference in) Outcome Improvement in Tolerators*** This estimand quantifies the degree of outcome improvement in subjects who tolerated and adhered to a particular treatment. This estimand concerns the subset of the population who initially began treatment and tolerated the treatment. One complication with this estimand is that it is difficult to identify the members of this subpopulation in advance in the field, and the assessed performance in a trial may therefore be an overestimate of the performance in practice.

A trial design that supports the use of this estimand is a targeted parallel-group randomized trial. An example is a design with an active treatment run-in period followed by placebo washout prior to randomization, limited to individuals who tolerated the active treatment during the run-in period. Outcome data are then collected on all randomized subjects. A trial design that does *not* support the use of this estimand is an untargeted parallel-group randomized trial in which outcome data are *not* collected on participants after they terminate or switch from the assigned treatment.

[2]To the extent that the subjects represent the population of patients with the specified disorder and to the extent that the health care of the subjects reflects the health care available in the population, the use of this estimand concerns the effectiveness of this treatment for the population.

3. ***(Difference in) Outcome Improvement If All Subjects Tolerated or Adhered*** This estimand quantifies the degree of outcome improvement in all subjects in the trial if they had all received treatment according to the protocol for the study duration. This estimand requires an imputation of what would have been the outcome if individuals who did not comply with the protocol had complied. Given that in many settings some amount of intolerability or nonadherence is unavoidable, this estimand reflects the effects of an infeasible treatment policy.

This estimand provides insight into the magnitude of improvement in efficacy that might be achieved if one could develop therapeutic strategies that produced very high levels of adherence in real-world settings; ultimately, such therapeutic strategies would then need to be evaluated in randomized clinical trials conducted to assess their effect on the estimand "outcome improvement for all randomized participants."

A trial design that supports the use of this estimand is a parallel-group randomized design in which all subjects are provided adjunctive or supportive therapies, assuming that such therapies are available and ensure tolerance and adherence. Outcome data are collected on all subjects. A trial design that does not support the use of this estimand is a parallel-group randomized design with no available and effective mechanism for ensuring adherence.

4. ***(Difference in) Areas Under the Outcome Curve During Adherence to Treatment*** This estimand compares the arm-specific means of the area under the outcome curve over the duration of protocol adherence. This estimand simultaneously quantifies the effect of treatment on both the outcome measure and the duration of tolerability or adherence in all subjects.

A trial design that supports the use of this estimand is a targeted parallel or parallel-group randomized trial. In such a trial with this estimand, there would be no need to collect outcome data after assigned treatment is discontinued or switched, other than to address secondary analysis issues, such as delayed side effects.

5. ***(Difference in) Outcome Improvement During Adherence to Treatment*** This estimand is the difference in mean outcomes from the beginning of the trial to the end of the trial or the end of adherence to the protocol, whichever occurs earlier. This estimand reflects both the duration of tolerability or adherence and outcome improvement in all subjects. A trial design that supports the use of this estimand is a parallel-group randomized trial. Again, in such a situation, estimating the primary estimand does not require collection of outcome data after the assigned treatment is discontinued.

Because estimands (1) outcome improvement, (4) area under the outcome curve during tolerated treatment, and (5) outcome improvement dur-

ing tolerated treatment may be influenced by both pharmacological efficacy and tolerance and adherence, they have the potential to be misinterpreted depending on whether the focus is on assessing intervention effectiveness or efficacy.

The advantage of (1) over (5) is that (1) has the alternative interpretation of the difference of two treatment policies. Estimand (5) is also problematic because it does not distinguish an immediately highly effective but extremely toxic treatment (i.e., one with no tolerability after a short period and high outcome difference over the short period of tolerability) from a nontoxic treatment with gradual outcome improvement (i.e., one with full tolerability and a difference outcome over the entire trial period of the same magnitude as that of the first treatment).

The choice of causal estimand and trial design needs to take into consideration the fact that clinical trials are often part of a larger strategy of exploring various features of an intervention prior to approval. For instance, for estimand (2) outcome improvement in tolerators we have described the benefits in limiting treatment dropout of a design with an active run-in period, followed by randomization of those who tolerated the treatment. However, in evaluating the results, it must be understood that the run-in period is part of the therapeutic strategy under study: consequently, for example, if the treatment has adverse effects during initial dosing, those risks would need to be assessed in other trials prior to approval.

In summary, the choice of outcome measure and estimand are crucial for clinical trial design and regulatory decision making. As the discussion above makes clear, there are a wide range of estimands that can be considered in a given situation, and each will involve tradeoffs between the representativeness of the population of study, the ease of study design and execution, and the sensitivity to missing data.

**Recommendation 1: The trial protocol should explicitly define (a) the objective(s) of the trial; (b) the associated primary outcome or outcomes; (c) how, when, and on whom the outcome or outcomes will be measured; and (d) the measures of intervention effects, that is, the causal estimands of primary interest. These measures should be meaningful for all study participants, and estimable with minimal assumptions. Concerning the latter, the protocol should address the potential impact and treatment of missing data.**

Given the resulting bias in the estimation of treatment effects caused by missing data, it is a serious concern that the actions recommended are not routine practice. However, it is our strong impression that these actions are not common, that rates of missing data remain high for a large fraction of trials, that protocols very often fail to devote attention to plans to com-

bat missing data, and that protocols are also often vague about the causal estimand. The failure to include a formal discussion of missing data in trial protocols should be viewed as a serious deficiency.

## MINIMIZING DROPOUTS IN TRIAL DESIGN

In this section, we describe a number of design elements for clinical trials that can help to reduce the number of participants who drop out due to lack of tolerability, lack of efficacy, or inability to provide the required measurement. The choice requires careful consideration because in some cases it will affect the generalizability of the study, that is, the population to which study conclusions are applicable. This section covers the following design elements: use of a run-in period or enrichment; flexible doses; target population selection; "add-on" studies; reduction of follow-up periods; allowing rescue medications; defining outcomes that can be ascertained in a high proportion of participants; and determining long-term efficacy in trials with randomized withdrawal.

***Use of Run-In Periods or Enrichment Before Randomization to Identify Participants Who Can Tolerate or Respond to the Study Treatment*** For studies in which the tolerability of treatments or adherence to study protocols is a concern, a run-in period can be used to establish short-term tolerability and adherence to the study treatment, followed by randomization of only those individuals who tolerated and adhered to therapy.

Such a design may result in a more efficient study with less missing data, but it likely will not adequately estimate the rate of adverse events in the broader population. Some clinical trials have also used a run-in period to identify participants who are likely to respond positively to the study treatment. This may also come at the cost of some external validity, reducing the ability to make estimates of the effectiveness of the treatment for the broader target population that might be given the treatment.

A related idea to run-in designs is the use of enrichment designs, which exclude participants based on initial indications that the response to the study treatment for them may be weaker or may be more difficult to tolerate. Enrichment designs have the advantage of clearly identifying the target population in advance of enrollment.[3]

***Flexible Dose (Titration) Studies*** Protocols that allow flexible dosing to accommodate individual differences in tolerability allow more participants to continue on the assigned treatment by reducing the frequency of dropout

[3]For a description of run-in periods and enrichment designs, see Institute of Medicine (2001) and Temple (2005).

because of adverse events or inadequate efficacy. Flexible-dose protocols are sometimes viewed as conflicting with the desire to assess the effects of specific doses, but giving investigators the flexibility to increase or decrease titration (when clinicians are allowed to individualize a patient's dosage) on the basis of a participant's ability to tolerate a drug may in fact be more reflective of real-life applications.

***Selection of Target Populations for Whom Treatment Is Indicated*** Participants who are doing well on their current treatments may not be good candidates for trial enrollment, in part because they are more likely to drop out because of lack of efficacy. Therefore, a good design approach is not to include participants who are receiving treatments that are proving effective in order to meet an enrollment target.

***Adding the Study Treatment to an Effective Treatment Considered to Be the Standard*** In many cases in which drug interactions are not a concern, dropout due to lack of efficacy can be reduced through the use of "add-on" study designs. In such a design, a new treatment or placebo may be added to an optimized background regimen for study participants. These designs may decrease the likelihood of missing data due to lack of efficacy.

***Reducing the Follow-Up Period*** Shorter follow-up periods may yield a reduction in dropouts, since fewer participants move out of the area, fewer develop intolerable adverse events, and the number and burden of clinical visits may be reduced. Modifying a trial design in this way may be preferable for assessing efficacy: essentially, it is trading off participants who respond more slowly to study treatment with participants who drop out early. Past experience of similar trials can provide guidance for evaluating this tradeoff in specific situations.

An alternative would be to define the primary outcome assessment for a shorter period of follow-up but retain a longer period of follow-up for safety and secondary outcome assessments. Use of shorter follow-up periods may be particularly useful with placebo control groups for which there are established effective active treatments. Longer follow-up periods clearly have advantages when short-term effects do not provide reliable assessments of the performance of an intervention, but the negative consequences in the form of missing data need to be recognized and taken into account.

***Allow Rescue Medication in the Event of Poor Response*** Dropout can be reduced by allowing alternative treatments for participants who are not responding to the study treatment. If this design option is adopted, the estimand and associated outcome measurements need to be carefully defined in the protocol. For example, time to treatment failure could include the

use of alternative therapy as an indicator of treatment failure. (In doing so, however, one needs to be careful to delineate the circumstances for switching to an alternative therapy in the trial protocol to support objective conclusions.)

***Define Outcomes That Can Be Ascertained in a High Proportion of Participants*** To avoid missing data caused by the use of outcomes that are undefined for some participants, it is useful to use primary outcomes that are ascertainable for all randomized participants.[4] This may require use of composite outcomes (e.g., outcomes that incorporate death as part of the outcome or incorporate use of rescue medication or surgery for initial poor response). At the analysis stage, for ordinal or for continuous outcomes, such events might be given a worst outcome rank. However, it is not always useful to use composite outcomes to avoid the occurrence of missing data, since composite outcomes can be difficult to interpret if individual components of the composite provide contrasting evidence about the intervention or if a weaker component dominates. In addition, primary outcome measures that require substantial invasive procedures (e.g., liver biopsies) are likely to result in significant missing data, and such outcome measures should be avoided whenever possible.

***Use of Randomized Withdrawal to Determine Long-Term Efficacy*** As noted above, trials with long-term follow-up may be more prone to missing data. In selected situations, this problem can be minimized by using randomized withdrawal designs. In such a design, all participants are initially treated with the intervention under study for a sufficiently long period to address the question of long-term efficacy. Only those who remain on and appear to have responded to therapy are then randomized for withdrawal or continuation. In cases in which loss of efficacy after withdrawal can be taken as evidence of drug efficacy, such a trial can generate long-term efficacy data.

**Recommendation 2: Investigators, sponsors, and regulators should design clinical trials consistent with the goal of maximizing the number of participants who are maintained on the protocol-specified intervention until the outcome data are collected.**

[4]One example is the use of the glomerular filtration rate to assess kidney performance for participants: some of them may progress to end-stage renal failure and undergo kidney transplantation.

## CONTINUING DATA COLLECTION FOR DROPOUTS

Even with careful attention to limiting missing data in the trial design, it is quite likely that some participants will not follow the protocol until the outcome data are collected. An important question is then what data to collect for participants who stop the assigned treatment. Sponsors and investigators may believe that the participants are no longer relevant to the study and so be reluctant to incur the costs of continued data collection. Yet continued data collection may inform statistical methods based on assumptions concerning the outcomes that participants might have had if they continued treatment. Continued data collection also allows exploration of whether the assigned therapy affects the efficacy of subsequent therapies (e.g., by improving the degree of tolerance to the treatment through exposure to a similar treatment, i.e., cross-resistance).

The correct decision on continued data collection depends on the selected estimand and study design. For example, if the primary estimand does not require the collection of the outcome after participants discontinue assigned treatment, such as with the estimand (4) above (area under the outcome curve during tolerated treatment), then the benefits of collecting additional outcome data after the primary outcome is reached needs to be weighed against the costs and potential drawbacks of the collection.

An additional advantage of data collection after subjects have switched to other treatments (or otherwise violated the protocol) is the ability to monitor side effects that occur after discontinuation of treatment. Although the cause of such side effects may be unclear (e.g., if a subject switches to another treatment), these data, when combined with long-term follow-up of other subjects in high-quality epidemiological studies, may help to determine treatment-associated risks that are not immediately apparent. We are convinced that in the large majority of settings, as has been argued by Lavori (1992) and Rubin (1992), the benefits of collecting outcomes after subjects have discontinued treatment outweigh the costs.

**Recommendation 3: Trial sponsors should continue to collect information on key outcomes on participants who discontinue their protocol-specified intervention in the course of the study, except in those cases for which a compelling cost-benefit analysis argues otherwise, and this information should be recorded and used in the analysis.**

**Recommendation 4: The trial design team should consider whether participants who discontinue the protocol intervention should have access to and be encouraged to use specific alternative treatments. Such treatments should be specified in the study protocol.**

**Recommendation 5: Data collection and information about all relevant treatments and key covariates should be recorded for all initial study participants, whether or not participants received the intervention specified in the protocol.**

## REFLECTING LOSS OF POWER FROM MISSING DATA

An important and relatively neglected issue in the design of clinical trials is how to account for the loss of power from missing data. (An additional impact of missing data is that the true significance level of the test of the treatment effect size could be larger than the specified level.) Currently, if any accommodation is done, it is simply to inflate the sample size that was initially planned to achieve a stated power by the inverse of one minus the anticipated dropout rate, as determined from other recent trials for similar interventions. If the dropouts provide no information about the treatment effect (which would not be the case for situations in which there was an interim outcome measure collected prior to participants' dropping out) and the data from dropouts are missing completely at random, then this approach is reasonable. However, in practice, dropouts may provide partial information about the treatment effect: that is, effects (or lack of effects) of the intervention often play a role in the decision to drop out. The missing completely at random assumption is generally too optimistic; therefore, power calculations should be based on more realistic missing at random or missing not at random assumptions. Under such assumptions, the effects of missing data on power cannot be easily assessed analytically: relatively involved simulation studies would be needed. This is rarely done and is an area for research.[5]

It should be added that the most worrisome effect of missing values on the inference for clinical trials is not the reduction of power, though that can be problematic, but biased estimation of the treatment effect. The bias from an unaccounted association between the indicator of missing values and the outcome of interest is not addressed by simply inflating the sample size. In particular, if the potential bias from missing data is similar in size to the anticipated size of the treatment effect, then detection of this effect is unlikely, regardless of the sample size chosen for the study. If some preliminary estimate of the potential nonresponse bias can be obtained,

[5]In some trials, consideration should be given to both potential nonadherence to the study treatment and also to the use of treatments, in the control group, that are similar to the study treatment (e.g., when the primary estimand is 1, above). This is more likely in a nonblinded study. Methods for inflating sample size to account for this type of noncompliance to the assigned treatments have been developed (see, e.g., Lakatos, 1988; Shih, 1995). In addition, it is also prudent at the design stage to inflate sample size to account for the fact that the primary outcome status may not be known for all study participants.

perhaps from a sensitivity analysis of the kind described in Chapter 5 on a related prior study, a simple strategy is to reduce the anticipated effect size by the anticipated size of the nonresponse bias and then power the study for this reduced effect size. If the adjusted effect size is too small to detect, it would be a strong incentive to design the study to reduce the degree of missingness.

## DESIGN ISSUES IN THE CASE STUDIES

We now return to the three case studies introduced in Chapter 1. These examples (chronic pain management, HIV, and mechanical devices for hearts) were used to illustrate how missing data arise in clinical trials. They are used in this section to illustrate how the design recommendations in this chapter can be carried out in these situations.

### Trials for Chronic Pain

Clinical trials for assessing interventions to relieve chronic pain are often subject to high rates of missing data because of inadequate efficacy and participants' inability to tolerate treatment. Participants who discontinue study treatment usually switch to a proven (approved) effective therapy, and it is typical for investigators to stop collecting pain response data on these individuals. The last observation carried forward approach is often used to impute missing outcome values.

#### Selection of (Causal) Estimand

As specified in Recommendation 1, a critical first step is to determine an appropriate estimand. Potential choices include

(a) (difference in) pain relief in all participants (e.g., degree of pain relief at 12 weeks [or more] in all patients in whom the treatment intervention is initiated [regardless of what is received throughout the course of the trial] [estimand 1, above]);
(b) (difference in) pain relief in tolerators (e.g., degree of pain relief in patients who tolerate and choose to receive 12 weeks of therapy [estimand 2, above]); and
(c) (difference in) treatment success rate (e.g., proportion of patients who can tolerate therapy, remain in study, and achieve adequate pain relief over 12 weeks [estimand 5, above]).

Option (a) addresses the anticipated outcomes in all patients who are randomized. Patients are managed according to a policy outlined in the

protocol and that reflects current practice. If the treatment policy reflects common practice in the clinical setting, this estimand may predict actual clinical outcomes. However, if many subjects receive effective alternative therapies, this estimand may shed only limited light on whether the treatment therapy is effective.

Option (b) addresses a key regulatory question, long-term efficacy in patients who will take the drug, but it fails to address other key questions, especially how well and how often the drug is tolerated and its efficacy in the total population receiving it, including those who do not take it for 12 weeks.

Option (c) addresses an important regulatory question and avoids missing data by defining a composite primary outcome. However, classifying all patients as either a treatment success or not may ignore important information, such as the extent of success or cause of failure. Also, counting patients who cannot tolerate therapy as failures may strongly weigh against drugs that are excellent in patients who tolerate them, even if there are significant subsets of patients who cannot tolerate them.

For example, in recent trials of trimethoprim sulfa (TS) against pentamidine for treatment of pneumocystis pneumonia in HIV-infected subjects, those who could not tolerate TS were typically switched to pentamidine. In evaluating TS, these treatment failures were "charged" to TS according to a traditional intent-to-treat analysis, ignoring the fact that almost all the TS failures were those who failed to tolerate the drug and not failures in efficacy, which is an important finding obscured by the composite outcome measure used. (For details, see Schneider et al., 1992.)

### Suggested Study Designs Paired with Estimands

For estimand (b) (pain relief in tolerators), two study designs that limit missing data are a randomized withdrawal design, in which patients are treated with the test treatment open-label for 12 weeks, and subjects who tolerate and have adequate response to the treatment are randomized to continue or to withdraw (e.g., are switched to placebo) and followed for some time, and a design that uses an active control run-in period followed by placebo washout, and then randomization of those patients who tolerated the active control and had relief. In this case, the outcome is pain control at 12 weeks.

These designs may limit missing data problems, but as previously noted they raise other issues, such as the inability to address safety and efficacy in all comers. However, it may well be best to address those questions in separate clinical trials rather than try to address all questions in one trial.

For estimand (c) (treatment success rate), one can include in the composite outcome patients who cannot tolerate therapy along with those who

have inadequate pain relief. In this approach, there will be minimal missing data for the primary outcome.

Finally, for the more traditional estimand a. (pain relief in all comers), there are some design alternatives that may help reduce the number of dropouts. For example, allowing dose modification will likely reduce dropouts in the treatment group because of inability to tolerate, since participants may tolerate a lower dose, or because of an inadequate response, since participants may respond adequately to a higher dose. Although such dose adjustments reflect common clinical practice, it is sometimes avoided in the regulatory setting due to requirements that are best addressed with fixed dosing regimens. (For example, if it is required to determine the minimal dose that provides significant response, dose adjustments may be inappropriate.) In such a case, it may be best nonetheless to allow dose adjustment to minimize missing data in trials that are estimating long-term pain relief in all comers and to address fixed dosing issues in a separate trial or trials. Finally, for this estimand, continuing to collect data through 12 weeks in all patients, including those who choose to switch therapies, and using these data in the analysis, is particularly important.

### Trials for Treatment of HIV

In many HIV trials, a noninferiority design is used to test whether a new drug is at least as safe and efficacious as the current standard of treatment. Since combination treatment is the norm for HIV, the typical design in this setting is new drug A plus background treatment compared with current drug B plus the same background treatment. A common primary outcome is often called time to loss of virologic response, but is in fact a composite measure that includes the following components: (1) death (and sometimes progression to an AIDS event), (2) discontinuation of study drug before 48 weeks, (3) loss to follow-up, and (4) HIV RNA level of greater than or equal to 50 copies/mL at or prior to 48 weeks on study drug.[6]

#### Suggested Estimands and Study Designs

Again, we emphasize the need to start by determining the causal estimand. Possible choices include the following:

(a) virologic response in all participants (e.g., the percentage with an HIV RNA level of less than 50 copies/mL after 48 weeks in all participants

[6]This composite outcome may be an example of the hazard mentioned in Chapter 1: if outcome measures are selected partially to reduce the frequency of missing data, they may also compromise the clinical value of the resulting inference.

randomized)—a "true" virologic outcome used for a comparison of treatment policies that ignores whether study treatment is discontinued;

(b) virologic response in tolerators (e.g., the percentage with an HIV RNA level of less than 50 copies/mL, among participants who were able to tolerate the treatment for 48 weeks (note that switches in antiretroviral therapy [ART] due to lack of efficacy need to be differentiated from switches due to side effects or lack of tolerability; and

(c) treatment success rate (e.g., the proportion of all randomized participants who stay on assigned treatment, remain in the study, and achieve an HIV RNA level of less than 50 copies/mL at 48 weeks), which is often the estimand in current practice.

Estimand (a) (virologic response in all participants) addresses anticipated outcomes in all participants who started on the therapy in question and were managed according to standard practice. This approach addresses a question about a specific efficacy outcome, and it compares two treatment policies (e.g., starting with a regimen using drug A with background treatment and starting with drug B with the same background treatment). This outcome is not often used in a regulatory setting because of concerns that estimation of the differences between drug A and drug B could be affected if more participants on one treatment group than another were switched to a virologically more potent regimen before 48 weeks.

Estimand (b) (virologic response in tolerators) addresses one key regulatory question, which is the efficacy in participants who will take the drug. However, it fails to address other key questions, for example, efficacy of the drug in the total population receiving it, including those who do not take it for 48 weeks. A run-in period is usually not practical because of concerns about HIV drug resistance. An analysis that excludes those who do not tolerate the study treatments may lead to biased estimates of treatment efficacy.

Estimand (c) (treatment success rate at 48 weeks) addresses an important regulatory question and avoids missing data through the use of a composite outcome. However, use of the composite outcome may mask important treatment differences, and in some circumstances may result in misleading results. For example, outcomes labeled as virologic failures may in fact reflect toxicity or losses to follow-up. Furthermore, counting participants who cannot tolerate therapy as failures may overly weigh against drugs that have excellent virologic efficacy in patients who do tolerate them, even if there are significant subsets of patients who cannot tolerate them.

### Study Designs That Minimize the Extent and Impact of Missing Data

The use of virologic response in all comers and the composite treatment success outcomes each have advantages and disadvantages. In the

regulatory setting, the latter is currently recommended. It avoids missing data by considering participants with missing data to be treatment failures. However, this choice of outcome gives equal weight to missing data, deaths, intolerance, and lack of virologic efficacy, creating difficulties in interpretation. If such an outcome is used, there may be advantages to continue to collect data after treatment discontinuation to the end of follow-up. These data may permit assessment of the consequences of treatment failure before 48 weeks due to intolerability or lack of virologic efficacy (e.g., the development of HIV drug resistance associated with virologic failure). Continued follow-up allows a separate assessment of each component of the composite outcome at or before 48 weeks (e.g., summaries of numbers assigned each treatment who failed virologically). Treatment policies, as in estimand 1, above, can also be compared.

### Trials for Mechanical Circulatory Devices for Severe Symptomatic Heart Failure

Device trials in patients with severe symptomatic heart failure have high rates of missing data for measures of functional status and health-related quality of life. The missing data arises because of deaths, some of which may be associated with the implantation procedure, failure to attend study follow-up examinations, and inability or unwillingness to perform functional tests or complete self-administered questionnaires. Unlike ascertainment of hospitalization events, measures of health status that include symptoms, functional status tests, and quality-of-life assessment require that patients be seen or complete a questionnaire.

In a trial of a left ventricular assist device (LVAD) as destination therapy for patients with severe symptomatic heart failure, it is critical to assess whether the device improves health status as well as survival. The ultimate approval and use of the device will depend on both outcomes. We consider possible estimands and study designs for assessing health status in a comparison of an LVAD with optimal medical management over a 2-year follow-up period. We assume that in such studies the four outcomes are of interest: death; disabling stroke; criteria met for implanting LVAD (e.g., based on previous trials in patients ineligible for transplant); and a self-administered quality-of-life assessment using a standard instrument (alternatively, or in addition to, a function measure test, such as a 6-minute walk time could be used). Some trials may also incorporate device removal or replacement in a primary composite outcome. Considerations in measuring and assessing health status in such trials have been summarized by a working group of the Heart Failure Society of America (see, e.g., Normand, 2005).

**Suggested Estimands and Study Designs**

Three potential choices for an estimand for evaluating health status include

(a) difference in quality of life between treatment groups for all randomized patients,
(b) difference in quality of life among survivors, and
(c) area under the quality-of-life curve while alive.

For the estimand (a) (difference in quality of life for all randomized patients), the quality-of-life comparison could be performed earlier than 2 years to maximize the number of patients in each treatment group under follow-up (e.g., at 6 months). Alternatively, patients who die or who are unable to complete the questionnaire for health reasons could be given a "worst rank" score. This latter strategy would likely affect the power of the resulting test statistic as some deaths would be expected to be unrelated to the treatments.

Estimand (b) (difference in quality of life among survivors) addresses the effect of the LVAD on an outcome that complements a composite outcome of death, disabling stroke, or progression to a specified criteria that indicate an LVAD should be implanted. However, a complete-case analysis or mixed-model approach to the analysis of the quality-of-life data may not be appropriate as it is unlikely the data are missing at random. Thus, the pattern of missing data should be considered, and other methods for modeling the missing data (e.g., pattern mixture models) should be used.

Estimand (c) (area under the quality of life curve while alive) has the advantage of simultaneously evaluating the LVAD for quality of life and duration of survival (or an expanded event-free outcome).

Irrespective of which estimand is used, in such trials it will be important to assess health status as objectively as possible. Since such trials cannot be blinded to patients or investigators caring for the patient, use of independent, trained evaluators (i.e., those not involved in the care of the patient), for whom it would be possible to be blinded to the treatment, should be considered. In addition, assessment by both patients' self-reports and clinicians' assessments should be considered. Prior to the initiation of a study, the importance of complete ascertainment of the health status measurements should be emphasized to patients in the informed consent process and to clinicians during protocol training. Also, plans should be developed for visiting patients to obtain the health status assessments if they cannot attend clinic examinations.

# 3

# Trial Strategies to Reduce the Frequency of Missing Data

This chapter discusses a number of strategies that can be applied to reduce the amount of missing data during trial implementation and conduct. The approaches in this chapter deal with the practical aspects of trial conduct, rather than the more fundamental design aspects covered in Chapter 2. We classify trial strategies into two types: (1) actions for design and management teams and (2) actions for investigators and site personnel. Before turning to those strategies, we briefly comment on the research literature on clinical trial dropouts, and we end the chapter with a look at setting and meeting targets for missing data.

## REASONS FOR DROPOUTS

The literature on the factors associated with and the effectiveness of various measures to reduce the occurrence of missing outcome values is relatively diffuse, possibly dependent on the medical condition, the intervention under study, and the population of interest. Thus, the literature is difficult to summarize and often not a great deal is known for particular situations. However, several statements can be supported: there is a lack of consensus regarding how to measure dropout rates; dropout can often be very substantial, sometimes more than 30 percent; and the rate of missing outcome data can sometimes be substantially reduced by applying some of the ideas suggested in Chapter 2 and below (see, e.g., Sprague et al., 2003; Oleske et al., 2007; Robinson et al., 2007; Snow et al., 2007; Warden et al., 2007; Williams et al., 2008).

## ACTIONS FOR DESIGN AND MANAGEMENT TEAMS

This section discusses some techniques that trial design and management teams can use to reduce the frequency of dropouts. First, designers and managers can limit participants' burden and inconvenience in the data collection stage. This can be done in at least five ways: (1) minimizing the number of visits and assessments, (2) collecting only the information that is needed at each visit, (3) using user-friendly case report forms, (4) using direct data capture that does not require a clinic visit whenever feasible, and (5) allowing a relatively large time window for each follow-up assessment. Examples of information not needed at each visit include aspects of the participant's medical history and contact information that were provided at earlier visits and information available from medical records. The overall aim is to balance the competing goals of reducing response burden and collecting sufficient information to fully support the analytic goals and to guide the next steps in treatment. (Regarding use of direct data capture to minimize the response burden, it would also be useful to attempt to collect whatever information is available from administrative records.)

Second, design and management teams can increase the incentives for participation and completion by the provision of effective treatments to participants after the trial. Such incentives might include continued access to effective study treatments on extension protocols until the treatment is licensed.

Third, designer and managers can select investigators with a good track record of both enrolling and following participants and collecting complete data in previous trials, and provide good training. The training (and on-study reinforcement) needs to emphasize the importance of complete data collection and the difference between discontinuing the study treatment and discontinuing data collection. Training should stress the value of collecting data after a participant discontinues the study (or the control) treatment.

As discussed in Chapter 2, many trial sponsors and investigators mistakenly assume that there is little reason for additional data collection when participants discontinue study treatment. But as we emphasize, the continued collection of data is important in many trials. The trial objectives and estimands need to be considered.

Training can also emphasize the importance of the informed consent process as a mechanism for ensuring that participants understand the commitment they are making, including their intent to complete the trial regardless of the treatment they are receiving. Training of investigators and research staff should also emphasize how to work with participants to minimize the extent of missing data (see later in this chapter). Finally, trainers need to know and explain to participants that any decision to withdraw consent is a participant's decision, not the investigator's decision. However,

when participants are dissatisfied with the conduct of the trial but have not yet withdrawn, the investigator should make an effort to address their concerns and retain them in the trial, rather than simply indicating that the participants withdrew consent. In doing so, investigators must be careful that their efforts do not cross over into coercion.

Fourth, designers and managers need to consider how investigators are paid. Paying investigators solely by the number of participants enrolled should be avoided because it places too much emphasis on enrollment and not enough on follow-up; payments should also reflect follow-up work (e.g., payment per visit or procedure). In addition, linking some additional compensation to the completeness of the data collection should be considered. It is acceptable and generally advisable to link a final payment to completion of forms at a study closeout visit (i.e., a final visit at the end of follow-up to assess the participant's status). However, care must be taken on this point. Providing extra compensation to investigators for encouraging participants to complete a study when the participants are thereby exposed to significant additional risks could create a conflict of interest on the part of the investigator and would therefore be unethical. But if there are minimal risks associated with data collection to the participant, it may be acceptable to provide financial incentives to the investigator to continue to collect data, whether or not the participant continues treatment.

Fifth, designers and managers can ensure that data collection is monitored and reported during the trial. Missing data and missed visits that could affect important outcomes need to be assessed in real time by site personnel during a clinical trial. The information from these assessments should be available and shown to investigators at regular meetings and on study websites, creating a climate to encourage other investigators to collect more complete data. Also, identification of poorly performing sites can help identify the need for some sort of remediation, including additional training, site visits, or even site closure. Site visits should be targeted on the basis of assessments of the amount of missing data, with the goal of helping to correct the problem.

## ACTIONS FOR INVESTIGATORS AND SITE PERSONNEL

Investigators and site personnel can also act in several ways to reduce the amount of missing data. First, in the informed consent process, they can emphasize to participants the importance of continued participation for the full duration of the trial. Similarly, they can ensure that the trial procedures allow for an informed withdrawal of consent so that participants recognize the importance of continued follow-up for data collection if they discontinue study treatment: see Box 3-1 for an example of language for withdrawal. Second, investigators and site personnel can provide incentives

**BOX 3-1**
**Example of Language for Withdrawal of Informed Consent**

- I no longer wish to take trial anti-HIV drugs but I am willing to attend follow-up visits.
- I no longer wish to take trial anti-HIV drugs and do not wish to attend further visits. I agree to my medical records being consulted in future to obtain clinical information for the Development of AntiRetroviral Therapy (DART) in Africa.
- I no longer wish to take trial anti-HIV drugs and do not wish to attend further visits. I do not agree to my medical records being consulted in future to obtain clinical information for DART.

for participants. In general, paying for voluntary participation in a clinical trial is regarded as ethical (see, e.g., Emmanuel, 2005). When compensation is to be provided, the Code of Federal Regulations requires that the responsible Institutional Review Board (IRB) ensure that the compensation is neither coercive nor at the level that would present undue influence (21 CFR 50.20). Providing cash is generally not viewed as being coercive, as it is a benefit. Most IRBs allow cash payments to be slightly backloaded (retaining a small proportion as an incentive for completion), but, generally, payments accrue as a study progresses in payment for participation activities that are completed. Compensating people for taking risks is not uncommon, and as noted it is generally acceptable if not judged as coercive. Payments for return visits of participants who have stopped taking medication are virtually always considered ethical, since the risk to the participant is zero or minimal. Study-branded gifts are also ethical and may have the added effect of increasing the participant's engagement with the trial.

Third, investigators and site personnel can collect information on which participants are at risk for dropping out and why: formal "intent-to-attend" questioning may help to identify reasons for dropout (see, e.g., Leon et al., 2007) and may yield useful covariates in missing data models. Factors influencing decisions to participate include: (i) time and duration of visits, (ii) need for assistance with transportation or child care, (iii) need for reminders, (iv) problems in relations with the staff, (v) problems with blood drawing or other procedures, (vi) side effects, and (vii) perceptions of intervention efficacy.

Fourth, investigators and site personnel can educate participants on the importance of continued engagement in the trial in order to help contribute

to important scientific knowledge. Mechanisms for such education include the production of a study newsletter, maintenance of a regularly updated website for trial participants, and providing access to interim papers and presentations on study progress and findings. (We note that IRBs may require approval for some communications with study participants.)

Fifth, investigators and site personnel can increase participants' engagement and retention in the study by such mechanisms as study-branded gifts; regular expressions of thanks, both verbal and written; social networking; and solicitation of input regarding relevant issues of study conduct. Other ways to encourage participation and involvement include reminders before a visit and after missed visits help to encourage participation.

Sixth, investigators and site personnel can make participation enjoyable in many ways, including: (i) development of a welcoming environment, (ii) hiring of friendly staff, (iii) operational practices that are respectful of participants' time and schedules, (iv) availability of on-site diversions for small children, and (v) valued education at the site.

Seventh, investigators and site personnel can ensure that participants' contact information is updated at each visit, recognizing that in some studies, home visits may be needed to keep all contact information current. The provision of transportation and child care costs can also improve retention.

For participants who want to discontinue treatment, it is important for site personnel to determine the reasons and to make sure that the participants understand the importance of continuing on the study for the purpose of data collection. If participants switch to alternative treatments due to intolerance, it is important for investigators and site personnel to document the changes because they may be useful in summarizing the study results.

**Recommendation 6: Study sponsors should explicitly anticipate potential problems of missing data. In particular, the trial protocol should contain a section that addresses missing data issues, including the anticipated amount of missing data, and steps taken in trial design and trial conduct to monitor and limit the impact of missing data.**

**Recommendation 7: Informed consent documents should emphasize the importance of collecting outcome data from individuals who choose to discontinue treatment during the study, and they should encourage participants to provide this information whether or not they complete the anticipated course of study treatment.**

## TARGETS FOR ACCEPTABLE RATES OF MISSING DATA

Although some missing data should be anticipated for every clinical trial, levels that are unacceptable given the design should be considered in

writing the protocol. One way to set target rates and maximally acceptable rates for missing data would be to use the results from similar trials to help determine what is reasonably achievable and did not excessively impact study conclusions and to determine how missing data can be further minimized. For example, using the findings from completed trials, some percentile could be used for the target and some higher percentile for the maximally acceptable value. Another possibility, which is likely not currently feasible but would be after a sufficient number of effective sensitivity analyses have been carried out (depending on the characteristics of the trials) is to observe what rates of missing values, in trials in which the primary analysis demonstrated a significant benefit, resulted in alternative analyses in which the treatment effect was no longer significant. In this way, one could try to limit the amount of missing data to ensure that a sensitivity analysis would not contradict the findings of the primary analysis.

Once goals are established, performance against these goals can be monitored, and the goals can be used to motivate investigators. Comparison of targets and current rates of missing data could also be used by a data monitoring committee to halt a trial for underperformance.

Establishing reasonable goals and adhering to them may not be an easy task for several reasons: (1) there may not be many similar trials to use as a basis for acceptable levels of missing data; (2) it may be difficult to determine the steps that trial investigators took to reduce missing data; and (3) it may be difficult to determine what, if any, sensitivity analyses were carried out for trials conducted by other sponsors. Nevertheless, it is important to set high standards for the participating investigators and monitor the amount of missing data for key outcomes in real time.

We note that one cannot be specific as to how to set target and maximally acceptable rates for missing data in all clinical trials. The amount of acceptable missing data will depend on many characteristics of the trial, including whether the assumption that the missing data are missing at random is reasonable, the size of the anticipated effect of the intervention under study, and the likelihood that a sensitivity analysis would render the results of the trial inconclusive. Applied research is needed on this topic, and techniques will need to evolve on the basis of that research.

**Recommendation 8: All trial protocols should recognize the importance of minimizing the amount of missing data, and, in particular, they should set a minimum rate of completeness for the primary outcome(s), based on what has been achievable in similar past trials.**

Finally, to monitor the degree of missing outcome data, the data and safety monitoring board (DSMB) for a trial should be aware of the trial's target for missing data, and the investigators should report to the DSMB

how they are doing relative to the target. If they are not doing well, the DSMB should discuss the issue with them. However, primary responsibility for ensuring that missing data are kept to a minimum should reside with the investigators, the protocol team, and the sponsor.

# 4

# Drawing Inferences from Incomplete Data

In this chapter, we review and comment on several approaches for drawing inferences from incomplete data. A substantial literature on this topic has developed over the last 30 years, and the range of approaches to modeling and inference is extremely broad. We make no attempt here to summarize that entire literature; rather, we focus on those methods that are most directly relevant to the design and analysis of regulatory clinical trials. We begin by presenting a set of principles for drawing inference from incomplete data. A major theme that we reiterate throughout the chapter is that inference from incomplete data relies on subjective, untestable assumptions about the distribution of missing values. On its face, this statement seems obvious. However, for a number of commonly used methods, users are not always aware of the assumptions that underlie the methods and the results drawn from applying them. This lack of awareness is particularly true of single imputation methods—such as last or baseline observation carried forward (LOCF or BOCF) and random effects (mixed effects) regression models—that rely on strong parametric assumptions.

In the second section of the chapter, we introduce a set of notation that is used throughout (and in Chapter 5). The third section summarizes the assumptions that underlie inference from incomplete data (missing completely at random, missing at random, etc.). The remaining sections describe commonly-used methods of analysis and offer comments and recommendations about their use in practice. In some cases, we offer recommendations for further research and investigation.

For both this chapter and the next, it is important to note the role of software. None of the techniques for either the primary analysis of clini-

cal trial data or for the subsequent sensitivity analysis that are described in the next chapter can be widely used, either at the U.S. Food and Drug Administration (FDA) or by trial sponsors, unless they are made available in one or more of the standard statistical software packages. It is beyond the scope of this report to describe and review specific software packages or routines. Many of the commonly used commercial and open-source packages used in the analysis of trials for the regulatory setting (SAS, SPSS, Stata, and R) allow for the analysis of incomplete data, using methods such as direct likelihood, Bayesian analysis, generalized estimating equations, inverse probability weighting, and multiple imputation.

Statistical software is evolving at a rapid pace to keep up with new developments in methodology and to implement proven methods. However, although progress is being made, the current suite of available tools remain lacking regarding augmented inverse probability weighting (IPW), missing not at random (MNAR) models, and analysis of the sensitivity to assumptions concerning the mechanism for missing outcome data. Given the urgency of the greater application of MNAR models and sensitivity analysis, we encourage the development and release of software tools to address these deficiencies. We again emphasize the importance of understanding and communicating the assumptions underlying analyses that are implemented in whatever software package is being used to draw inference about treatment effects. In most cases, communication of this information will necessitate referring to technical documentation for a specific analysis routine or procedure.

## PRINCIPLES

There is no universal method for handling incomplete data in a clinical trial. Each trial has its own set of design and measurement characteristics. There is, however, a set of six principles that can be applied in a wide variety of settings.

First, it needs to be determined whether missingness of a particular value hides a true underlying value that is meaningful for analysis. This may seem obvious but is not always the case. For example, consider a longitudinal analysis of CD4 counts in a clinical trial for AIDS. For subjects who leave the study because they move to a different location, it makes sense to consider the CD4 counts that would have been recorded if they had remained in the study. For subjects who die during the course of the study, it is less clear whether it is reasonable to consider CD4 counts after time of death as missing values.

Second, the analysis must be formulated to draw inference about an appropriate and well-defined causal estimand (see Chapter 2). The causal estimand should be defined in terms of the full data (i.e., the data that were

intended to be collected). It is important to distinguish between the estimand and the method of estimation, the latter of which may vary depending on assumptions.

Third, reasons for missing data must be documented as much as possible. This includes full and detailed documentation for each individual of the reasons for missing records or missing observations. Knowing the reason for missingness permits formulation of sensible assumptions about observations that are missing, including whether those observations are well defined.

Fourth, the trial designers should decide on a primary set of assumptions about the missing data mechanism. Those primary assumptions then serve as an anchor point for the sensitivity analyses. In many cases, the primary assumptions can be missing at random (MAR) (see Chapter 1). Assumptions about the missing data mechanism must be transparent and accessible to clinicians.

Fifth, the trial sponsors should conduct a statistically valid analysis under the primary missing data assumptions. If the assumptions hold, a statistically valid analysis yields consistent estimates, and standard errors and confidence intervals account for both sampling variability and for the added uncertainty associated with missing observations.

Sixth, the analysts should assess the robustness of the treatment effect inferences by conducting a sensitivity analysis. The sensitivity analysis should relate treatment effect inferences to one or more parameters that capture departures from the primary missing data assumption (e.g., MAR). Other departures from standard assumptions should also be examined, such as sensitivity to outliers.

## NOTATION

Throughout this and the next chapter, we use the following conventions. Let $X$ represent treatment indicators and baseline (i.e., pretreatment) covariates that are fully observed and conditioned on in the primary statistical analysis (such as study center and stratification variables). Another way to characterize $X$ is as the *design variables* that would be adjusted for or conditioned on in the final analysis. Let $Y$ denote the primary outcome variable, which may be a single outcome, a vector of repeated measurements, or a time to event. Auxiliary variables are denoted by $V$; these variables are distinct from design variables $X$ and may represent individual-level characteristics (either pre- or posttreatment) that aid in drawing inference from incomplete response data. Information on compliance or side effects of treatments that may be useful for modeling the missing data but are not included in the primary analytic model may be included in $V$. (We note that the collection and use of all available covariate information that is predic-

tive of the outcome in the full data model, and the occurrence of missing outcome data in the missing data model, is important and can dramatically improve the associated inference.)

In the absence of missing data, let $Z$ denote the values of $(V,Y)$ for an individual participant. For simplicity, we assume throughout that observations on $(V,Y)$ are independent within levels of $X$.

To distinguish between missing and observed data, let $M$ denote the indicator of whether $Y$ is missing. In repeated measures studies, we include a subscript for repeated measures. That is, if the intended outcome measures are $Y = (Y_1,Y_2,\ldots,Y_K)$, the corresponding missingness indicators are $M = (M_1,M_2,\ldots,M_K)$, where $M_j = 1$ if $Y_j$ is missing, and $M_j = 0$ if it is observed. We generally will assume that $Y$ and $V$ have the same missing data pattern, though in practice this restriction can be relaxed.

In many situations, missing values can be denoted by a single value, such as $M = 1$; in other settings, it may be useful to allow more than one missing-value code to indicate different types of missing data, such as $M = 1$ for lack of efficacy, $M = 2$ for inability to tolerate a drug because of side effects, $M = 3$ for a missed clinic visit, and so on. This notation allows for different modeling assumptions for the different causes of missing data.

## ASSUMPTIONS ABOUT MISSING DATA AND MISSING DATA MECHANISMS

The general missing data taxonomy described in this section is fully presented in Rubin (1976) and Little and Rubin (2002). Elaboration on the sequential versions of these for longitudinal data can be found in Robins et al. (1995) and Scharfstein et al. (1999). Discussion of the more general notion of *coarsening* can be found in Heitjan (1993) and Tsiatis (2006).

### Missing Data Patterns and Missing Data Mechanisms

It is useful to distinguish the *pattern* of the missing data from the missing data *mechanism*. The pattern simply defines which values in the data set are observed and which are missing, as described for an individual by the vector of indicators $M = (M_1,\ldots,M_K)$. Some methods for handling missing data apply to any pattern of missing data; other methods assume a special pattern.

A simple example of a special pattern is univariate missing data, where missingness is confined to a single variable. Another special pattern is *monotone missing data*, where the variables can be arranged so that $Y_{j+1}$ is missing for all cases where $Y_j$ is missing. This pattern commonly arises in longitudinal data, when the sole cause of missingness is attrition or dropouts, and there are no intermittently missing values.

The missing data *mechanism* relates to why values are missing and the connection of those reasons with treatment outcomes. The missing data mechanism can be represented in terms of the conditional distribution $[M \mid X,V,Y]$[1] for the missing data indicators given the values of the study variables that were *intended* to be collected. To emphasize that this distribution may depend both on observed and missing values of $V$ and $Y$, this is sometimes written as $[M \mid X,V_{obs},V_{mis},Y_{obs},Y_{mis}]$.

### Missing Completely at Random

Missing data are missing completely at random (MCAR) if missingness does not depend on values of the covariates, auxiliary and outcome variables $(X,V,Y)$. That is,

$$[M \mid X,V_{obs},V_{mis},Y_{obs},Y_{mis}] = [M]. \quad (1)$$

MCAR is generally a very strong assumption, unlikely to hold in many clinical trials. Situations in which MCAR might be plausible include *administrative censoring*, when outcomes are censored because a study is terminated at a planned date, and the outcome has not yet occurred for late accruals; and *designed missing data*, when expensive or onerous measurements are recorded only for a random subsample of participants. A closely related concept is conditional MCAR, which allows for the independence of the missing values, but is conditional on covariates $X$. Finally, it is useful to mention that MCAR is unique in that one can test whether the missing outcomes are MCAR if they are at least missing at random, which is discussed below.

### Missing at Random

A more realistic condition than MCAR for many studies is MAR, which requires that missingness is independent of missing responses $Y_{mis}$ and $V_{mis}$, conditionally on observed responses $(Y_{obs},V_{obs})$ and covariates $X$. That is,

$$[M \mid X,V_{obs},V_{mis},Y_{obs},Y_{mis}] = [M \mid X,V_{obs},Y_{obs}]. \quad (2)$$

If $Y$ and $V$ are considered to be random variables with distributions based on a model, then one can show that condition (2) is equivalent to

$$[Y_{mis},V_{mis} \mid X,V_{obs},Y_{obs},M] = [Y_{mis},V_{mis} \mid X,V_{obs},Y_{obs}], \quad (3)$$

[1]The notation $[a \mid b, c]$ (e.g., $[M \mid X,V,Y]$ ) is used to denote the conditional distribution of $a$ given the joint distribution of $b$ and $c$.

which implies that the predictive distribution of the missing variables given the observed variables does not depend on the pattern of missing values. This version of MAR is relevant from an analysis perspective because it characterizes the predictive distribution of the missing values, which is the basis for principled methods of imputation.

As we describe below, many standard analysis methods for incomplete data operate under the MAR assumption. It is therefore imperative that both the MAR assumption and the assumptions underlying the full data model (e.g., multivariate normality) be thoroughly justified before results from these models can be considered valid for treatment comparisons. In general: (a) even under MAR, different assumptions about the full data model will lead to different predictive distributions; (b) with incomplete data, assumptions about both the missing data mechanism and the full data model are unverifiable from the data; and (c) nevertheless, inference and therefore decisions about treatment effect often crucially depend upon them.

### MAR for Monotone Missing Data Patterns

With longitudinal repeated measures, and even for event time outcomes, the MAR assumption is not always intuitive for a general pattern of missing values.

However, it has a simple interpretation in the case of *monotone* missing data, such as that caused by dropouts. Suppose the data intended to be collected comprise repeated measures on an outcome $Y$, denoted by $Y_1,\ldots,Y_K$. Let $M_j = 1$ if $Y_j$ is missing, and let $M_j = 0$ if $Y_j$ is observed. Under monotone missingness, if observation $j$ is missing ($M_j = 1$), then all subsequent observations also are missing ($M_{j+1} = \ldots = M_K = 1$).

At any given time $j$, let $Y_j^- = (Y_1,\ldots,Y_{j-1})$ denote the history of measurements up to but not including time $j$, and let $Y_j^+ = (Y_j,\ldots,Y_K)$ denote the future measurements scheduled, including and after time $j$. At time $j$, the predictive distribution of future values given the observed history is denoted by $[Y_j^+ \mid Y_j^-,X,M_j = 0]$. The MAR condition holds if predictions of future measurements for those who drop out at time $j$ are equivalent in distribution to predictions for those who have observed data at and after time $j$. Formally, MAR is equivalent to

$$[Y_j^+ \mid Y_j^-,X,M_j = 1] = [Y_j^+ \mid Y_j^-,X,M_j = 0]. \quad (4)$$

Hence, under MAR, missing values at time $j$ and beyond can be predicted sequentially from the histories of participants still in the study at time $j$.

MAR for monotone missing data patterns also can be written in terms of the probability of dropouts at each measurement occasion. At time $j$,

the dropout probability is $P(M_j = 1 \mid M_{j-1} = 0)$. In general, this probability can depend on any aspect of the observations intended to be collected. MAR states that the dropout probability can only depend on observed data history,

$$P(M_j = 1 \mid M_{j-1} = 0, Y_j^-, Y_j^+, X) = P(M_j = 1 \mid M_{j-1} = 0, Y_j^-, X). \quad (5)$$

This representation shows that one can think of the MAR assumption as a *sequentially random dropout process*, where the decision to drop out at time $j$ is like the flip of a coin, with probability of 'heads' (dropout) depending on the measurements recorded through time $j - 1$.

Both (4) and (5) can be generalized by allowing the past measurements to include auxiliary covariates. Specifically, let $Z_j^- = (Y_1,\ldots,Y_{j-1},V_1,\ldots,V_{j-1})$ denote the observed history of both outcomes and auxiliaries. Then MAR can be restated by replacing $Y_j^-$ with $Z_j^-$ in (4) and (5). In fact, the MAR assumptions (4, 5) change depending on the set of auxiliary variables $V$ included in the analysis. The validity of the MAR assumption can be improved by measuring and including auxiliary variables that are predictive of whether the outcome variables are missing and predictive of the values of the missing variables.

### Missing Not at Random

MAR will fail to hold if missingness or dropout depends on the values of missing variables after conditioning on the observed variables. When MAR fails to hold, missing data are said to be MNAR.

For a monotone missing data pattern, missingness will be MNAR if there exists, for any $j$, at least one value of $Z_j^-$ for which

$$[Y_j^+ \mid Z_j^-, X, M_j = 1] \neq [Y_j^+ \mid Z_j^-, X, M_j = 0], \quad (6)$$

or equivalently, there exists, for any $j$, at least one value of $Y_j^+$, such that

$$P(M_j = 1 \mid M_{j-1} = 0, Z_j^-, Y_j^+, X) \neq P(M_j = 1 \mid M_{j-1} = 0, Z_j^-, X). \quad (7)$$

For (6), the consequence of MNAR is that the prediction of future observations for those who drop out cannot be reliably predicted using data observed prior to dropping out; or, that the distribution $[Y_j^+ \mid Z_j^-, X_j^-]$ differs between those who do and do not drop out at time $j$. Because these differences cannot be estimated from the observed data, they are entirely assumption driven. This is the central problem of missing data analysis in clinical trials.

### Example: Hypertension Trial with Planned and Unplanned Missing Data

Murray and Findlay (1988) describe data from a large multicenter trial of metopropol and ketanserin, two antihypertensive agents for patients with mild to moderate hypertension, with diastolic blood pressure as the outcome measure of interest. The double-blind treatment phase lasted 12 weeks, with clinic visits scheduled for weeks 0, 2, 4, 8, and 12. The protocol stated that patients with diastolic blood pressure exceeding 110 mmHg at either the 4- or 8-week visit should "jump" to an open follow-up phase—a form of planned dropout. In total, 39 of the 218 metopropol patients and 55 of the 211 ketanserin patients jumped to open follow-up.

In addition, 17 metopropol patients and 20 ketanserin patients had missing data for other reasons, including side effects. Analyses of the observed data clearly showed that those with missing blood pressure readings differed systematically from the patients who remained in the study, as would be predicted by the protocol for jumping to the open phase. This example provides an illustration of the importance of defining what is represented by a missing outcome. For the participants who were removed from protocol, it is possible to treat the missing values as values that would be observed had the individuals remained on treatment. The mechanism for those with missing values is MAR because missing outcomes resulted from the value of a recorded intermediate outcome variable for blood pressure, and are therefore a function of an observed value.

### Summary

1. Inferences from incomplete data, whether model-based or not, rely on assumptions—known as missing data mechanisms—that cannot be tested from the observed data.

2. A formal taxonomy exists for classifying missing data mechanisms, including for longitudinal and event history data. The mechanisms can be classified as MCAR, MAR, and MNAR.

3. Missing data mechanisms describe the relationship between the missing data indicator(s) $M$, the full outcome data $Y = (Y_{obs}, Y_{mis})$, design variables $X$, and auxiliary covariates $V$. Traditionally, these assumptions characterize restrictions on the distribution of $M$ given $(Y_{obs}, Y_{mis}, X, V)$. Each has an equivalent representation in terms of the predictive distribution of missing responses, namely $Y_{mis}$ given $(M, Y_{obs}, X, V)$.

## COMMONLY USED ANALYTIC METHODS UNDER MAR

Three common approaches to the analysis of missing data can be distinguished: (1) discarding incomplete cases and analyzing the remainder

(complete-case analysis); (2) imputing or filling in the missing values and then analyzing the filled-in data; and (3) analyzing the incomplete data by a method that does not require a complete (i.e., a rectangular) data set. Examples of (3) include likelihood-based methods, such as maximum likelihood (ML), restricted ML, and Bayesian methods; moment-based methods, such as generalized estimating equations and their variants; and semiparametric models for survival data, such as the Cox proportional hazards model. Multiple imputation (Rubin, 1987; Little and Rubin, 2002), an extension of single imputation that allows uncertainty in the imputations to be reflected appropriately in the analysis, is closely related to Bayesian methods (discussed later in this chapter).

### Deletion of Cases with Missing Data

A simple approach to missing data is complete-case analysis, also known as listwise deletion, in which incomplete cases are discarded and standard analysis methods are applied to the complete cases. In many statistical packages, it is the default analysis.

Although it is possible to list conditions under which an analysis of complete cases provides a valid inference (essentially, conditional MCAR), this method is generally inappropriate for a regulatory setting. When missingness is in the outcome, the MAR assumption is generally weaker and can reduce bias from deviations from MCAR by making use of the information from incomplete data. Furthermore, when missingness is appreciable, rejection of incomplete cases will involve a substantial waste of information and increase the potential for significant bias.[2]

In addition, if data are not collected after withdrawal from treatment, then the MAR assumption relies only on information accumulated while subjects are on treatment. Hence, any method that relies on MAR is estimating the mean under the condition that everyone had remained on treatment. This generally will not provide a valid estimator of the intention-to-treat effect. On the other hand, if data are collected after withdrawal from treatment, this information can be used either within inverse probability weighting (IPW) or in an imputation context to estimate an intention-to-treat effect under MAR (Hogan and Liard, 1996). It is for this

[2]When data are not MCAR, the bias of complete-case analysis depends on the degree of deviation from MCAR, the amount of missing data, and the specifics of the analysis. In particular, the bias in estimating the mean of a variable is the difference in the means for complete and incomplete cases multiplied by the fraction of incomplete cases. Thus, the potential for bias increases with the fraction of missing data. With respect to regression models, complete-case analysis yields valid inferences in regression if the model is correctly specified and missingness depends on the predictor variables, observed or missing, but not on the outcome. (For details, see Little and Rubin [2002].)

reason that we emphatically recommend aggressive collection of outcome data after individuals withdraw from treatment.

## Inverse Probability Weighting

### Univariate Outcome

When data are MAR but not MCAR, a modification of complete-case analysis is to assign a sampling weight to the complete cases. This tends to reduce bias, to the extent that the probability of being observed is a function of the other measured variables. Consider the simple case in which the intended outcome is $Y$, the design variables are $X$, and some auxiliary variables $V$ are available. As usual, $M = 1$ indicates that $Y$ is missing. To describe IPW, it is useful to introduce a response indicator, $R = 1 - M$, such that $R = 1$ when $Y$ is observed and $R = 0$ when it is missing.

An IPW estimator for the mean of $Y$ can be computed as follows:

1. Specify and fit a model for $\pi(X,V,\theta) = P_\theta(R = 1 \mid X,V)$, for example using logistic regression.

2. Estimate the mean of $Y$ using the weighted average

$$\hat{\mu} = (1/n)\sum_i \frac{R_i Y_i}{\pi\left(X_i, V_i, \hat{\theta}\right)}; \tag{8}$$

that is, the average of the observed $Y$ weighted inversely by the probability of being observed.

3. Standard error estimators can be computed analytically or by bootstrap methods. (For details on the bootstrap estimator of variance, see Efron and Tibshirani, 1993.)

For large samples, this method properly adjusts for bias when the data are MAR, provided the model for $\pi(X,V,\theta)$ is correctly specified. In finite samples, the method can yield mean estimates that have high variance when some individual-specific weights are high (i.e., when $\pi$ is close to zero). An alternative is to create strata based on the predicted probability of being complete and then weight respondents by the inverse of the response rate within these strata. Strata can be chosen to limit the size of the weights and hence control variance.

In addition to the MAR assumption, the IPW method requires two other key assumptions: (1) there are no covariate profiles $(X,V)$ within which $Y$ cannot be observed and (2) the support of the missing data distri-

bution is the same as that for the observed data distribution. Technically, (1) stipulates that $P(R = 1 \mid X,V) > 0$ for all possible realizations of $(X,V)$. A potential restriction imposed by (2) is that individual missing values cannot be imputed outside the range of observed values.

### IPW Regression for Repeated Measures

With repeated measures, a convenient way to estimate the treatment effect is through a regression model for the mean of the outcome vector conditional on the design variables $X$. With fully observed data, repeated measures regression models can be fit using generalized estimating equations (GEE) (Zeger and Liang, 1986).

With fully observed data, a desirable property of regression parameter estimates from GEE is that they retain such properties as consistency and asymptotic normality regardless of the assumed within-subject (longitudinal) correlation structure. When data are missing, this property no longer holds, and regression estimates may depend strongly on the assumed correlation structure (see Hogan et al., 2004, for an empirical example).

When missingness is MAR and follows a monotone pattern, the IPW method can be used to obtain consistent estimates of regression parameters using a specified procedure. Here, we emphasize that auxiliary information should be included in the observed-data history, $Z_j^- = (Y_1,\ldots,Y_{j-1},V_1,\ldots,V_{j-1})$ and the model for $\pi(X,V,\theta)$. The procedure is as follows:

1. Specify the regression model that would be used had all the intended data been collected.
2. Let $\phi_j(X,Z_j^-;\theta) = P(R_j = 1 \mid R_{j-1} = 1,X,Z_j^-;\theta)$ denote the probability that $Y_j$ is observed.
3. Specify and fit a model for $\phi_j$; denote the estimated parameters by $\hat{\theta}$.
4. Let $\pi_j\left(X,Z_j^-;\theta\right)=\prod_{k=1}^{j}\phi_j\left(X,Z_k^-;\theta\right)$ denote the probability that an individual has remained in the study to time $j$.
5. Fit the regression specified in Step 1, and weight individual contributions to the model by $\left\{\pi_j\left(X,Z_j^-;\hat{\theta}\right)\right\}^{-1}$. Use the assumed independence correlation structure.
6. Use the bootstrap technique for standard error estimation.

In large samples, the IPW GEE yields consistent estimators when the response probability model is correctly specified, but again may have high variance when individual weights are large. The augmented IPW GEE procedure (discussed below) can be used to partially remedy this weakness.

An example of this approach comes from Hogan et al. (2004): an analysis of repeated binary data from a smoking cessation study. The authors used inverse probability weighting to estimate the effect of a behavioral intervention involving supervised exercise on the rate of smoking cessation among 300 women. The primary outcome was smoking status, assessed weekly over 12 weeks. This example shows how to construct sequential weights for model fitting, and it illustrates how unweighted GEE estimators can vary under different choices for the assumed correlation when longitudinal data are incomplete. The authors include analyses using LOCF for comparison, and a critique is also included. The publication includes SAS code for fitting the model.

**Augmented IPW Estimation Under MAR**

The IPW GEE method does not make full use of the information in incomplete cases. The augmented IPW GEE procedure remedies this weakness. The procedure is best understood in the simple case in which only the values for the last time point are missing for some cases, and they are MAR. Suppose one wishes to estimate $\mu = E(Y_K)$, the mean outcome in a particular treatment arm. As before, it will be convenient to introduce the variable $R_K = 1 - M_K$, where now $R_K = 1$ if $Y_K$ is observed. Denote the observed history at time $K$ as $Z_K^-$, and let $\pi(Z_K^-;\theta) = P(R_K = 1 \mid Z_K^-; \theta)$, which can be modeled as described in the previous section.

The augmented IPW (AIPW) estimator of $\mu$ is

$$\hat{\mu} = (1/n)\sum_i \frac{R_{iK}Y_{iK}}{\pi_i\left(Z_{iK};\hat{\theta}\right)} + (1/n)\sum_i \left\{\frac{R_{iK}}{\pi_i\left(Z_{iK};\hat{\theta}\right)} - 1\right\} g\left(Z_{iK}^-, X_i\right), \qquad (9)$$

where $g\left(Z_{iK}^-\right)$ is some function of the observed-data history up to $K - 1$.

The first term is just the IPW estimator of $\mu$. The second (augmentation) term has mean zero, so that $\hat{\mu}$ is still a consistent estimator. However, the variance of $\hat{\mu}$ will depend on the choice of $g$, where the optimal choice is $E(Y_K \mid Z_K^-, X)$. The precise form of this expectation is unknown, but it can be estimated, for example, from a regression of $Y_K$ on $(Z_K^-, X)$ for those with observed $Y_K$. In that case, $g(Z_{iK}^-, X_i)$ is replaced by the regression prediction $\hat{Y}_{iK}$. If this outcome regression model is approximately correct, then $\hat{\mu}$ can be substantially more efficient than the standard IPW estimator. Since we can only estimate $g$, it is comforting that it appears to be the case that even if $g$ is only approximately known, there is likely to be a substantial gain. However, more research is needed to fully justify this conjecture.

In summary, augmented IPW estimators are obtained by adding to the IPW estimating functions an augmentation term that depends on unknown functions of the observed data. Appropriate choices of these functions can lead to substantial efficiency improvements over standard IPW GEE.

### Double Robustness Property of Augmented IPW Estimators

An important limitation of standard IPW estimation is that it yields biased estimates if the missingness model is incorrect. Remarkably, the augmented IPW estimator $\hat{\mu}$ with the optimal choice for $g$ offers not only efficiency improvements over IPW, but also bias protection against misspecification of the model for the missingness probabilities (for accessible accounts, see Tsiatis, 2006; Rotnitzky et al., 2009).

Methods that yield valid inferences when either one or the other of the outcome regression or the missingness model is correct are said to have a double-robustness property. Double robustness is a useful theoretical property, although it does not necessarily translate into good performance when neither model is correctly specified (Kang and Schafer, 2007). Empirical and theoretical studies of these estimators have begun to appear in the statistics and econometrics literature. With more published applications to real data and carefully designed simulation studies, the use of doubly robust estimators could become more commonly used in the near future. However, at present, the operating characteristics of this method in applied settings with finite samples need to become more completely understood.

### Advantages and Disadvantages of IPW Methods

The IPW method is generally simple to implement when the missing values have a monotone pattern, and can be carried out in any software package that allows weighted analyses. A key advantage is that, under a correctly specified model for missingness, information on many auxiliary variables can be accommodated, including information on previously observed outcomes. IPW methods can be extended to estimation targets other than the mean, such as the median. Potential disadvantages include (a) the need to correctly specify the nonresponse model, and (b) potential instabilities associated with weights that are very large, leading to biased estimation and high variance in finite samples. Double robust estimators have the potential to alleviate both of these limitations.

## Likelihood Methods

Likelihood approaches are dependent on a parametric model for the full-data distribution. Inference is based on the likelihood function of the

observed data. To describe these approaches, we define $Y$,$V$,$X$, and $M$ as above. The objective is to draw inference about a parameter $\theta$ in a model $p(y|x;\ \theta)$ for the response data that were intended to be collected but that might not be fully observed.

Under MAR, methods such as ML and Bayesian posterior inference can be used to draw inference about $\theta$ without having to specify an explicit model that relates $M$ to ($Y$,$V$,$X$). (To avoid distracting technicalities, we assume that $V$ is not present.)

To understand this approach, first write the model for the joint distribution [$Y$, $M$ | $X$] as

$$p(y,m \mid x;\theta,\psi) = p(y \mid x;\theta)p(m \mid y,x;\psi), \tag{10}$$

where $p(y \mid x;\theta)$ is the model for the *full* response data (i.e., the data that were intended to be collected), $p(m \mid y,x;\psi)$ is the model for the missing data mechanism, and ($\theta$,$\psi$) are unknown parameters.

Now let $D$ = ($Y_{obs}$,$X$,$M$) denote the observed data for an individual. The individual contribution to the likelihood function for ($\theta$,$\psi$) given $D$ is obtained by averaging (or integrating) over all possible realizations of the missing observations $Y_{mis}$,

$$L(\theta,\psi \mid D) = \int p(y_{obs},y_{mis} \mid x;\theta)p(m \mid y_{obs},y_{mis},x;\psi)dy_{mis}. \tag{11}$$

Under MAR, the second term under the integral simplifies as $p(m \mid y_{obs},y_{mis},x;\psi) = p(m \mid y_{obs},x;\psi)$. If one further assumes that $\theta$ and $\psi$ are functionally distinct (i.e., $\theta$ is not a function of $\psi$ and vice versa), then the likelihood factors as

$$L(\theta,\psi \mid D) = p(m \mid y_{obs},x;\psi)\int p(y_{obs},y_{mis} \mid x;\theta)dy_{mis} \tag{12}$$

$$= p(m \mid y_{obs},x;\psi)p(y_{obs} \mid x;\theta). \tag{13}$$

An immediate consequence is that inference about $\theta$ no longer depends on the functional form of the missing data mechanism because the likelihood for $\theta$ is directly proportional to $p(y_{obs}|x;\theta)$. When missingness is MAR and when $\theta$ and $\psi$ are functionally distinct, the missing data mechanism is said to be *ignorable* because it does not have to be modeled in order to draw inferences about $\theta$. In practice, MAR is the key condition.

### Maximum Likelihood Inference

Large-sample inferences about $\theta$ under the ignorability assumption follow from standard results in ML theory, which state that under some regularity conditions $\hat{\theta}$ follows a normal distribution having mean $\theta$ and variance equal to the inverse information matrix $I^{-1}(\theta)$. The variance can be estimated by $I^{-1}(\hat{\theta})$ or by such other techniques as bootstrap or sandwich estimators. (For details on the sandwich estimator of variance, see Kauermann and Carroll, 2001.) Numerical methods for maximizing the likelihood include factored likelihood methods for monotone patterns and iterative methods like the EM algorithm and its extensions for general patterns; see Little and Rubin (2002) for details.

### Bayesian Inference and Data Augmentation

ML is most useful when sample sizes are large because the parameter estimates are consistent, and standard errors are well approximated by the large sample variance-covariance matrix. However, Bayesian inference offers a useful alternative in some settings. First, when sample sizes are small, a useful alternative approach is to add a prior distribution for the parameters and compute the posterior distribution of the parameters of interest. Second, when computation of the observed-data likelihood is difficult, data augmentation embedded in a posterior sampling algorithm can make computation of posterior modes much simpler. Although there are notable exceptions, for many standard models it is possible to specify priors that are diffuse enough so that inferences from ML and from a posterior are essentially equivalent.[3]

For ignorable models, this posterior distribution is

$$p(\theta \mid D) \propto p(\theta) L(\theta \mid D), \tag{14}$$

where $p(\theta)$ is the prior and $L(\theta \mid D)$ is the same likelihood as in (13)—the full-data likelihood averaged over all possible realizations of the missing data.

Because the posterior distribution rarely has a simple analytic form for incomplete-data problems, simulation methods are often used to generate draws of $\theta$. Keeping in mind that the parameter $\theta$ indexes the distribution $p(y \mid x;\theta)$ of the *full* response data, generating from the posterior can be made easier by embedding a data augmentation step within the sampling algorithm; this approach effectively imputes the missing responses under

[3]Note, however, that assessing sensitivity to the prior is an important part of any Bayesian analysis.

the model $p(y \mid x;\theta)$ specified for [$Y \mid X$] (see Tanner, 1996; Daniels and Hogan, 2008).

For monotone missing data patterns, when the parameters of the factored distributions are distinct, the factored likelihood methods yield simple direct simulation methods; for general patterns, iterative simulation methods like data augmentation and the Gibbs' sampler play a prominent role (Little and Rubin, 2002). In particular, the Gibbs' sampler starts with an initial draw $\theta^{(0)}$ from an approximation to the posterior distribution of $\theta$. Given a value $\theta^{(t)}$ of $\theta$, drawn at iteration $t$,

1. draw $Y_{\text{mis}}^{(t+1)}$ from the distribution $p\left(y_{\text{mis}} \mid D;\theta^{(t)}\right)$, and
2. draw $\theta^{(t+1)}$ from the distribution $p\left(\theta \mid D, Y_{\text{mis}}^{(t+1)};\theta^{(t)}\right)$.

The iterative procedure can be shown in the limit to yield draws from the posterior distribution of $\theta$ given $D$, integrating over $Y_{\text{mis}}$. This algorithm can be run independently $L$ times to generate $L$ independent, identically distributed draws from the approximate joint posterior distribution of $(\theta, Y_{\text{mis}})$.

***Example: Multivariate Normal Regression*** For continuous outcomes and ignorable missing data, a common approach is to use a regression model based on a multivariate normal distribution. The general specification assumes that the vector $Y = (Y_1,\ldots,Y_K)^T$ follows a multivariate normal distribution, conditionally on design variables $X$. Hence, one assumes

$$Y \mid X = x \sim N\left(\mu(x), \Sigma(x)\right), \tag{15}$$

where $\mu(x)$ is a $K$-dimensional mean vector, and $\Sigma(x)$ is a $K \times K$ variance-covariance matrix. Especially if $K$ is large or if $X$ is high-dimensional, some choices must be made about the functional form and structure of both $\mu(x)$ and $\Sigma(x)$ to ensure that the model can be fit to finite samples. For example, it is possible to assume $\mu(x)$ has a linear trend over time, separately by treatment group, or that $\Sigma(x)$ has a simplified structure that can be described using a small number of parameters. As with the normality assumption itself, however, with incomplete data these choices will not lend themselves to empirical checking. The predicting distribution of the missing responses can differ substantially according to the specification.

In fact, with incomplete data, two models having the same mean specification but different specifications of $\Sigma$ will yield different inferences about the mean.

In short, for a multivariate normal model (including random effects models), inference about the mean from incomplete data depends on specification of the variance-covariance matrix (Meng and Rubin, 1993; Daniels and Hogan, 2008).[4] This dependence should be assessed, along with that for outlying or influential observations and most importantly for the missing data mechanism using sensitivity analyses, as described in Chapter 5.

***Example: Random Effects Models (Mixed Models)*** For longitudinal data, random effects models provide a parsimonious way to specify a multivariate distribution. The simplest versions are specified on two levels. The first (top) level specifies the distribution of responses for an individual, conditional on individual-specific random effects. For example, if it is assumed that each individual's repeated measures are governed by an underlying individual-specific mean, then variations in individual means can be represented through random effects $u$, and the top level model might be

$$\left[ Y_{ij} \mid X = x_i \right] = \left( \alpha_j + u_i \right) + \beta x_i + e_{ij}, \tag{16}$$

where $u_i$ is the individual-specific random effect, $e_{ij}$ is residual error, and both have mean zero. In most random effects models, $u$ and $e$ are assumed to be independent. In this simple model, the mean of $Y_{ij}$ at time $j$ is $\alpha_j + u_i$ for those with $x_i = 0$, and $\alpha_j + \beta + u_i$ for those having $x_i = 1$. The average treatment effect for an individual is $\beta$.

The second level specifies a distribution for the random effects; typically this is chosen to be the normal distribution, such as $u_i \sim N(0,\tau^2)$. If one combines this with an assumption of normality for $e_{ij}$, the distribution of the responses is itself normal, with variance partitioned into its within- and between-individual components.

Importantly, regression coefficients in the random effects model are *within-subject* effects—or the average effect for a typical individual—because they are being estimated conditionally on the random effect $u$. By contrast, the treatment effect estimated in a multivariate normal regression or other model not having random effects is a *between-subject* effect—the average effect between the groups of individuals randomized to treatment and control.

When both the error and random effects distributions are normal, the random effects model coincides with a multivariate normal regression with a constrained parameterization of $\Sigma$. In this special case, the between- and

[4]This may seem counterintuitive because for datasets having no missing values, estimates of the mean typically do not depend on specification of the variance matrix. This is not true for incomplete data.

within-subject treatment effects are equivalent (though their standard errors will differ). However, in general nonlinear mixed models, the treatment effect estimated from a random effects model does not coincide with the standard between-subjects contrast that is typically of direct interest for regulatory decision making (see Diggle et al., 2002).

***Likelihood-Based Models for Binary Data*** A substantial literature describing likelihood-based models for repeated binary data has developed over the last 10-15 years. Many of these models are based on a loglinear formulation and are parameterized directly in terms of between-subject treatment effects; see Fitzmaurice et al. (1993), Heagerty (1999, 2002).

### Advantages and Disadvantages of Likelihood-Based Methods

If missingness is ignorable, ML and Bayesian approaches under ignorability provide valid inferences using models that are generally easy to fit using commercial software. Random effects models can be very useful for simplifying a highly multivariate distribution using a small number of parameters.

However, with incomplete data, there are several reasons that inference about a treatment effect should not be limited to a single likelihood-based model:

- In addition to the ignorability assumption, inference relies on parametric assumptions about the model $p(y \mid x;\theta)$. These assumptions cannot be jointly checked from observed data, and it is difficult to ascertain the degree to which these different assumptions may be driving the inference.
- When using Bayesian posterior inference, results can sometimes be dependent on prior specifications, such as variance components (see Gelman, 2006).
- In the case of random effects models, parametric assumptions include those being made about the random effects. Frequently, these are made primarily for convenience of computation, as when the random effects are assumed to be normally distributed.
- With random effects models, two estimands can be distinguished: a between-subject (population-averaged) and within-subject (or subject-specific) treatment effect. For models where the mean is a linear combination of treatment and covariates, these quantities are equal, but the standard errors will differ. For models where the mean is nonlinear in treatment and covariates (such as random effects logistic regression), the between- and within-subject treatment effects will differ. In either case, the estimand must be stated and justified in advance of conducting the analysis. For more discussion on this point, see Diggle et al. (2002).

For more information on specification and inference for likelihood-based models for repeated measures, see Verbeke and Molenberghs (2000), Diggle et al. (2002), Fitzmaurice et al. (2004), and Daniels and Hogan (2008). More information on Bayesian inference can be found in Carlin and Louis (2000) and Gelman et al. (2003).

### Imputation-Based Approaches

Methods that impute or fill in the missing values have the advantage that, unlike complete-case analysis, the information from observed values in the incomplete cases is retained. Single imputation approaches include (a) regression imputation, which imputes the predictions from a regression of the missing variables on the observed variables; (b) hot deck imputation, which matches the case with missing values to a case with values observed that is similar with respect to observed variables and then imputes the observed values of the respondent; and (c) LOCF or BOCF methods for repeated measures, which impute the last observed value or the baseline value of the outcome variable. Two problems with single imputation are (1) inferences (tests and confidence intervals) based on the filled-in data can be distorted by bias if the assumptions underlying the imputation method are invalid, and (2) statistical precision is overstated because the imputed values are assumed to be true.

#### Example: LOCF Imputation

A common single imputation method is LOCF, which is based on the strong assumption that the outcome of a participant does not change after drop out. Even if this model is scientifically reasonable, attention needs to be paid to whether a single imputation by LOCF propagates imputation uncertainty in a way that yields valid tests and confidence intervals. For certain estimands, like the change from baseline to a fixed time after baseline, single LOCF imputation appropriately reflects uncertainty, since under the assumed model, the observation at time of dropout is essentially exchangeable with the observation at the end of the study, if observed. In general, however, this will not be the case. For instance, if imputed values are used to estimate a slope or area under the curve, statistical uncertainty from LOCF may be underestimated.

LOCF is sometimes mistakenly considered to be valid under MCAR or MAR, but in general it makes an MNAR assumption. Suppose for simplicity that there are $K$ repeated measures, that missing data are confined to the last time point $K$, and that there are no auxiliary variables. The MAR assumption is

$$p\left(y_K \mid x, y_{\text{obs}}, m=1\right) = p\left(y_K \mid x, y_{\text{obs}}, m=0\right); \tag{17}$$

that is, that the predictive distribution of $Y_K$, based on design variables $X$ and observed response data $Y_{\text{obs}} = (Y_1,\ldots,Y_{K-1})$, is the same for those with missing and observed $Y_K$. By contrast, LOCF assumes that, for those with missing $Y_K$, the predicted value is $Y_{K-1}$ with probability one. Formally, it assumes[5]

$$P\left(y_K = y_{K-1} \mid X = x, y_{\text{obs}}\right) = 1. \tag{18}$$

Hence, $p\left(y_K \mid x, y_{\text{obs}}, m=1\right) \neq p\left(y_K \mid x, y_{\text{obs}}, m=0\right)$ in general, and LOCF is neither MCAR nor MAR. Moreover, the LOCF method attaches no uncertainty to the filled-in value of $Y_K$ resulting in an artificial increase of sample size for some analyses. Similar comments apply to BOCF (Molenberghs and Kenward, 2007).

Although the utility of these methods does rest on the plausibility of the assumptions underpinning these estimators, the pragmatic justification often stems from the sometimes mistaken view that they provide a simple and conservative imputation that will help prevent approval of ineffective treatments. However, this is not necessarily the case, since, for example, LOCF is anticonservative in situations where participants off study treatment generally do worse over time. In such cases, if many participants discontinue study treatment due to problems with tolerability, the treatment can be made to look much better than the control by such an imputation strategy.

### Multiple Imputation

Multiple imputation is designed to fill in missing data under one or more models and to properly reflect uncertainty associated with the "filling-in" process (Rubin, 1987). Instead of imputing a single value for each missing observation, a set of $S$ (say $S = 10$) values for each missing observation is generated from its predictive distribution,[6] resulting in $S$ distinct filled-in

[5]Another characterization of this assumption is that the predicted mean of $Y_K$ is $E(Y_K \mid Y_1,\ldots,Y_{K-1}) = Y_{K-1}$, and that the variance of this prediction is zero: i.e., $\text{var}(Y_K \mid Y_1,\ldots,Y_{K-1}) = 0$.

[6]One transparent way to represent the missing data assumptions is through the predictive distribution of missing outcomes, given the observed outcomes and whatever modeling assumptions are being used for inference. The predictive distribution of a missing outcome for an individual is $p(y_{\text{mis}} \mid y_{\text{obs}}, x_i, m_i, v_i)$. To provide more concreteness to this notion, using multiple imputation, the predictive distribution of missing responses can be characterized by the imputations themselves. If a parametric model is being used under MAR, the predictive

datasets. One then follows an analysis that would have been used on the full data for each of the *S* datasets and combines the results in a simple way. The multiple imputation estimate is the average of the estimates from the *S* datasets, and the variance of the estimate is the average of the variances from the *S* datasets plus the between-sample variance of the estimates over the *S* datasets.

The between-imputation variance estimates the contribution to the overall variance from imputation uncertainty, which is missed by single imputation methods. Another benefit of multiple imputation is that the averaging over datasets reduces sampling variance and therefore results in more efficient point estimates than does single random imputation. Usually, multiple imputation is not much more difficult than single imputation—the additional computing from repeating an analysis *S* times is not a major burden. Moreover, the methods for combining inferences are straightforward and implemented in many commercially available software packages. Most of the work is in generating good predictive distributions for the missing values. This aspect has been addressed in a variety of widely-available software packages. We emphasize here the need to fully understand and communicate the assumptions underlying any imputation procedures used for drawing inferences about treatment effects.

An important advantage of multiple imputation in the clinical trial setting is that auxiliary variables that are not included in the final analysis model can be used in the imputation model. For example, consider a longitudinal study of HIV, for which the primary outcome *Y* is longitudinal CD4 count and that some CD4 counts are missing. Further, assume the presence of auxiliary information *V* in the form of longitudinal viral load. If *V* is not included in the model, the MAR condition requires the analysis to assume that, conditional on observed CD4 history, missing outcome data are unrelated to the CD4 count that would have been measured; this assumption may be unrealistic. However, if the investigator can confidently specify the relationship between CD4 count and viral load (e.g., based on knowledge of disease progression dynamics) and if viral load values are observed for all cases, then MAR implies that the predictive distribution of missing CD4 counts given the observed CD4 counts and viral load values is the same for cases with CD4 missing as for cases with CD4 observed, which may be a much more acceptable assumption.

---

distribution can be characterized by examining the predicted values of the missing observations under the model. For some methods, such as inverse probability weighting, this will not always be a straightforward exercise. However, when it is possible to compute in a straightforward manner, examining the predictive distribution can be an important diagnostic for understanding whether the missing data assumptions yield realistic values of the missing observations themselves.

The missing values of CD4 can be multiply imputed using a model that includes the auxiliary variables, and the multiple imputation inference applied to an analysis model that does not condition on the auxiliary variables. Note that we assume here complete data on the auxiliary variables, so cases that have viral load measured but not historical CD4 levels are not useful. However, if the auxiliary variables are incomplete but measured in a substantial number of cases for which CD4 is missing, then multiple imputation can still be applied productively, multiply imputing the missing values of both CD4 and the missing auxiliary variable values in the imputation step.

The predictive distribution of the missing values can be based on a parametric model for the joint distribution of $V$ and $Y$ given $X$, using the Bayesian paradigm described in Section 4.3.2. Extension to more robust spline-based models are considered in Zhang and Little (2009). Implementation assuming MAR is as follows:

1. Specify an analysis model $p(y \mid x;\theta)$ for $[Y \mid X]$, the data that was *intended* to be collected. This is the model for the full response data.
2. Specify an imputation model $p(y \mid x,v;\phi)$ for $[Y \mid X,V]$ that will be used to impute missing values of $Y$. Fit the model to $\{Y,V,X; R = 1\}$; that is, the data on observation times where $Y$ is observed.
3. For those with $R = 0$, generate $S$ predicted values $\hat{Y}$ from the predictive distribution of $p(y \mid x,v) = \int p(y \mid x,v;\phi)p(\phi)d\phi$. (Specific approaches to drawing from the predictive distribution, are give above.) This creates $S$ completed datasets.
4. Fit the model in step 1 to each of the $S$ completed datasets, generating parameter estimates $\hat{\theta}_1,\ldots,\hat{\theta}_S$ and associated standard errors.
5. Combine the estimates and standard errors into a summary inference about $\theta$, using rules cited above.

**Advantages and Disadvantages of Imputation-Based Procedures**

Single-imputation methods such as LOCF and BOCF are simple to implement, but they generally do not conform to well-recognized statistical principles for drawing inference, especially in terms of reflecting all sources of uncertainty in an inference about treatment effect. Two popular misconceptions about LOCF and BOCF are (1) that they reflect an MAR or MCAR mechanism, and (2) that they result in estimates of treatment effect that tend to favor placebo or standard therapy over the experimental treatment. Claim (2) is generally not true because the methods do not always

yield conservative effect estimators, and standard errors and confidence interval widths can be underestimated when uncertainty about the imputation process is neglected.

Multiple imputation methods address concerns about (b), and enable the use of large amounts of auxiliary information. They can be relatively straightforward to implement, without special programming needs, and can handle arbitrary patterns of missing data. Moreover, the missing data assumptions are made explicit in the imputation model.

Although multiple imputation is a principled method, it does rely on parametric assumptions. Moreover, the data model $p(y \mid x;\theta)$ may be incompatible with the imputation model $p(y \mid v,x;\phi)$. Compatible models have the property that when auxiliary variables $v$ are integrated out of the imputation model, the result is the data model: that is, $\int p(y \mid v,x;\phi)p(v \mid x;\phi)dv = p(y \mid x;\phi)$. To verify compatibility, a model for [$V$ | $X$] is needed (see Meng, 1994).

As an example, with multivariate data, it is often possible to formulate a set of conditional distributions relating each variable to a set of the other variables, which are reasonable when taken one at a time, but incoherent in the sense that they cannot be derived from a single joint distribution. Such models, even when incoherent, may still be useful for creating multiple imputations (see, e.g., Baccini et al., 2010). Finally, robustness of multiple imputation inference to model form has also been investigated (see Zhang and Little, 2009), but more research would be valuable.

For key references, Rubin (1987, 1996) provides comprehensive reviews of multiple imputation. Discussions of the issues related to clinical trials are found in Glynn et al. (1993), Heitjan (1997), Liu et al. (2000), and Molenberghs and Kenward (2007).

***MCAR-MAR Diagnostics*** As indicated above, while the observed data cannot be used to distinguish between MAR and MNAR missing data mechanisms, they can be used to distinguish between MCAR and MAR models and between competing MAR models. To this end, it is standard to assess the degree to which the treated and control groups differ in how the outcome variable $Y$ relates to the various design variables $X$ and auxiliary variables $V$. Therefore, various standard summary statistics and graphs should be used to assess the extent to which this is the case. Examples would include graphs for the treated and control groups of $Y$ against $X$ variables or $V$ variables and means for the treated and control groups of $Y$ for individuals with specific ranges of values of $X$ and $V$ in comparison with individuals with other ranges, and correlations of $Y$ for the treated and control groups versus $X$ or $V$.

### Event Time Analyses

In event time analysis, missingness can take at least two important forms. First, with repeated event times, it is possible that some event times in a sequence are unobserved. This situation is entirely similar to the one described above. Second, it is common for event times to be unobserved because the event has not occurred when follow-up terminates. This situation is referred to as right censoring in the time-to-event (or survival) literature and has received considerable attention. It is a special case of coarsened data (Heitjan, 1993), which also includes left-censored and interval-censored data, heaped and grouped data.

The term "noninformative censoring" does not have a consistent definition, but it often refers to a censoring mechanism that is independent of the unobserved time to event, given an appropriate set of covariates (see Rotnitzky et al. [2009] for a discussion of this topic). Noninformative censoring implies that, among those still at risk at $t$, the hazard of death (or event) is equivalent between those who are censored at $t$ and those who are not. Typically this assumption is made conditionally on an observed covariate history up to time $t$. A related and more general concept is "coarsened at random," which extends the concept of MAR to coarsened data (Heitjan, 1993). Connections between missing data assumptions used in repeated measures and event time analyses are discussed in Scharfstein and Robins (2002).

## ANALYTIC METHODS UNDER MNAR

MNAR models apply when missingness depends on the missing values after conditioning on observed information. For example, if a subject drops out of a longitudinal clinical trial when his blood pressure got too high and one did not observe that blood pressure, or if in an analgesic study measuring pain, the subject dropped out when the pain was high and one did not observe that pain value, missingness depends on the missing value.

Any analysis, whether based on likelihoods or moment assumptions, must be based on correct specification of the association between $Y$ and $M$ (given $X$ and possibly $V$). Under MAR, the methods provided above will provide valid inferences. However, the MAR assumption cannot be verified from observed data, and even with modeling assumptions, the information to simultaneously estimate the parameters of the missing-data mechanism and the parameters of the complete-data model is very limited. Hence, model-based estimates tend to be very sensitive to misspecification of the model. In many if not most cases, a sensitivity analysis is needed to see how much the answers change for various alternative assumptions about the missing-data mechanism.

### Definitions: Full Data, Full Response Data, and Observed Data

Analytic approaches for handling MNAR depend on making assumptions about the joint distribution $[Y,M \mid X]$. To describe analytic approaches for handling MNAR, it is necessary to carefully distinguish between full data, observed data, and their associated models. We describe both the full and observed data in terms of responses only, but these can be easily extended to include auxiliary variables $V$.

***Full Data*** Full data refers to the sample that was intended to be collected. Importantly, the full data includes the missing data indicators. With univariate $Y$, the full data are $(Y,X,M)$ for every individual.

***Observed Data*** Observed data refers to the response, covariate, and missing data indicators that were actually observed. In a simple case in which $Y$ is univariate and no covariates are missing, the observed data is $(Y, X, M = 1)$ for those with observed response, and it is $(X, M = 0)$ for those with missing response.

With multivariate $Y$, it is useful to partition $Y$ as $(Y_{obs}, Y_{mis})$. In particular, if $K$ observations were intended to be taken on each individual, and the missing data pattern is monotone, then the observed data comprise $Y_{obs} = (Y_1,\ldots,Y_j)$ for some $j \leq K$, and $M = (1,\ldots,1,0,\ldots,0)$ where the first $j$ elements of $M$ are 1's and the remaining elements are 0's. Notice that $j$ will vary across individuals.

***Full-Data Model*** The full-data model is the probability model that governs the joint distribution $[Y,M \mid X]$. Regardless of the form of this joint distribution, it can be written as $p(y,m \mid x) = p(y_{obs},y_{mis},m \mid x)$, and factored as $p(y_{obs},y_{mis},m \mid x) = p(y_{obs},m \mid x) \times p(y_{mis} \mid y_{obs},m,x)$.

This factorization makes clear which parts of the full-data model can be inferred from observed data and which cannot. Specifically, notice that the first factor on the right-hand side is a model for the distribution of variables that are *observed*. Generally speaking, the data analyst can estimate this distribution from the observed data, possibly by making modeling assumptions that can be evaluated using standard goodness-of-fit methods.

The second factor on the right-hand side is the model for the distribution of *missing* observations, and it cannot be inferred from observed data alone for the simple reason that no assumptions about the distribution can be checked from observed data. This factorization makes clear that inference from incomplete data requires the analyst to specify a model (or set of assumptions) for the observed-data distribution and to combine that with a set of untestable and unverifiable assumptions that describe how missing data are to be extrapolated from the observed data.

***Target Distribution (or Parameter)*** Most often in clinical trials, primary interest centers on the distribution $[Y \mid X] = [Y_{obs}, Y_{mis} \mid X]$, where $X$ includes the treatment group and possibly other design variables. The target distribution is related to the full-data distribution through the identity:

$$p(y_{obs}, y_{mis} \mid x) = \sum_m p(y_{obs}, y_{mis}, m \mid x) = \sum_m p(y_{obs}, m \mid x) p(y_{mis} \mid y_{obs}, m, x). \quad (19)$$

Hence, inference about the target distribution relies critically on the untestable assumptions being made about $p(y_{mis} \mid y_{obs}, m, x)$.

***Selection and Pattern Mixture Models*** Two broad classes of models for the joint distribution of $Y$ and $M$ are selection models, which factor the full data distribution as

$$[Y_{obs}, Y_{mis}, M \mid X] = [M \mid Y_{obs}, Y_{mis}, X] \times [Y_{obs}, Y_{mis} \mid X] \quad (20)$$

and pattern mixture models, which factor the full-data distribution as

$$[Y_{obs}, Y_{mis}, M \mid X] = [Y_{obs}, Y_{mis} \mid M, X] \times [M \mid X]. \quad (21)$$

Pattern mixture models can be factored to make the missing data extrapolation explicit within missing data pattern $M$, that is

$$[Y_{obs}, Y_{mis}, M \mid X] = [Y_{mis} \mid Y_{obs}, M, X] \times [Y_{obs} \mid M, X] \times [M \mid X]. \quad (22)$$

## Selection Models

Selection models can be divided into two types, (1) parametric and (2) semiparametric. Parametric selection models were first proposed by Rubin (1974) and Heckman (1976), based on parametric assumptions for the joint distribution of the full data (usually, a normal distribution for responses and a probit regression for the missing data indicators). For repeated measures, parametric selection models were described by Diggle and Kenward (1994), and semiparametric models were proposed by Robins et al. (1995) and Rotnitzky et al. (1998).

To illustrate a standard formulation, assume the full-response data comprise $(Y_1, Y_2)$, and the objective is to capture the mean of $Y_2$ in each treatment group. Further, assume $Y_2$ is missing on some individuals. A parametric selection model might assume that the full-response data follows a bivariate normal distribution:

$$(Y_1, Y_2) \mid X = x \sim N(\mu(x), \Sigma(x)), \quad (23)$$

and the "selection mechanism" part of the model follows a logistic regression

$$\text{logit}\left\{P\left(M=0 \mid Y_1, Y_2, X\right)\right\} = \alpha_0 + \alpha_1 Y_1 + \alpha_2 Y_2. \quad (24)$$

Parametric selection models can be fit to observed data, even though there appears to be no empirical information about several of the model parameters. Specifically, there is no information about the association between $M$ and $Y_2$ because $Y_2$ is missing. Likewise, there is no information about the mean, variance, and covariance parameters involving $Y_2$.

The model can be fit because of the parametric and structural assumptions being imposed on the full-data distribution. This can be seen as both beneficial or as a reason to exercise extreme caution. Convenience is the primary benefit, especially if the model can be justified on scientific grounds. The reason for caution is that, again, none of the assumptions underlying this parametric model can be checked from the observed data. In parametric selection models fit under the MNAR assumption, identification of parameters and sensitivity to assumptions raises serious problems: see, for example, Kenward (1998), Little and Rubin (2002, Chapter 15), the discussion of Diggle and Kenward (1994), and Daniels and Hogan (2008, Chapter 9).

*Semiparametric* selection models do not assume a parametric model for the full-data response distribution, so they are therefore somewhat less sensitive to these assumptions. These models are discussed in greater detail in Chapter 5.

### Pattern Mixture Models

Pattern mixture models were proposed for repeated measures data by Little (1993, 1994); a number of extensions and generalizations have followed. The connection between pattern mixture and selection models is described in Little and Wang (1996), in Molenberghs et al. (1998), and in Birmingham et al. (2003).

The models can be viewed from an imputation perspective, in which missing values $Y_{\text{mis}}$ are imputed from their predictive distribution given the observed data including $M$; that is,

$$p(y_{\text{mis}} \mid y_{\text{obs}}, x, M). \quad (25)$$

Under MAR, this equals $p(y_{\text{mis}} \mid y_{\text{obs}}, x)$. However, if data are not MAR, the predictive distribution (25) is a direct by-product of the pattern mixture formulation because it conditions on the missing data indicators. This more direct relationship between the pattern mixture formulation and the

predictive distribution for imputations yields gains in transparency and computational simplicity in some situations, as illustrated in Kenward and Carpenter (2008, Section 4.6).

Under MNAR, the selection model factorization requires full specification of the model for the missing data mechanism. Some pattern mixture models avoid specification of the model for the missing data mechanism in MNAR situations by using assumptions about the mechanism to yield restrictions on the model parameters (Little, 1994; Little and Wang, 1996; Hogan and Laird, 1997).

Many pattern mixture formulations are well suited to sensitivity analysis because they explicitly separate the observed data distribution from the predictive distribution of missing data given observed data. Sensitivity analyses can be formulated in terms of differences in mean (or other parameter) between those with observed and those with missing responses.

**Advantages and Disadvantages of Selection and Pattern Mixture Models**

Substantively, it seems more natural to assume a model for the full-data response, as is done in selection models. For example, if the outcome is blood pressure, it may seem natural to assume the combined distribution of blood pressures over observed and missing cases follows a single distribution, such as the normal distribution. Moreover, if MAR is plausible, a likelihood-based selection formulation leads directly to inference based solely on the model for the full-data response, and inference can proceed by ML.

However, it may not be intuitive to specify the relationship between nonresponse probability and the outcome of interest, which typically has to be done in the logit or probit scale. Moreover, the predictive distribution of missing responses typically is intractable, so it can be difficult to understand in simple terms how the missing observations are being imputed under a given model. And, as indicated above, selection models are highly sensitive to parametric assumptions about the full data distribution. This concern can be alleviated to some degree by the use of semiparametric selection models.

Specification of pattern mixture models also appeals to intuition in the sense that it is natural to think of respondents and nonrespondents having different outcome distributions. The models are transparent with respect to how missing observations are being imputed because the within-pattern models specify the predictive distribution directly.

Pattern mixture models can present computational difficulties for estimating treatment effects because of the need to average over missing data patterns; this is particularly true of pattern mixture specifications involving regression models within each pattern.

**Examples: Pattern Mixture Model for Continuous Outcomes**

Daniels and Hogan (2008, Chapter 10) use pattern mixture models to analyze data from a randomized trial of recombinant human growth hormone (rHGH) on muscle strength in elderly people. More than 120 people were randomized to four different treatment arms. The primary outcome in this trial was quadriceps strength, assessed at baseline, 6 months, and 12 months. A pattern mixture model was fit under MAR and parameterized to represent departures from MAR. The example shows how to construct sensitivity plots to assess the effect of departures from MAR on the inferences about treatment effect. An important feature of the model is that the fit to the observed data is unchanged at different values of the sensitivity parameters. However, the model does rely on parametric assumptions, such as normality. These assumptions can be checked for the observed data, but have to be subjectively justified for the missing data.

**Example: Pattern Mixture Model for Binary Outcomes**

Daniels and Hogan (2008, Chapter 10) use pattern mixture models to analyze data from an intervention study for smoking cessation among substance abusers. The primary outcome was smoking status, assessed at baseline, 1 month, 6 months, and 1 year. A pattern mixture model was fit under MAR and expanded to allow for MNAR missingness. In addition to presenting sensitivity analysis, the example shows how to incorporate prior information about the smoking rate of dropouts to obtain a summary inference about treatment effect.

**Sensitivity of Parametric Selection Models**

The sensitivity of MNAR selection models to distributional assumptions is illustrated by Verbeke and Molenberghs (2000, Chapter 17), who show that, in the context of an onychomycosis study, excluding a small amount of measurement error drastically changes the likelihood ratio test statistics for the MAR null hypothesis. In a separate example, Kenward (1998) revisited the analysis of data from a study on milk yield performed by Diggle and Kenward (1994). In this study, the milk yields of 107 cows were to be recorded during 2 consecutive years. Data were complete in the first year, but 27 measurements were missing in year 2 because these cows developed mastitis, which seriously affected their milk yield and therefore deemed missing for the purposes of the study. Although in the initial paper there was some evidence for MNAR, Kenward (1998) showed that removing two anomalous profiles from the 107 completely eliminated this evidence. Kenward also showed that changing the conditional distribution

of the year 2 yield, given the year 1 yield, from a normal to a heavy-tailed *t* distribution led to a similar conclusion.

Several authors have advocated using local influence tools for purposes of sensitivity analysis (Thijs et al., 2000; Molenberghs et al., 2001; Van Steen et al., 2001; Verbeke et al., 2001; Jansen et al., 2006). In particular, Molenberghs et al. (2001) revisited the mastitis example. They were able to identify the same two cows also found by Kenward (1998), in addition to another one. However, it is noteworthy that all three are cows with *complete* information, even though local influence methods were originally intended to identify subjects with other than MAR mechanisms of missingness. Thus, an important question concerns the combined nature of the data and model that leads to apparent evidence for an MNAR process. Jansen et al. (2006) showed that a number of features or aspects, but not necessarily the (outlying) nature of the missingness mechanism in one or a few subjects, may be responsible for an apparent MNAR mechanism.

### Selection and Pattern Mixture Models: Literature

The literature covering selection and pattern mixture models is extensive. Review papers that describe, compare, and critique these models include Little (1995), Hogan and Laird (1997, 2004), Kenward and Molenberghs (1999), Fitzmaurice (2003), and Ibrahim and Molenberghs (2009). The models are also discussed in some detail in Little and Rubin (2002), Diggle et al. (2002), Fitzmaurice et al. (2004), Molenberghs and Kenward (2007), and Daniels and Hogan (2008).

An extensive literature also exists on extensions of these models involving random effects, sometimes called shared-parameter or random-coefficient-dependent models. Reviews are given by Little (1995) and Molenberghs and Kenward (2007). Although these models can be enormously useful for complex data structures, they need to be used with extreme caution in a regulatory setting because of the many layers of assumptions needed to fit the models to data.

## Recommendations

**Recommendation 9: Statistical methods for handling missing data should be specified by clinical trial sponsors in study protocols, and their associated assumptions stated in a way that can be understood by clinicians.**

Since one cannot assess whether the assumptions concerning missing data are or are not valid after the data are collected, one cannot assert that the choice of missing data model made prior to data collection needs to be

modified as a result of a lack of fit. Thus, one needs to carry out a sensitivity analysis. Of course, model fitting diagnostics can be used to demonstrate that the complete data model may need to be adjusted, but the missing data model raises no additional complexities.

**Recommendation 10: Single imputation methods like last observation carried forward and baseline observation carried forward should not be used as the primary approach to the treatment of missing data unless the assumptions that underlie them are scientifically justified.**

Single imputation methods do not account for uncertainty associated with filling in the missing responses. Further, LOCF and BOCF do not reflect MAR data mechanisms.

Single imputation methods are sometimes used not as a method for imputation but rather as a convenient method of sensitivity analysis when they provide a clearly conservative treatment of the missing data. This can obviously be accomplished by using a best possible outcome for the missing values in the control group and a worst possible outcome for the missing values in the treatment group. If the result of such a technique is to demonstrate that the results of the primary analysis do not depend on the treatment of the missing data, such an approach can be useful. However, techniques that are often viewed as being conservative and therefore useful in such an approach, are sometimes not conservative and so care is required.

**Recommendation 11: Parametric models in general, and random effects models in particular, should be used with caution, with all their assumptions clearly spelled out and justified. Models relying on parametric assumptions should be accompanied by goodness-of-fit procedures.**

We acknowledge that this is an area where the current toolkit is somewhat lacking, and therefore more research is needed. Some contributions to this area include Verbeke et al. (2001, 2008), Gelman et al. (2005), and He and Raghunathan (2009).

**Recommendation 12: It is important that the primary analysis of the data from a clinical trial should account for the uncertainty attributable to missing data, so that under the stated missing data assumptions the associated significance tests have valid type I error rates and the confidence intervals have the nominal coverage properties. For inverse probability weighting and maximum likelihood methods, this can be accomplished by appropriate computation of standard errors, using either asymptotic results or the bootstrap. For imputation, it**

**is necessary to use appropriate rules for multiply imputing missing responses and combining results across imputed datasets because single imputation does not account for all sources of variability.**

**Recommendation 13: Weighted generalized estimating equations methods should be more widely used in settings when missing at random can be well justified and a stable weight model can be determined, as a possibly useful alternative to parametric modeling.**

**Recommendation 14: When substantial missing data are anticipated, auxiliary information should be collected that is believed to be associated with reasons for missing values and with the outcomes of interest. This could improve the primary analysis through use of a more appropriate missing at random model or help to carry out sensitivity analyses to assess the impact of missing data on estimates of treatment differences. In addition, investigators should seriously consider following up all or a random sample of trial dropouts, who have not withdrawn consent, to ask them to indicate why they dropped out of the study, and, if they are willing, to collect outcome measurements from them.**

## INSTRUMENTAL VARIABLE METHODS FOR ESTIMATING TREATMENT EFFECTS AMONG COMPLIERS

Estimates of treatment effects for all individuals randomized as in intention-to-treat analysis are protected against bias by the randomization. In this estimand, individuals who are assigned a treatment but never comply with it, perhaps because they cannot tolerate treatment side effects, are treated in the same way as individuals who comply with the treatment. Sometimes, particularly in secondary analyses, interest lies in the treatment effect in the subpopulation of individuals who would comply with a treatment if assigned to it. The average treatment effect in this population is called the complier-average causal effect (CACE) (Baer and Lindeman, 1994; Angrist et al., 1996; Imbens and Rubin, 1997a, 1997b; Little and Yau, 1998; White, 2005).

An alternative estimand to the CACE is the average treatment effect (ATE) (Robins, 1989; Robins and Greenland, 1996). It is defined as the difference in mean outcome if all individuals had been assigned and complied with the treatment ($T = 1$) and the mean if all individuals had been assigned and complied with the control treatment ($T = 0$). The ATE is defined for the whole target population, and it requires assumptions about the treatment outcome for noncompliers had they complied with the treatment. Whether this counterfactual event is meaningful typically depends on context. For example, noncompliance to a behavioral treatment, such as an exercise

regime, might plausibly be changed by increased motivation, as might occur if evidence of success of the treatment becomes widely known. In contrast, if noncompliance to a drug is the result of intolerable side effects, then compliance may require a reformulation of the drug to remove the side effects. Such reformulation may change the properties of the drug, and estimation of the ATE is consequently more speculative.

Simple approaches to estimating the CACE or the ATE include as-treated analysis, in which participants are classified according to the treatment actually received, and per-protocol analysis, which restricts analysis to participants who comply with the assigned treatment. These analyses are subject to selection bias in that participants who comply with a treatment may be a biased sample of participants randomized to that treatment. The bias may be reduced by adjustment for covariates, but it remains a major concern.

Although this is often characterized as a problem of selection bias, recent approaches have suggested alternatives to as-treated and per-protocol analyses by applying a missing-data perspective. Consider a binary variable C(T) taking the value 1 if an individual would comply with a particular treatment T if assigned to it, and 0 otherwise. We call this variable *principal compliance*, to distinguish it from observed compliance, which depends on the treatment actually assigned. It is a special case of *principal stratification* (Frangakis and Rubin, 2002). Principal compliance C(T) is observed for participants who are assigned to treatment T, but it is not observed for participants assigned other treatments, T′, so for these individuals the values of C(T) can be regarded as missing.

In simple trials involving an active treatment and a control treatment, an alternative to as-treated and per-protocol estimates is based on the idea of treating the randomization as an instrumental variable (IV), in economic parlance. The IV estimator yields a direct estimate of the CACE, and it is protected from selection bias by the randomization. However, it requires certain assumptions to be valid, and it also yields estimators with potentially high variance, particularly if the treatment compliance rate is low. Model-based versions of the IV estimator based on treating C as missing for some participants have been proposed that are potentially more efficient, although they make stronger distributional assumptions. For a nontechnical article comparing this approach with as-treated and per-protocol estimates, see Little et al. (2009) for a discussion of extensions to two or more active treatments, see Long, Little, and Lin (in press).

An example illustrating the above discussion and a number of associated issues is provided by the evaluation of a trial to assess the effect of an influenza vaccine (Hirano et al., 2000). The trial randomly assigned physicians to encouragement (T1) or no encouragement (T2) to vaccinate their patients against influenza. The primary endpoint was hospitalization, and

the intention-to-treat estimates showed of those encouraged, 7.8 percent were hospitalized and of those not encouraged 9.2 percent were hospitalized. However, the trial had only a weak effect on the actual taking of the vaccine: of those encouraged, 31 percent of patients received the vaccine; of those not encouraged, 19 percent of patients received the vaccine.

Therefore, to better understand the trial results, at least a secondary estimand of interest was CACE, that is, in this case, the effect of encouragement on hospitalization for the patients who would have been vaccinated if their physician had been encouraged but not vaccinated if their physician had not been encouraged. Assuming the standard exclusion restrictions of IV, CACE was estimated as an 8.2 percent reduction in hospitalization. Yet, even this turned out to represent only part of a better understanding of the trial results.

In this study, there were a number of good baseline predictors of compliance under both arms, C(T1) and C(T2), and thus, the effect of compliers could be in part identified without the need of exclusion restrictions. When these restrictions were relaxed, the effect of encouragement on compliers was estimated at 3.7 percent, but there was at least as large of an estimated effect (5.3 percent) of encouragement on hospitalization for always-takers. Later commentaries on these results suggested that the latter effect is explainable by the earlier time in the season at which the always-takers likely receive the vaccine when encouraged, compared to when not encouraged. Since this effect is comparable to CACE, it suggested that the effect of vaccination lies more in its timing and not only on its receipt.

To further explicate this method, we offer an example of coprimary outcomes that induce missing data. For randomized controlled trials with two (or more) coprimary outcomes, say $E$ and $Y$, values of $E$ can determine whether $Y$ has a meaning as a measurement. This effect presents a challenge in the very definition of the effect between the two interventions, say T1 and T2, on $Y$, because the existence of $Y$ is determined after the intervention. This problem can be treated in principle in the context of missing information, not of $Y$ (which is sometimes undefined) but of certain strata, called principal strata. Our example involves clinical trials for HIV.

The idea of cell-mediated immunity is to train the killer cells to recognize and attack a protein that human CD4 cells create when the CD4 cells are infected (as opposed to targeting the virus directly, whose identification is difficult due to mutations over time). For this reason, randomized trials for cell-mediated immunity vaccines should be designed to assess two coprimary outcomes: reducing primary infection (say, $E$), and, if a person is infected ($E = 1$), keeping low viral load (say $Y$). Work by Gilbert et al. (2003) and then by Mehrotra et al. (2006) showed how principal stratification (Frangakis and Rubin, 2002) can be used to formulate the target hypotheses with such coprimary outcomes. Specifically, the first coprimary

research hypothesis is that changing treatment T1 (placebo) to T2 (vaccine) changes the primary infection rate *E*. The second coprimary research hypothesis should capture that the vaccine can also affect viral load when infected. However, the viral load distributions between infectees under the placebo condition and those infected under the vaccine condition could be different simply because the immune system is inherently different between the two groups. (In fact, if the vaccine prevents some primary infections, infectees under vaccine are expected to have weaker baseline immune system than infectees under placebo.) One can disentangle baseline differences from vaccine effects if one focuses on the people who would have been infected regardless of receiving the vaccine or the placebo. This stratum is known as a principal stratum because membership to it does not change depending on assignment to different interventions. Thus, the second coprimary research hypothesis can be that changing treatment T1 (placebo) to T2 (vaccine) will change the viral load for those for whom changing T1 to T2 does not prevent primary infection.

For a person under placebo who gets infected ($E$(T1) = infected), one does not know if the person would have been also infected under vaccine ($E$(T2) = infected), so membership to the principal stratum—$E$(T1) = $E$(T2) = infected—is partly missing. (Estimation of the effect of vaccine on viral load $Y$ for this stratum is discussed above.)

Additional examples of randomized controlled trials with coprimary outcomes using principal stratification include determining if the immune response to a vaccine is causing reduction in infection rates (Follmann, 2006); assessing more general surrogate outcomes in vaccine trials (Qin et al., 2008); and evaluating the effect of an intervention on severity of a disease (e.g., of prostate cancer) when a person does get the disease (Shepherd et al., 2008).

## MISSING DATA IN AUXILIARY VARIABLES

The assumptions and models discussed above have been limited to outcome variables. Usually, there are many auxiliary variables collected at each visit that can be useful to incorporate into the analysis. Specifically, these variables are useful because they both help explain the reasons for future nonresponse as well as help predict the missing outcomes (and so help improve the efficiency with which the treatment effects are estimated). They can also serve to make the MAR assumption more tenable. We have assumed throughout that the collection of auxiliary variable data is complete, which is clearly not always the case. We do note that the above approaches can be modified to incorporate missing auxiliary data by augmenting the missing outcome variable with a missing $V$. Although including $V$ along with the missing outcome variable will often address the problem,

the literature on missing data in longitudinal settings is fairly limited, and more research on dealing with missing auxiliary data would be useful. We do believe that many of the above approaches can be easily modified to incorporate auxiliaries by replacing $Y_k^-$ in the conditional means and probabilities with $Z_k^-$, which includes $(Y_1,\ldots,Y_{k-1},V_1,\ldots,V_{k-1})$. An excellent example of the use of this method is Liu et al. (2009).

# 5

# Principles and Methods of Sensitivity Analyses

This chapter concerns principles and methods for sensitivity analyses that quantify the robustness of inferences to departures from underlying assumptions. Unlike the well-developed literature on drawing inferences from incomplete data, the literature on the assessment of sensitivity to various assumptions is relatively new. Because it is an active area of research, it is more difficult to identify a clear consensus about how sensitivity analyses should be conducted. However, in this chapter we articulate a consensus set of principles and describe methods that respect those principles.

We begin by describing in some detail the difficulties posed by reliance on untestable assumptions. We then demonstrate how sensitivity to these assumptions can be represented and investigated in the context of two popular models, selection and pattern mixture models. We also provide case study illustrations to suggest a format for conducting sensitivity analyses, recognizing that these case studies cannot cover the broad range of types and designs of clinical trials. Because the literature on sensitivity analysis is evolving, the primary objective of this chapter is to assert the importance of conducting some form of sensitivity analysis and to illustrate principles in some simple cases. We close the chapter with recommendations for further research on specific aspects of sensitivity analysis methodology.

## BACKGROUND

There are fundamental issues involved with selecting a model and assessing its fit to incomplete data that do not apply to inference from complete data. Such issues occur even in the missing at random (MAR)

case, but they are compounded under missing not at random (MNAR). We believe that, especially when the primary analysis assumes MAR, the fit of an MAR model can often be addressed by standard model-checking diagnostics, leaving the sensitivity analysis to MNAR models that deviate from MAR. This approach is suggested in order not to overburden the primary analysis. The discussion in Chapter 4 provides some references for model-checking of MAR models. In addition, with MAR missingness mechanisms that deviate markedly from missing completely at random (MCAR), as in the hypertension example in Chapter 4, analyses with incomplete data are potentially less robust to violations of parametric assumptions than analyses with complete data, so checking them is even more critical.

The data can never rule out an MNAR mechanism, and when the data are potentially MNAR, issues of sensitivity to modeling asumptions are even more serious than under MAR. One approach could be to estimate from the available data the parameters of a model representing an MNAR mechanism. However, the data typically do not contain information on the parameters of the particular model chosen (Jansen et al., 2006).

In fact, different MNAR models may fit the observed data equally well but have quite different implications for the unobserved measurements and hence for the conclusions to be drawn from the respective analyses. Without additional information, one cannot usefully distinguish between such MNAR models based solely on their fit to the observed data, and so goodness-of-fit tools alone do not provide a relevant means of choosing between such models.

These considerations point to the necessity of sensitivity analysis. In a broad sense, one can define a sensitivity analysis as one in which several statistical models are considered simultaneously or in which a statistical model is further scrutinized using specialized tools, such as diagnostic measures. This rather loose and very general definition encompasses a wide variety of useful approaches.

A simple procedure is to fit a selected number of (MNAR) models, all of which are deemed plausible and have equivalent or nearly equivalent fit to the observed data; alternatively, a preferred (primary) analysis can be supplemented with a number of modifications. The degree to which conclusions (inferences) are stable across such analyses provides an indication of the confidence that can be placed in them.

Modifications to a basic model can be constructed in different ways. One obvious strategy is to consider various dependencies of the missing data process on the outcomes or the covariates. One can choose to supplement an analysis within the selection modeling framework, say, with one or several in the pattern mixture modeling framework, which explicitly models the missing responses at any given time given the previously observed responses. Alternatively, the distributional assumptions of the models can be altered.

The vast range of models and methods for handling missing data highlights the need for sensitivity analysis. Indeed, research on methodology has shifted from formulation of ever more complex models to methods for assessing sensitivity of specific models and their underlying assumptions. The paradigm shift to sensitivity analysis is, therefore, welcome. Prior to focused research on sensitivity, many methods used in practice were potentially useful but ad hoc (e.g., comparing several incompatible MNAR models to each other). Although informal sensitivity analyses are an indispensable step in the analysis of incomplete longitudinal data, it is desirable to have more formal frameworks within which to develop such analyses.

It is possible to assess model sensitivities of several different types, including sensitivity to: (a) distributional assumptions for the full data, (b) outlying or influential observations, and (c) assumptions about the missing data mechanism. Assessment of (a) can be partially carried out to the extent that one can compare observed and fitted values for the *observables* under the model specified for the full data. However, distributional assumptions for the missing data cannot be checked. Assessment of (b) can be used to identify observations that are outliers in the observed-data distribution or that may be driving weakly identified parts of an MNAR model (Molenberghs and Kenward, 2007). This chapter focuses on (c), sensitivity to assumptions about the missing data mechanism.

## FRAMEWORK

To focus ideas, we restrict consideration to follow-up randomized study designs with repeated measures. We consider the case in which interest is focused on treatment comparisons of visit-specific means of the repeated measures. With incomplete data, inference about the treatment arm means requires two types of assumptions: (i) untestable assumptions about the distribution of missing outcomes data, and (ii) testable assumptions about the distribution of observed outcomes. Recall that the full-data distribution, described in Chapter 4, can be factored as

$$[Y_{obs}, Y_{mis}, M \mid X] = [Y_{obs}, M \mid X] \times [Y_{mis} \mid Y_{obs}, M, X]. \quad (1)$$

Type (i) assumptions are needed to estimate the distribution $[Y_{mis} \mid Y_{obs}, M, X]$, while type (ii) assumptions are used, if necessary, to model the observables $[Y_{obs}, M \mid X]$ in a parsimonious way.

Type (i) assumptions are necessary to identify the treatment-specific means. Informally, a parameter is *identified* if one can write its estimator as a function that depends only on the observed data. When a parameter is not identified, it would not be possible to obtain a point estimate even if the sample size were infinite. It is therefore essential to conduct a sensitivity

analysis, whereby the data analysis is repeated under different type (i) assumptions, in order to clarify the extent to which the conclusions of the trial are dependent on unverifiable assumptions. The usefulness of a sensitivity analysis ultimately depends on the transparency and plausibility of the unverifiable assumptions. It is key that any sensitivity analysis methodology allow the formulation of these assumptions in a transparent and easy-to-communicate manner.

Ultimately, type (i) assumptions describe how missing outcomes are being "imputed" under a given model. A reasonable way to formulate these assumptions is in terms of the connection (or link) between the distributions of those having missing and those having observed outcomes but similar covariate profiles. Making this difference explicit is a feature of pattern mixture models. Examples discussed in this chapter illustrate both pattern mixture and selection modeling approaches.

In general, it is also necessary to impose type (ii) assumptions. An important consideration is that modeling assumptions of type (ii), which apply to the distribution of *observed data*, can be supported and scrutinized with standard model-checking techniques.

Broadly speaking, there are two approaches for combining type (i) and (ii) assumptions to draw inferences about the treatment-specific means: pattern mixture and selection modeling. To illustrate these approaches, the next four sections present four example designs of increasing complexity. The first two examples involve a single outcome, without and then with auxiliary data. These examples are meant to illustrate when and why the assumptions of type (i) and (ii) are needed. The third and fourth examples extend the designs to those with repeated measures, with monotone and non-monotone missing data, respectively, with and without auxiliary data.

Our examples are not meant to be prescriptive as to how every sensitivity analysis should be conducted, but rather to illustrate principles that can guide practice. Type (i) assumptions can only be justified on substantive grounds. As the clinical contexts vary between studies, so too will the specific form of the sensitivity analysis.

## EXAMPLE: SINGLE OUTCOME, NO AUXILIARY DATA

We start with the simple case in which the trial records no baseline covariate data, and the only measurement to be obtained in the study is that of the outcome $Y$, taken at a specified time after randomization. We assume that the treatment-arm-specific means of $Y$ form the basis for treatment comparisons and that in each arm there are some study participants on whom $Y$ is missing. We let $R = 1$ if $Y$ is observed and $R = 0$ otherwise.

Because estimation of each treatment arm mean relies solely on data from subjects assigned to that arm, the problem reduces to estimation of a mean $E(Y)$ based on a random sample with $Y$ missing in some units. Thus, formally, the problem is to estimate $\mu = E(Y)$ from the observed data, which comprises the list of indicators $R$, and the value of $Y$ for those having $R = 1$.

The MAR assumption described in Chapter 4 is a type (i) assumption. In this setting, MAR means that, within each treatment arm, the distribution of $Y$ among respondents (i.e., those with $R = 1$) is the same as that for nonrespondents (i.e., with $R = 0$).

This example illustrates several key ideas. First, it vividly illustrates the meaning of an untestable assumption. Let $\mu_1 = E(Y \mid R = 1)$ denote the mean among respondents, $\mu_0 = E(Y \mid R = 0)$ the mean among nonrespondents, and $\pi = P(R=1)$ the proportion of those responding. The full-data mean $\mu$ is a weighted average

$$\mu = \pi\mu_1 + (1 - \pi)\,\mu_0, \tag{2}$$

but there is no information in the data about the value of $\mu_0$. Hence, any assumption one makes about the distribution for the nonrespondents will be untestable from the data available. In particular, the MAR assumption—that $\mu_1 = \mu_0$—is untestable.

Second, this example also illustrates the *identifiability* (or lack thereof) of a parameter. Without making assumptions about $\mu_0$, the full-data mean $\mu$ cannot be identified (estimated) from the observed data. However, if one is prepared to adopt an untestable assumption, $\mu$ will be identified. For example, one can assume MAR is equivalent to setting $\mu_1 = \mu_0$. From (2), MAR implies that $\mu = \mu_1$, or that the full-data mean is equal to the mean among those with observed $Y$. Hence, under MAR, a valid estimate of $\mu_1$ is also valid for $\mu$. A natural choice is the sample mean among those with observed data, namely, $\hat{\mu}_1 = \sum_i R_i Y_i \Big/ \sum_i R_i$.

Third, this example is the simplest version of a pattern mixture model: the full-data distribution is written as a mixture—or weighted average—of the observed and missing data distributions. Under MAR, their means are equal. However, it is more typical to use pattern mixture models when the means are not assumed to be equal (MNAR).

By contrast, in the selection model approach, type (ii) assumptions are made in terms of how the probability of nonresponse relates to the possibly unobserved outcome. The full-data mean can be estimated using a weighted average of the observed outcomes, where the weights are individual-specific and correspond to the conditional probability of being observed given the observed outcome value. The reweighting serves to create a "pseudo-

population" of individuals who are representative of the intended full-data sample of outcomes.

Importantly, there is a one-to-one relationship between the specification of a selection model and specification of a pattern-mixture model. The key distinction ultimately arises in how type (ii) assumptions are imposed. As it turns out, the two approaches generate equivalent estimators in this simple example, but for more complex models that rely on type (i) assumptions to model the observed data, that is not the case.

### Pattern Mixture Model Approach

Because we are only interested in the mean of $Y$, it suffices to make assumptions about how the mean of $Y$ among nonresponders links to the mean of $Y$ among respondents. A simple way to accomplish this is by introducing a sensitivity parameter $\Delta$ that satisfies $\mu_0 = \mu_1 + \Delta$, or

$$E(Y \mid R = 0) = E(Y \mid R = 1) + \Delta. \tag{3}$$

It is easy to see that $\Delta = \mu_0 - \mu_1$, the difference in means between respondents and nonrespondents. To accommodate general measurement scales, the model should be parameterized so that the sensitivity parameter satisfies an identity such as

$$\mu_0 = g^{-1}\{g(\mu_1) + \Delta\}, \tag{4}$$

where $g(\,)$ is a function, specified by the data analyst, that is strictly increasing and maps values from the range of $Y$ to the real line. The function $g$ determines the investigator's choice of scale for comparisons between the respondents' and nonrespondents' means and is often guided by the nature of the outcome.

For a continuous outcome, one might choose $g(u) = u$, which reduces to the simple contrast in means given by (3), where $\Delta$ represents the difference in mean between nonrespondents and respondents.

For binary outcomes, a convenient choice is $g(u) = \log(u/(1-u))$, which ensures that the $\mu_0$ lies between 0 and 1. Here, $\Delta$ is the log odds ratio comparing the odds of $Y = 1$ between respondents and nonrespondents.

Each value of $\Delta$ corresponds to a different unverifiable assumption about the mean of $Y$ in the nonrespondents. Any specific value of $\Delta$ corresponds to an estimate of $\mu$ because $\mu$ can be written as the weighted average

$$\mu = \pi\mu_1 + (1-\pi)g^{-1}\{g(\mu_1) + \Delta\}. \tag{5}$$

After fixing $\Delta$, one can estimate $\mu$ by replacing $\mu_1$ and $\pi$ with their sample estimators $\hat{\mu}_1$ and $\hat{\pi}$. Formulas for standard error estimators can be derived from standard Taylor expansions (delta method), or one can use the bootstrap.

To examine how inferences concerning $\mu$ depend on unverifiable assumptions about the missing data distribution, notice that $\mu$ is actually a function of $\Delta$ in (5). Hence, one can proceed by generating an estimate of $\mu$ for each value of $\Delta$ that is thought to be plausible. In this model, $\Delta = 0$ corresponds to MAR; hence, examining inferences about $\mu$ over a set or range for $\Delta$ that includes $\Delta = 0$ will summarize the effects of departures from MAR on inferences about $\mu$.

For fixed $\Delta$, assumption (4) is of type (i). In this simple setting, type (ii) assumptions are not needed because $\mu_1$ and $\pi$ can be estimated with sample means, and no modeling is needed.

Finally, to test for treatment effects between two arms, one adopts a value $\Delta_0$ for the first arm and a value $\Delta_1$ for the second arm. One then estimates each mean separately under the adopted values of $\Delta$ and conducts a Wald test that their difference is zero. To investigate how the conclusions depend on the adopted values of $\Delta$, one repeats the testing over a range of plausible values for the pair $(\Delta_0, \Delta_1)$.

### Selection Model Approach

A second option for conducting sensitivity analysis is to assume that one knows how the odds of nonresponse change with the values of the outcome $Y$. For example, one can assume that the log odds of nonresponse differs by $\alpha$ for those who differ by one unit on $Y$. This is equivalent to assuming that one knows the value of $\alpha$ (but not $h$) in the logistic regression model

$$\text{logit}\,\{P[R = 0 \mid Y = y]\} = h + \alpha\, y. \tag{6}$$

Models like (6) are called selection models because they model the probability of nonresponse (or selection) as a function of the outcome. Each unique value of $\alpha$ corresponds to a different unverifiable assumption about how the probability of nonresponse changes with the outcome.

The model in (6) is also equivalent to assuming that

$$p\,(y \mid R = 0) = p(y \mid R = 1) \times \exp(\alpha\, y) \times const. \tag{7}$$

Adopting a value of $\alpha$ is equivalent to adopting a known link between the distribution of the respondents and that of the nonrespondents, because one

cannot use the data to learn anything about the nonrespondent distribution or to check the value of $\alpha$. Moreover, one cannot check two other important assumptions: that the log odds of nonresponse is linear in $y$ and that the support of the distribution of $Y$ among nonrespondents is the same as that among respondents (as implied by (7)).

Although not immediately apparent, once a value of $\alpha$ is adopted, one can estimate $\mu = E[Y]$ consistently. A sensitivity analysis consists of repeating the estimation of $\mu$ at different plausible values of $\alpha$ so as to assess the sensitivity of inferences about $\mu$ to assumptions about the missing data mechanism as encoded by $\alpha$ and model (6).

Estimation of $\mu$ relies on the identity

$$E(Y) = E\left\{\frac{R \times Y}{P(R=1 \mid Y)}\right\}, \tag{8}$$

which suggests estimation of $\mu$ through inverse probability weighting (see below); in this case, the weights can depend on missing values of $Y$. The inverse probability weighting estimator is

$$\hat{\mu}_{IPW} = \sum_i \frac{R_i Y_i}{1 - \text{expit}\left(\hat{h} + \alpha Y_i\right)}, \tag{9}$$

where $\text{expit}(u) = \text{logit}^{-1}(u) = \exp(u) / \{1 + \exp(u)\}$. To compute $\hat{h}$, one solves the unbiased estimating equation

$$\sum_i \left\{\frac{R_i}{1 - \text{expit}\left(h + \alpha Y_i\right)}\right\} = 0 \tag{10}$$

for $h$.[1] Analytic formulas for consistent standard error estimators are available (e.g., Rotnitzky et al., 1998), but bootstrap resampling can be used. Sensitivity analysis for tests of treatment effects proceeds by repeating the test over a set of plausible values for $\alpha$, where different values of $\alpha$ can be chosen for each arm.

With the selection model approach described here we can conduct sensitivity analysis, not just about the mean but about any other component of the distribution of $Y$, for example, the median of $Y$. Just as in the preceding pattern mixture approach, the data structure in this setting is so simple that we need not worry about postulating type (ii) assumptions.

[1]Estimation of $h$ by standard logistic regression of $R$ on $Y$ is not feasible because $Y$ is missing when $R = 0$; the estimator $\hat{h}$ exploits the identity $E\left[\frac{R}{p\{R=1 \mid Y\}}\right] = 1$.

## EXAMPLE: SINGLE OUTCOME WITH AUXILIARY DATA

We next consider a setting in which individuals are scheduled to have a measurement $Y_0$ at baseline, which we assume is never missing (this constitutes the auxiliary data), and a second measurement $Y_1$ at some specified follow-up time, which is missing in some subjects. We let $R_1 = 1$ if $Y_1$ is observed and $R_1 = 0$ otherwise. As in the preceding example, we limit our discussion to estimation of the arm-specific mean of $Y_1$, denoted now by $\mu = E(Y_1)$.

In this example, the type (i) MAR assumption states that, within each treatment group and within levels of $Y_0$, the distribution of $Y_1$ among nonrespondents is the same as the distribution of $Y_1$ among respondents. That is,

$$[Y_1 \mid Y_0, X, R = 1] = [Y_1 \mid Y_0, X, R = 0]. \tag{11}$$

### Pattern Mixture Model Approach

In this and the next section, we demonstrate sensitivity analysis under MNAR. Under the pattern mixture approach one specifies a link between the distribution of $Y_1$ in the nonrespondents and respondents who share the same value of $Y_0$. One can specify, for example, that

$$E(Y_1 \mid Y_0, R_1 = 0) = g^{-1}[g\{\eta(Y_0)\} + \Delta], \tag{12}$$

where $\eta(Y_0) = E(Y_1 \mid R_1 = 1, Y_0)$ and $g$ is defined as in the example above.

***Example: Continuous Values of Y*** Suppose $Y_1$ is continuous. One needs a specification of both the sensitivity analysis function $g$ and the relationship between $Y_1$ and $Y_0$, represented by $\eta(Y_0)$. A simple version of $\eta$ is a regression of $Y_1$ on $Y_0$,

$$\eta(Y_0) = E(Y_1 \mid Y_0, R = 1) \tag{13}$$

$$= \beta_0 + \beta_1 Y_0. \tag{14}$$

Now let $g(u) = u$ as in the first example above. In this case, using (12), the mean of the missing $Y_1$ are imputed as regression predictions of $Y_1$ plus a shift $\Delta$,

$$E(Y_1 \mid Y_0, R_1 = 0) = \beta_0 + \beta_1 Y_0 + \Delta. \tag{15}$$

Hence, at a fixed value of $\Delta$, an estimator of $E_\Delta(Y_1 \mid R_1 = 0)$ can be derived as the sample mean of the regression predictions $\hat{Y}_1 = \hat{\beta}_0 + \hat{\beta}_1 Y_0 + \Delta$ among those with $R_1 = 0$. The estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ come from a regression of $Y_1$ on $Y_0$ among those with $R_1 = 1$.

In this case, $\Delta$ represents the baseline adjusted difference in the mean of $Y_1$ between nonrespondents and respondents. If $\Delta > 0$ ($< 0$), then for any fixed value of $Y_0$, the mean of $Y_1$ among nonrespondents is $\Delta$ units higher (lower) than the mean of $Y_1$ among respondents.

A few comments are in order for this example:

- Model (12) assumes that mean differences do not depend on $Y_0$. If one believes that they do, then one may choose a more complex version of the $g$ function, such as

$$E(Y_1 \mid Y_0, R_1 = 0) = g^{-1}[g\{\eta(Y_0)\} + \Delta_0 + \Delta_1 Y_0]. \tag{16}$$

If this version is coupled with a linear regression for $\eta(Y_0)$, then both the slope and the intercept of that regression will differ for respondents and nonrespondents.

- In general, any user-specified sensitivity function $d(Y_0,\Delta)$ can be posited, including the simple versions $d(Y_0,\Delta) = \Delta$ and $d(Y_0,\Delta) = \Delta_0 + \Delta_1 Y_0$. Importantly, no version of $d(Y_0,\Delta)$ can be checked using the observed data. The choice of $d$ function is a type (i) assumption.
- Likewise, more general choices can be made for the form of $\eta(Y_0)$, including versions that are nonlinear in $Y_0$. The choice of $\eta$ is a type (ii) assumption; it can be critiqued by standard goodness-of-fit procedures using the observed data.

***Example: Binary Outcome Y*** If $Y$ is binary, the functional form of $g$ and $\eta$ will need to be different than in the continuous case. Choosing $g(u) = \log(u/(1 + u))$ implies that $\Delta$ is the log odds ratio comparing the odds of $Y_1 = 1$ between respondents and nonrespondents, conditional on $Y_0$. As with the continuous case, $\Delta > 0$ ($\Delta < 0$) implies that, for every level of $Y_0$, nonrespondents are more (less) likely to have $Y_1 = 1$ than respondents.

The function $\eta(Y_0)$, which describes $E(Y_1 \mid Y_0, R = 1)$, should be specified in terms of a model that is appropriate for binary outcomes. For example, a simple logistic specification is

$$\text{logit}\{\eta(Y_0)\} = \lambda_0 + \lambda_1 Y_0, \tag{17}$$

which is equivalent to writing

$$\eta(Y_0) = \frac{\exp(\lambda_0 + \lambda_1 Y_0)}{1 + \exp(\lambda_0 + \lambda_1 Y_0)}. \quad (18)$$

When $Y_0$ is binary, this model is saturated. But when $Y_0$ is continuous, or includes other auxiliary covariates, model choice for $\eta$ will take on added importance.

***Inference*** A sensitivity analysis to examine how inferences are impacted by the choice of $\Delta$ consists of repeating the inference over a set or range of values of $\Delta$ deemed to be plausible. It can proceed in the following manner:

Step 1. Specify models for $\eta(Y_0)$ and $d(Y_0,\Delta)$.

Step 2. Fit the model $\eta(Y_0)$ to those with $R_1 = 0$, and obtain the estimated function $\hat{\eta}(Y_0)$.

Step 3. The full-data mean $\mu = E(Y_1)$ is

$$\mu = \pi E\{\eta(Y_0) \mid R=1\} + (1-\pi)E\left(g^{-1}\left[g\{\eta(Y_0)\} + d(Y_0,\Delta)\right] \mid R=0\right), \quad (19)$$

where expectations are taken over the distribution of $Y_0 \mid R$. Although the general formula looks complex, it is easily computed for a fixed value of $\Delta$ once the model for $\eta$ has been fit to data. Specifically,

Step 3a. The estimate of $E\{\eta(Y_0) \mid R=1\}$ is the sample mean $\sum_i R_i \hat{\eta}(Y_{0i}) \Big/ \sum_i R_i$.

Step 3b. The estimate of $E\left(g^{-1}\left[g\{\eta(Y_0)\} + d(Y_0,\Delta)\right] \mid R_1 = 0\right)$ also is computed as a sample mean,

$$\frac{\sum_i (1-R_i) g^{-1}\left[g\{\hat{\eta}(Y_{0i})\} + d(Y_{0i},\Delta)\right]}{\sum_i (1-R_i)}. \quad (20)$$

Step 3c. The estimate of $\pi$ is $\hat{\pi} = (1/n)\sum_i R_i$.

Step 3d. The estimate $\hat{\mu}$ of $\mu$ is computed by replacing parameters in (19) by their estimators described in the previous steps.

Step 4. Standard errors are computed using bootstrap resampling.

Step 5. Inferences about $\mu$ are carried out for a plausible set or range of values of $\Delta$. Because each unique value of $\Delta$ yields an estimator $\hat{\mu}_\Delta$, it is possible to construct a contour plot of $Z$-scores, p-values, or confidence

intervals for treatment effect as a function of Δ. An illustration, computed using data from a diabetic neuropathy trial, appears in Figure 5-1.

### Selection Model Approach

In parallel to the first example, with no auxiliary data, another way to postulate type (i) assumptions about the nature of selection bias is by postulating a model for the dependence of the probability of nonresponse on the (possibly missing) outcome $Y_1$, within levels of $Y_0$. For example, one can assume that, conditionally on $Y_0$, each unit increase in $Y_1$ is associated with an increase of $\alpha$ in the log odds of nonresponse. That is,

$$\frac{odds\left(R_1 = 0 \mid Y_0, Y_1 + 1\right)}{odds\left(R_1 = 0 \mid Y_0, Y_1\right)} = \exp\left(\alpha\right), \tag{21}$$

or, equivalently,

$$\text{logit}\left\{P\left(R = 0 \mid Y_0, Y_1\right)\right\} = h\left(Y_0\right) + \alpha Y_1, \tag{22}$$

where $h(Y_0)$ is an unknown function of $Y_0$. This can also be written as

$$p\left(y_1 \mid R_1 = 0, Y_0\right) = p\left(y_1 \mid R_1 = 1, Y_0\right) \times \exp\left(\alpha y_1\right) \times h\left(Y_0\right). \tag{23}$$

In this latter form,[2] one can see that the observed data have no information about $\alpha$. The choice of $\alpha = 0$ specifies that within levels of $Y_0$, $R_1$ and $Y_1$ are independent (i.e., MAR). Values of $\alpha \neq 0$ reflect residual association between missingness and nonresponse after adjusting for $Y_0$.[3]

Analogous to the example with no auxiliary data, estimation of $\mu = E(Y_1)$ relies on the identity

$$E\left(Y_1\right) = E\left\{\frac{R_1 Y_1}{P\left(R_1 = 1 \mid Y_1, Y_0\right)}\right\}, \tag{24}$$

which suggests the inverse probability weighted (IPW) estimator

[2]The constant $h(Y_0)$ is $[E\{\exp(\alpha Y_1) \mid R_1 = 1, Y_0\}]^{-1}$, which ensures that $p(Y_1 \mid R_1 = 0, Y_0)$ is a density.

[3]If one believes that the association between nonresponse and outcome varies according to the baseline measurement $Y_0$, one can replace $\alpha$ with a known function of $Y_0$—for instance, $\alpha_0 + \alpha_1 Y_0$, with $\alpha_0$ and $\alpha_1$ having specified values. Regardless of the choice, once the values of $\alpha$ are fixed, $\mu = E(Y_1)$ can be written purely in terms of the distribution of the observed data and is therefore identified.

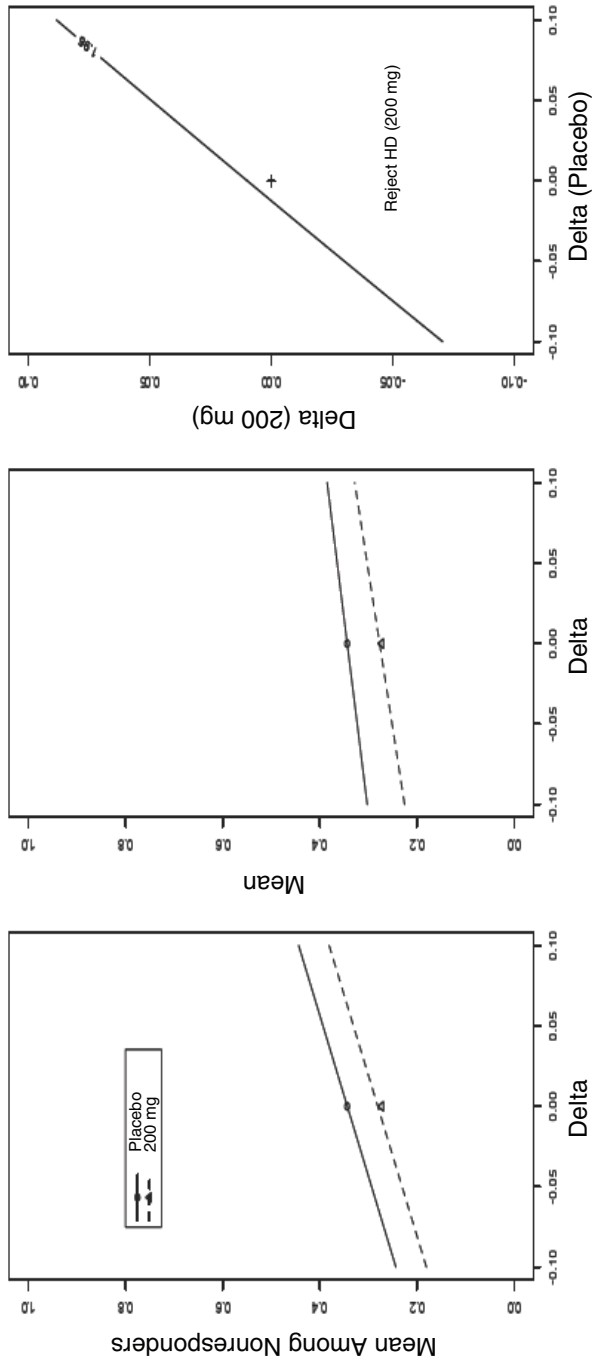


**FIGURE 5-1** Pattern mixture sensitivity analysis. Left panel: plot of mean outcome among nonrespondents as a function of sensitivity parameter Δ, where Δ = 0 corresponds to MAR. Center panel: plot of full-data mean as function Δ. Right panel: contour of Z statistic for comparing placebo to active treatment, where Δ is varied separately by treatment.

$$\hat{\mu}_{IPW} = (1/n)\sum_i \frac{R_i Y_i}{1-\text{expit}\left\{\hat{h}(Y_{0i}) + \alpha Y_{1i}\right\}}, \tag{25}$$

where $\hat{h}(Y_0)$ is an estimator of $h(Y_0)$.

Unless $Y_0$ is discrete with a few levels, estimation of $h(Y_0)$ requires the assumption that $h(Y_0)$ takes a known form, such as $h(Y_0;\gamma) = \gamma_0 + \gamma_1 Y_0$. (Note that if one adopts this model, one is assuming that the probability of response follows a logistic regression model on $Y_0$ and $Y_1$ with a given specified value for the coefficient $\alpha$ of $Y_1$.) Specifying $h(Y_0)$ is a type (ii) assumption that is technically not needed to identify $\mu$ but is needed in practical situations involving finite samples.

One can compute an estimator $\hat{\gamma}$ of $\gamma$ by solving a set of estimating equations[4] for $\gamma$,

$$\sum_i \frac{\partial h(Y_{0i},\gamma)}{\partial \gamma}\left[\frac{R_{1i}}{1-\text{expit}\left\{h(Y_{0i};\gamma) + \alpha Y_{1i}\right\}} - 1\right] = 0. \tag{26}$$

Formulas for sandwich-type standard error estimators are available, but the bootstrap can also be used to compute standard error estimates. Hypothesis-testing sensitivity analysis is conducted in a manner similar to the one described in the example above with no auxiliary data.

As with the pattern mixture models, by repeating the estimation of $\mu$ at a set or interval of known $\alpha$ values, one can examine how different degrees of residual association between nonresponse and the outcome $Y_1$ affect inferences concerning $E(Y_1)$. A plot similar to the one constructed for the pattern mixture model is given in Figure 5-2.

## EXAMPLE: GENERAL REPEATED MEASURES SETTING

As the number of planned measurement occasions increases, the complexity of the sensitivity analysis grows because the number of missing data patterns grows. As a result, there can be limitless ways of specifying models.

Consider a study with $K$ scheduled postbaseline visits. In the special case of monotone missing data, there are $(K + 1)$ patterns representing each of the visits at which a subject might last be seen, that is, 0,...,$K$. The

[4]As with the selection approach of the example with no auxiliary data, to estimate $\gamma$ one cannot fit a logistic regression model because $Y_1$ is missing when $R_1 = 0$. The estimator $\hat{\gamma}$ exploits the identity $E\left[\frac{R_1}{P(R_1 = 1 \mid Y_0, Y_1)} \mid Y_0\right] = 1.$

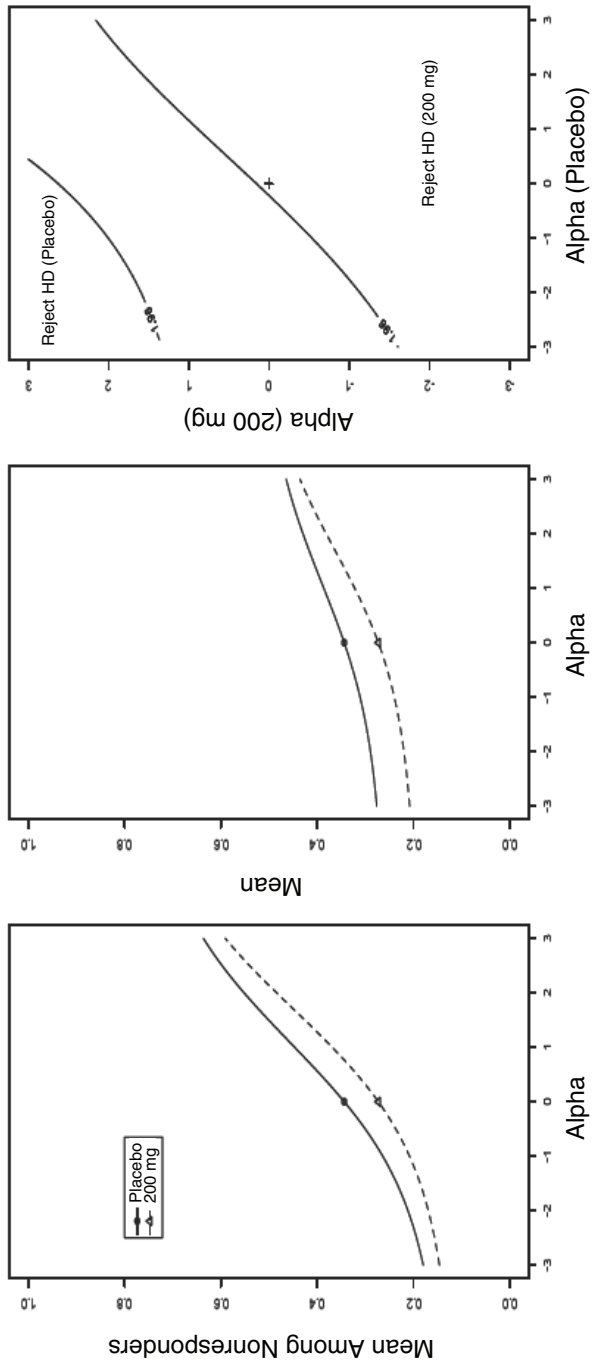


**FIGURE 5-2** Selection model sensitivity analysis. Left panel: plot of mean outcome among nonrespondents as a function of sensitivity parameter $\alpha$, where $\alpha = 0$ corresponds to MAR. Center panel: plot of full-data mean as function of $\alpha$. Right panel: contour of Z statistic for comparing placebo to active treatment where $\alpha$ is varied separately by treatment.

$(K + 1)^{st}$ pattern represent subjects with complete data, while the other $K$ patterns represent those with varying degrees of missing data. In the general setting, there are many ways to specify pattern models—the models that link the distribution of missing outcomes to the distribution of observed outcomes within specified strata—and it is generally necessary to look for simplifications of the model structure.

For example, one could link the conditional (on a shared history of observed outcomes through visit $k - 1$) distribution of missing outcomes at visit $k$ among those who were last seen at visit $k - 1$ to (a) the distribution of outcomes at visit $k$ among those who complete the study, (b) the distribution of outcomes at visit $k$ among those who are in the study through visit $k$, or (c) the distribution of outcomes at visit $k$ among those who are last seen at visit $k$.

Let $Y_k$ denote the outcome scheduled to be measured at visit $k$, with visit 0 denoting the baseline measure. We use the notation $Y_k^- = (Y_0,\ldots,Y_k)$ to denote the history of the outcomes through visit $k$ and $Y_k^+ = (Y_{k+1},\ldots,Y_K)$ to denote the future outcomes after visit $k$. We let $R_k$ denote the indicator that $Y_k$ is observed, so that $R_k = 1$ if $Y_k$ is observed and $R_k = 0$ otherwise. We assume that $Y_0$ is observed on all individuals so that $R_0 = 1$. As above, we focus on inference about the mean $\mu = E(Y_K)$ of the intended outcome at the last visit $K$.

### Monotone Missing Data

Under monotone missingness, if the outcome at visit $k$ is missing, then the outcome at visit $k + 1$ is missing. If we let $L$ be the last visit that a subject has a measurement observed, then the observed data for a subject is $Y_L^- = (Y_0,\ldots,Y_L)$, where $L \leq K$.

### A Pattern Mixture Model Approach

As noted above, there are many pattern models that can be specified. Here, we discuss inference in one such model. Recall that both type (i) and type (ii) assumptions are needed. We first address type (i) specification, illustrating a way to link distributions with those having missing observations to those with observed data.

The general strategy is illustrated for the case $K = 3$, which relies on an assumption known as "nonfuture dependence" (Kenward et al., 2003). In simple terms, the nonfuture dependence assumption states that the probability of drop out at time $L$ can only depend on observed data up to $L$ and the possibly missing value of $Y_L$, but not future values of $L$.

In the model used here, we assume there is a link between $\left[Y_k \mid Y_k^-, L=k-1\right]$ and $\left[Y_k \mid Y_k^-, L>k-1\right]$, which are, respectively, the distributions of $Y_k$ among those who do and do not drop out at time $k-1$. The idea is to use the distribution of those still in the study at time $k-1$ to identify the distribution of those who drop out at $k-1$.

It can be shown that $\mu = E(Y_K)$ can be estimated by a recursion algorithm, provided the following observed-data distributions are estimated:

$$\left[Y_3 \mid Y_0, Y_1, Y_2, L=4\right], \left[Y_2 \mid Y_0, Y_1, L\geq 3\right], \left[Y_1 \mid Y_0, L\geq 2\right], \left[Y_0 \mid L\geq 1\right], \quad (27)$$

and the following dropout probabilities

$$\begin{gathered} P\left(L=4 \mid L\geq 3, Y_0, Y_1, Y_2, Y_3\right), P\left(L=3 \mid L\geq 2, Y_0, Y_1, Y_2\right), \\ P\left(L=2 \mid L\geq 1, Y_0, Y_1\right), P\left(L=1 \mid Y_0\right), \end{gathered} \quad (28)$$

can also be estimated. Each is identified from observed data when missingness is monotone.

What is needed to implement the estimation of $\mu = E(Y_K)$ is a model that links the distributions with observed data (27) to the distributions having missing observations. One simple way to do this is to assume the distribution of $Y_k$ among recent dropouts, $\left[Y_k \mid Y_{k-1}^-, L=k-1\right]$, follows the same parametric model as the distribution of $Y_k$ among respondents, $\left[Y_k \mid Y_{k-1}^-, L>k-1\right]$, but with a different—or shifted—parameter value. This assumption cannot be verified and may not be realistic in all studies; we use it here simply as an illustration.

To be more concrete, suppose that the outcomes $Y_0,\ldots,Y_3$ are continuous. One can assume regression models for each of (27) as follows,

$$E\left(Y_0 \mid L\geq 1\right) = \mu_0, \quad (29)$$

$$E\left(Y_1 \mid Y_0, L\geq 2\right) = \mu_1 + \beta_1 Y_0, \quad (30)$$

$$E\left(Y_2 \mid Y_0, Y_1, L\geq 3\right) = \mu_2 + \beta_2 \left(Y_0, Y_1\right)^T, \quad (31)$$

$$E\left(Y_3 \mid Y_0, Y_1, Y_2, L=4\right) = \mu_3 + \beta_3 \left(Y_0, Y_1, Y_2\right)^T. \quad (32)$$

This modeling of the observed data distribution comprises our type (i) assumptions. These can (and must) be checked using the observables.

Using type (ii) assumptions, the distributions of missing $Y$ can be linked in a way similar to those for the first example above. For example,

those with $L = 1$ are missing $Y_1$. One can link the observed-data regression $E(Y_1 \mid Y_0, L \geq 2)$ to the missing-data regression $E(Y_1 \mid Y_0, L = 1)$ through

$$E(Y_1 \mid Y_0, L = 1) = \mu_1^* + \beta_1^* Y_0, \tag{33}$$

where, say, $\mu_1^* = \mu_1 + \Delta_{\mu_1}$ and $\beta_1^* = \beta_1 + \Delta_{\beta_1}$. Models for missing $Y_2$ and $Y_3$ can be specified similarly.

As with the previous cases, (33) is a type (ii) assumption and cannot be checked with data. Moreover, even using a simple structure like (33), the number of sensitivity parameters grows large very quickly with the number of repeated measures. Hence, it is important to consider simplifications, such as setting $\Delta_\beta = 0$, assuming $\Delta_\mu$ is equivalent across patterns, or some combination of the two.

Note that under our assumptions, $\Delta_{\mu_k}$ is the difference between the mean of $Y_k$ among those who drop out at $k - 1$ and those who remain beyond $k - 1$, conditional on observed data history up to $k - 1$. In this example, the assumption of linearity in the regression models, combined with an assumption that $\Delta_{\beta_k} = 0$ for all $k$, means that one does not need a model for $P(L = k \mid L \geq k, Y_k^-)$ to implement the estimation via recursion algorithm.

A sensitivity analysis consists of estimating $\mu$ and its standard error repeatedly over a range of plausible values of specified $\Delta$ parameters. For this illustration, setting $\Delta = 0$ implies MAR.[5]

### Selection Model

Another way to posit type (i) assumptions in this setting is to postulate a model for how the odds of dropping out between visits $k$ and $k + 1$, depends on the (possibly missing) future outcomes, $Y_k^+$, given the recorded history $Y_k^-$. That is,

[5]An attractive feature of the pattern mixture approach we consider here (the one that links the distribution of outcomes between dropouts at a given time and those who remain in the study at that time) is that the special choice of link that specifies that these two distributions are the same is tantamount to the MAR assumption (i.e., the assumption that at any given occasion the past recorded data are the only predictors of the future outcomes that are used to decide whether or not to drop out of the study at that time). This feature does not hold with other choices of pattern mixture models. Thus, in our example, exploring how inferences about $\mu$ change as $\Delta_k$ moves away from $\Delta_k = 0$ is tantamount to exploring the impact of distinct degrees of residual dependence between the missing outcomes and dropping out on our inferences about $\mu$. In more general pattern mixture models, $\Delta = 0$ is only sufficient, but not necessary, for MAR to hold. It is possible to find other unique combinations of $\Delta$ that correspond to MAR.

$$odds\left(L=k \mid L\geq k, Y_k^-, Y_k^+\right)=\frac{P\left(L=k \mid L\geq k, Y_k^-, Y_k^+\right)}{P\left(L>k \mid L\geq k, Y_k^-, Y_k^+\right)}.$$

The MAR assumption states that the odds do not depend on the future outcomes $Y_k^+$. The nonfuture dependence assumption above states that it depends only on the future through $Y_{k+1}$. That is,

$$odds\left(L=k \mid L\geq k, Y_k^-, Y_k^+\right)=odds\left(L=k \mid L\geq k, Y_k^-, Y_{k+1}\right) \tag{34}$$

and is equivalent to assuming that after adjusting for the recorded history, the outcome to be measured at visit $k + 1$ is the only predictor of all future missing outcomes that is associated with the odds of dropping out between visits $k$ and $k + 1$.

This last assumption, coupled with an assumption that quantifies the dependence of the odds in the right hand side on $Y_{k+1}$, suffices to identify $\mu = E(Y_K)$: in fact, it suffices to identify $E(Y_k)$ for any $k = 1,...,K$. For example, one might assume

$$\frac{odds\left(L=k \mid L\geq k, Y_k^-, Y_{k+1}+1\right)}{odds\left(L=k \mid L\geq k, Y_k^-, Y_{k+1}\right)}=\exp(\alpha), \tag{35}$$

that is, that each unit increase in $Y_{k+1}$ is associated with a constant increase in the log odds of nonresponse of $\alpha$, the same for all values of $Y_k^-$ and all visits $k$.

Under (34), $\alpha = 0$ implies MAR. One would make this choice if it is believed that the recorded history $Y_k^-$ encodes all the predictors of $Y_{k+1}$ that are associated with missingness. Values of $\alpha \neq 0$ reflect residual association of dropping out between visits $k$ and $k + 1$ and the possibly unobserved outcome $Y_{k+1}$, after adjusting for previous outcomes, and hence the belief that dropping out cannot be entirely explained by the observed recorded history $Y_k^-$. By repeating estimation of the vector $\mu$ for each fixed $\alpha$, one can examine how different degrees of residual association between dropping out and outcome at each occasion after adjusting for the influence of recorded history affects inferences concerning $\mu$.

Assumptions (34) and (35) together are equivalent to specifying that

$$\text{logit}\left\{P\left(L=k \mid L\geq k, Y_k^-, Y_k^+\right)\right\}=h_k\left(Y_k^-\right)+\alpha Y_{k+1}, \tag{36}$$

where $h_k\left(Y_k^-\right)$ is an unknown function of $Y$. This, in turn, is equivalent to the pattern mixture model

$$p\left(y_{k+1} \mid L=k, Y_k^-\right)=p\left(y_{k+1} \mid L\geq k+1, Y_k^-\right)\times\exp\left(\alpha y_{k+1}\right)\times const. \quad (37)$$

In this latter form, one can see that there is no evidence in the data regarding $\alpha$ since it serves as the link between the conditional (on $Y_k^-$) distribution of $Y_{k+1}$ among those who drop out between visits $k$ and $k + 1$ and those who remain through visit $k + 1$. If one believes that the association between dropping out and future outcomes depends solely on the current outcome but varies according to the recorded history, one can replace $\alpha$ with a known function of $Y_k^-$.

For instance, replacing in equation (36) the constant $\alpha$ with the function $\alpha_0 + \alpha_1 Y_k$ with $\alpha_0$ and $\alpha_1$ specified, encodes the belief that the residual association between dropping out between $k$ and $k + 1$ and the outcome $Y_{k+1}$ may be stronger for individuals with, say, higher (if $\alpha_1 > 0$) values of the outcome at visit $k$. As an example, if $Y_k$ is a strong predictor of $Y_{k+1}$ and that lower values of $Y_{k+1}$ are preferable (e.g., HIV-RNA viral load), then it is reasonable to postulate that subjects with low values of $Y_k$ drop out for reasons unrelated to the drug efficacy (and, in particular, then to their outcome $Y_{k+1}$) while subjects with higher values of $Y_k$ drop out for reasons related to drug efficacy and hence to their outcome $Y_{k+1}$.

Regardless of how the residual dependence is specified, $\mu$ can be expressed in terms of the distribution of the observed data, that is, it is identified. Estimation of $\mu = E[Y_K]$ relies on the identity

$$E\left[Y_K\right]=E\left\{\frac{R_K Y_K}{\pi_K\left(Y_K^-;h_1,\ldots,h_K;\alpha\right)}\right\}, \quad (38)$$

where

$$\pi_K\left(Y_K^-;h_1,\ldots,h_K;\alpha\right)=\prod_{k=0}^{K-1}\left[1-\text{expit}\left\{h_k\left(Y_K^-\right)+\alpha Y_{k+1}\right\}\right].$$

This formula suggests that one can estimate $\mu$ with the IPW estimator

$$\hat{\mu}_{IPW}=\left(1/n\right)\sum_i \frac{R_{iK} Y_{iK}}{\pi_k\left(Y_{iK}^-;\hat{h}_{i1},\ldots,\hat{h}_{iK};\alpha\right)}. \quad (39)$$

This estimator relies on estimators $\hat{h}_{ik}=h_k\left(Y_{ik}^-\right)$. In order to estimate these functions, one needs to impose type (ii) modeling assumptions on $h_k\left(Y_k^-\right)$, that is, $h_k\left(Y_k^-\right)=h_k\left(Y_k^-;\gamma_k\right)$. For example, one can assume that $h_k\left(Y_k^-\right)=\gamma_{0,k}+\gamma_{1,k}Y_k$ (adopting such a model would be tantamount to assuming that the probability of dropping out at each time follows a logistic regression model on just the immediately preceding recorded data and on the current outcome).

As with the selection approach of the two preceding examples, to estimate $\gamma_k$, one cannot fit a logistic regression model because $Y_{K+1}$ is missing when $L = k$. However, one can estimate it instead by solving the estimating equations

$$\sum_{i=1}^{n}\sum_{k=0}^{K-1} R_{ik}\frac{\partial h_k\left(Y_{ik}^{-};\gamma_k\right)}{\partial \gamma_k}\left[\frac{R_{i,k+1}}{1-\operatorname{expit}\left\{Y_{ik}^{-};\gamma_k\right\}+\alpha Y_{i,k+1}}-1\right]. \qquad (40)$$

for $\gamma$, justified on similar grounds as the estimators of $h$ functions in the previous examples.

Formula for sandwich-type standard error estimators are available (see Rotnitzky et al., 1997), but the bootstrap can also be used to compute standard error estimates. Sensitivity analysis with regard to hypothesis testing is conducted in a manner similar to the one described in the first example above.

### Nonmonotone Missing Data

A typical way to analyze nonmonotone missing data is to treat the time of dropout as the key missing data variable and then to assume MAR within dropout pattern (or conditional on dropout time). The advantage of this approach is purely practical: It interpolates missing data under a specified model. That said, however, the current literature suggests that MAR within pattern does not easily correspond to realistic mechanisms for generating the data. This raises concern among members of the panel that nonmonotone dropouts may require more specialized methods for modeling the missing data mechanism, and accounting for departures from MAR.

This topic has not been deeply studied in the extant statistical literature, and in particular numerical studies are lacking. We recommend this as a key area of investigation that will: (a) examine the appropriateness of existing models and in particular the potential pitfalls of assuming MAR within missing data pattern; and (b) develop and apply novel, appropriate methods of model specification and sensitivity analysis to handle nonmonotone missing data patterns.

## COMPARING PATTERN MIXTURE AND SELECTION APPROACHES

The main appeal of the selection model approach is that, since it models the probability of nonresponse rather than the distribution of outcomes, it can easily accommodate vectors of auxiliary factors with components that can be of all types, discrete, categorical, and continuous.

Two disadvantages of the selection approach as they relate to drawing inferences are (a) the inverse weighting estimation procedure, which can yield relatively inefficient inferences (i.e., large standard errors), and (b) that model checking of the type (ii) assumptions must be conducted for each unique value of the sensitivity analysis parameters. Formal model checking procedures have yet to be formally developed for this setting. The inefficiencies associated with the inverse weighting procedure are mitigated in settings with a sizable fraction of missing data, as the sampling variability is often of less concern than the range of type (i) assumptions that are entertained. To address (b), one should fit a highly flexible model for the $h$ function in the selection model.

Another potential disadvantage of selection models relates to interpretation of the sensitivity parameter. Particularly for continuous measures, it may be difficult to interpret nonresponse rates on the odds scale and to specify reasonable ranges for the sensitivity parameter. Plots such as those shown in Figure 5-2 (above) can be helpful in understanding how values of the sensitivity parameter correspond to imputed means for the missing outcomes.

Advantages of the pattern mixture model include transparent interpretation of sensitivity parameters and straightforward model checking for the observed-data distribution. The sensitivity parameters are typically specified in terms of differences in mean between respondents and nonrespondents, which appeal directly to intuition and contributes to formulating plausible ranges for the parameter. Pattern mixture models also can be specified so that the fit to the observed data is identical across all values of the sensitivity parameters; hence, model checking will be straightforward and does not depend on the assumed missing data assumption.

Disadvantages of pattern mixture modeling include difficulties in including auxiliary information, which will generally require additional modeling. Computation of the weighted averages across patterns for models of large numbers of repeated measures also can become complex without significant simplifying assumptions.

## TIME-TO-EVENT DATA

A major challenge in the analysis of time-to-event outcomes in randomized trials is to properly account for censoring that may be informative. Different approaches have been proposed in the research literature to address this issue. When no auxiliary prognostic factors are available, the general strategy has been to impose nonidentifiable assumptions concerning the dependence between failure and censoring times and then vary these assumptions in order to assess the sensitivity of inferences on the estimated survivor function. When prognostic factors are recorded, Robins and colleagues in a series of papers (Robins and Rotnitzky, 1992; Robins, 1993;

Robins and Finkelstein, 2000) proposed a general estimation strategy under the assumption that all measured prognostic factors that predict censoring are recorded in the database. Scharfstein and Robins (2002) proposed a method for conducting sensitivity analysis under the assumption that some but not all joint prognostic factors for censoring and survival are available. Their approach is to repeat inference under different values of a nonidentifiable censoring bias parameter that encodes the magnitude of the residual association between survival and censoring after adjusting for measured prognostic factors.

In randomized studies, censoring typically occurs for several reasons, some noninformative, others informative. For instance, in studies with staggered entry, the administrative end of the follow-up period typically induces noninformative censoring. However, loss to follow-up due to dropouts induces a competing censoring mechanism that is likely to be informative. Treatment discontinuation might induce yet another informative censoring process.

Under the Scharfstein and Robins methodology, the analyst specifies a range for the parameter encoding the residual dependence of the hazard of the minimum of competing censoring times on the censored outcome. However, this range might be rather difficult to specify if the reasons that each censoring might occur are quite different, more so if some censoring processes are informative and some are not. To ameliorate this problem in studies with staggered entry, one can eliminate the censoring by the administrative end of the follow-up period (typically a source of noninformative censoring) by restricting the follow-up period to a shorter interval in which (with probability) no subject is administratively censored. However, in doing so, one would lose valuable information on the survival experience of the study patients who remain at risk at the end of the reduced analysis interval. Rotnitzky et al. (2007) provide estimators of the survival function under separate models for the competing censoring mechanisms, including both informative and noninformative censoring. The methods can be used to exploit the data recorded throughout the entire follow-up period and, in particular, beyond the end of the reduced analysis interval discussed above.

## DECISION MAKING

Even after model fitting and sensitivity analysis, investigators have to decide about how important the treatment effect is. Unfortunately, there is no scientific consensus on how to synthesize information from a sensitivity analysis into a single decision about treatment effect. At least three possibilities can be considered.

One possibility is to specify a plausible region for the sensitivity parameters and report estimates of the lower and upper bounds from this range. These

endpoints form bounds on the estimated treatment effect and would be used in place of point estimates. Accompanying these bounds would be a 95 percent confidence region. This procedure can be viewed as accounting for both sampling variability and variability due to model uncertainty (i.e., uncertainty about the sensitivity parameter value): see Molenberghs and Kenward (2007) for more detailed discussion and recommendations for computing a 95 percent confidence region.

A second possibility is to carry out inference under MAR and determine the set of sensitivity parameter values that would lead to overturning the conclusion from MAR. Results can be viewed as equivocal if the inference about treatment effects could be overturned for values of the sensitivity parameter that are plausible.

The third possibility is to derive a summary inference that averages over values of the sensitivity parameters in some principled fashion. This approach could be viewed as appropriate in settings in which reliable prior information about the sensitivity parameter value is known in advance.

Regardless of the specific approach taken to decision making, the key issue is weighting the results, either formally or informally, from both the primary analysis and each alternative analysis by assessing the reasonableness of the assumptions made in conjunction with each analysis. The analyses should be given little weight when the associated assumptions are viewed as being extreme and should be given substantial weight when the associated assumptions are viewed as being comparably plausible to those for the primary analysis. Therefore, in situations in which there are alternative analyses as part of the sensitivity analysis that support contrary inferences to that of the primary analysis, if the associated assumptions are viewed as being fairly extreme, it would be reasonable to continue to support the inference from the primary analysis.

## RECOMMENDATION

**Recommendation 15: Sensitivity analyses should be part of the primary reporting of findings from clinical trials. Examining sensitivity to the assumptions about the missing data mechanism should be a mandatory component of reporting.**

We note that there are some often-used models for the analysis of missing data in clinical trials for which the form of a sensitivity analysis has not been fully developed in the literature. Although we have provided principles for the broad development of sensitivity analyses, we have not been prescriptive for many individual models. It is important that additional research be carried out so that methods to carry out sensitivity analyses for all of the standard models are available.

# 6

# Conclusions and Recommendations

Missing data in clinical trials can seriously undermine the benefits provided by randomization into control and treatment groups. Two approaches to the problem are to reduce the frequency of missing data in the first place and to use appropriate statistical techniques that account for the missing data. The former approach is preferred, since the choice of statistical method requires unverifiable assumptions concerning the mechanism that causes the missing data, and so always involves some degree of subjectivity.

In Chapters 2 and 3, we detail some of the causes of missing data in clinical trials and discuss how to reduce the amount of missing data. However, because it is impossible to eliminate all occurrences of missing data, in Chapters 4 and 5 we discuss analysis methods that properly account for the missing outcome values.

In this concluding chapter, we bring all our recommendations together from the preceding chapters and offer three additional broad recommendations, two addressed to the U.S. Food and Drug Administration (FDA) and the companies that sponsor clinical trials. One is for the FDA and the National Institutes of Health to use their extensive database to develop a better understanding of the various causes of dropout from clinical trials, the typical extent of missing data in different types of trials, and the reductions in the rates of missing data that can be anticipated from the application of various alternative trial designs and techniques for trial conduct. A second recommendation is for the training of analysts in the latest techniques for the treatment of missing data in clinical trials. Finally, a third recommendation is for various research problems to be pursued.

## TRIAL OBJECTIVES

Questions such as whether to continue to collect trial outcome data after a participant has discontinued use of the study treatment, whether to use a single or composite outcome measure, how long to measure outcome data, all depend on the estimation goal of the trial. This estimation includes not only the outcome of interest, but also whether the focus is on short- or long-term effects of the intervention and the target population of interest. Possibilities for the latter include the "intent-to-treat" population, or the population of treatment compliers. Before selecting a trial design, it is important to decide on the primary parameter and population of interest, the "causal estimand." Once the estimand is decided, the clinical trial design can be optimized for the measurement of that estimand.

**Recommendation 1: The trial protocol should explicitly define (a) the objective(s) of the trial; (b) the associated primary outcome or outcomes; (c) how, when, and on whom the outcome or outcomes will be measured; and (d) the measures of intervention effects, that is, the causal estimands of primary interest. These measures should be meaningful for all study participants, and estimable with minimal assumptions. Concerning the latter, the protocol should address the potential impact and treatment of missing data.**

## REDUCING DROPOUTS THROUGH TRIAL DESIGN

The interpretation of the trial findings is more difficult when participants discontinue their assigned interventions before the end of the study. Therefore, the trial design should be selected to maximize the number of participants who are maintained on the study intervention throughout the duration of the trial.

**Recommendation 2: Investigators, sponsors, and regulators should design clinical trials consistent with the goal of maximizing the number of participants who are maintained on the protocol-specified intervention until the outcome data are collected.**

There is a key distinction between treatment dropout and analysis dropout, and although there are trials in which treatment dropout will understandably be substantial, there is very little reason for substantial amount of missing data, that is, analysis dropouts. Furthermore, for many trial estimands, the benefits of retaining participants in the study can be substantial, including to support an analysis of effectiveness (comparison

of treatment policies) and to be able to monitor side effects that occur after discontinuation of treatment.

**Recommendation 3: Trial sponsors should continue to collect information on key outcomes on participants who discontinue their protocol-specified intervention in the course of the study, except in those cases for which a compelling cost-benefit analysis argues otherwise, and this information should be recorded and used in the analysis.**

**Recommendation 4: The trial design team should consider whether participants who discontinue the protocol intervention should have access to and be encouraged to use specific alternative treatments. Such treatments should be specified in the study protocol.**

**Recommendation 5: Data collection and information about all relevant treatments and key covariates should be recorded for all initial study participants, whether or not participants received the intervention specified in the protocol.**

## REDUCING DROPOUTS THROUGH TRIAL CONDUCT

In addition to trial design, aspects of trial conduct can also substantially reduce the amount of missing data. Chapter 3 outlines specific trial conduct techniques that should be considered. Given the importance of reducing the frequency of missing data, the monitoring of missing data from the design stage throughout the conduct of a trial needs to be accounted for in the trial protocol.

**Recommendation 6: Study sponsors should explicitly anticipate potential problems of missing data. In particular, the trial protocol should contain a section that addresses missing data issues, including the anticipated amount of missing data, and steps taken in trial design and trial conduct to monitor and limit the impact of missing data.**

**Recommendation 7: Informed consent documents should emphasize the importance of collecting outcome data from individuals who choose to discontinue treatment during the study, and they should encourage participants to provide this information whether or not they complete the anticipated course of study treatment.**

**Recommendation 8: All trial protocols should recognize the importance of minimizing the amount of missing data, and, in particular, they**

**should set a minimum rate of completeness for the primary outcome(s), based on what has been achievable in similar past trials.**

## TREATING MISSING DATA

Missing data are often unavoidable, despite best efforts to reduce their occurrence in trial design and conduct. The validity of assumptions concerning the source of missing data can only be assessed jointly by both data analysts and clinicians. Therefore, it is important that the assumptions underlying any selected analysis technique be clearly articulated so that they can be evaluated by clinicians as well as by statistical analysts.

**Recommendation 9: Statistical methods for handling missing data should be specified by clinical trial sponsors in study protocols, and their associated assumptions stated in a way that can be understood by clinicians.**

Methods like last observation carried forward (LOCF) and baseline observation carried forward (BOCF) are commonly applied in situations in which their underlying assumptions are unrealistic. The analysis methods used should yield confidence intervals for the treatment effect that have the claimed coverage properties and tests should have their nominal size when data are missing.

**Recommendation 10: Single imputation methods like last observation carried forward and baseline observation carried forward should not be used as the primary approach to the treatment of missing data unless the assumptions that underlie them are scientifically justified.**

**Recommendation 11: Parametric models in general, and random effects models in particular, should be used with caution, with all their assumptions clearly spelled out and justified. Models relying on parametric assumptions should be accompanied by goodness-of-fit procedures.**

**Recommendation 12: It is important that the primary analysis of the data from a clinical trial should account for the uncertainty attributable to missing data, so that under the stated missing data assumptions the associated significance tests have valid type I error rates and the confidence intervals have the nominal coverage properties. For inverse probability weighting and maximum likelihood methods, this analysis can be accomplished by appropriate computation of standard errors, using either asymptotic results or the bootstrap. For imputation, it**

**is necessary to use appropriate rules for multiply imputing missing responses and combining results across imputed datasets because single imputation does not account for all sources of variability.**

**Recommendation 13: Weighted generalized estimating equations methods should be more widely used in settings when missing at random can be well justified and a stable weight model can be determined, as a possibly useful alternative to parametric modeling.**

One very useful source of information that appears to have been rarely used is the follow-up of a sample of participants who withdrew from a study. Such data could be very useful in determining reasons for withdrawal and their missing outcome measurements.

**Recommendation 14: When substantial missing data are anticipated, auxiliary information should be collected that is believed to be associated with reasons for missing values and with the outcomes of interest. This could improve the primary analysis through use of a more appropriate missing at random model or help to carry out sensitivity analyses to assess the impact of missing data on estimates of treatment differences. In addition, investigators should seriously consider following up all or a random sample of trial dropouts, who have not withdrawn consent, to ask them to indicate why they dropped out of the study, and, if they are willing, to collect outcome measurements from them.**

Given that the assumptions for the missing data mechanism cannot be validated, the sensitivity of inferences for treatment effects in clinical trials to those assumptions needs to be assessed.

**Recommendation 15: Sensitivity analyses should be part of the primary reporting of findings from clinical trials. Examining sensitivity to the assumptions about the missing data mechanism should be a mandatory component of reporting.**

## UNDERSTANDING THE CAUSES AND DEGREE OF DROPOUTS IN CLINICAL TRIALS

A crucial issue that sponsors must wrestle with in planning a clinical trial is how much missing data they are likely to experience, how much could be reduced through the use of various techniques (such as those outlined in this report), and consequently if they implement these various techniques, what degree of missingness is likely to remain. The answers to

these questions will help trial sponsors decide on how to plan for missing data when determining sample size, whether steps are needed to reduce the amount of missing data (some of which may be resource intensive), and the potential for lack of robustness of final estimates of intervention effects to assumptions about missing data. In addition, analysts need to know what assumptions about the missing data mechanisms are scientifically defensible.

Information from previously collected clinical studies would help in answering these questions. Although FDA retains data from all clinical trials over which it has oversight, the data are confidential to the company that sponsored the trial and are therefore not shared. And although there is some research on why participants drop out of different kinds of clinical trials, empirical evidence is lacking for many types of trials. There is a need for more standardized data collection, documentation, and analysis of the reasons for and the frequency of missing data. Systematic investigations of factors related to treatment dropout and withdrawal and to missing data more generally are needed.

A pharmaceutical company that has been researching interventions in a particular area for a long time may have internal data that can provide some of this information. However, if a company is small or has limited prior experience, having access to information from prior clinical trials would be extremely useful in trial design.

**Recommendation 16: The U.S. Food and Drug Administration and the National Institutes of Health should make use of their extensive clinical trial databases to carry out a program of research, both internal and external, to identify common rates and causes of missing data in different domains and how different models perform in different settings. The results of such research can be used to inform future study designs and protocols.**

While it is difficult to be specific, characteristics of a trial that has failed because of missing data concerns include (a) rates of missing data that are two to three times as large as the difference in rate of successful outcome between the groups; (b) differential rates of missing data across treatment arms, indicating a high likelihood that biases would not cancel out in the treatment comparison; (c) lack of a record about the reasons for missing data, making it unclear whether the mechanisms are MAR; (d) lack of auxiliary data that would be the basis for plausible missing data adjustments; (e) marginal treatment effects that might be easily overturned by uncertainty from missing data; and (f) inadequate analysis methods that do not reflect uncertainty from missing data.

Many of the analysis techniques described in Chapters 4 and 5 have been explored both theoretically and in applications in the research literature over the past 20 years. However, their applications to clinical trials have been limited. There seems to be a reticence on the part of biostatisticians, both at the drug and device companies and at FDA, to embrace these various techniques. We conjecture that this reticence may be a by-product of the regulatory environment, a result of the limited development of supporting software for newer methods, or a result of a lack of training and education. We believe that once these techniques are implemented in a few major clinical trials, any conservatism from the regulatory environment would be offset by the more effective use of the information provided by these techniques. Also, as pointed out in Chapters 4 and 5, software now exists for most of the techniques as part of the most commonly used statistical packages. The remaining possibility, training and education, can be addressed by making education at FDA a higher priority. It is also important that FDA's clinical reviewers have some understanding of modern analysis methods so that they can assist in the judgment as to the reasonableness of assumed missing data mechanisms.

**Recommendation 17: The U.S. Food and Drug Administration (FDA) and drug, device, and biologic companies that sponsor clinical trials should carry out continued training of their analysts to keep abreast of up-to-date techniques for missing data analysis. FDA should also encourage continued training of their clinical reviewers to make them broadly familiar with missing data terminology and missing data methods.**

Throughout this report, we have advocated that further research be carried out in a number of important areas. We have decided to bring together those calls for additional research here in this final chapter. Areas in need of further research include

(1) designs for the follow-up of participants in clinical trials who have dropped out of the study (e.g., referred to here as analysis dropouts) and who have not withdrawn their consent,

(2) collecting the typical rates and likely causes of missing data in various kinds of clinical trials,

(3) the effect of missing data on the power of clinical trials,

(4) how to set useful target rates and acceptable rates of missing data in clinical trials,

(5) the robustness of missing data methods such as inverse probability weighting methods and multiple imputation methods to assumptions,

(6) the assessment of goodness-of-fit for the parametric models used to analyze data from clinical trials (when there is missing data),

(7) the performance of double-robust procedures in comparison to more commonly used procedures,

(8) the impact of missingness in auxiliary variables on the various current methods, and ways of reducing the associated bias,

(9) methods of sensitivity analysis in clinical trials, particularly for nonmonotone patterns in longitudinal data,

(10) methods for assessing and limiting the impact of informative censoring for time-to-event outcomes, and

(11) how to develop effective decision rules based on the input from sensitivity analyses.

We have collected the highest priority of these calls for additional research in a final recommendation, adding to that a call for the development of the associated software tools.

**Recommendation 18: The treatment of missing data in clinical trials, being a crucial issue, should have a higher priority for sponsors of statistical research, such as the National Institutes of Health and the National Science Foundation. There remain several important areas where progress is particularly needed, namely: (1) methods for sensitivity analysis and principled decision making based on the results from sensitivity analyses, (2) analysis of data where the missingness pattern is nonmonotone, (3) sample size calculations in the presence of missing data, and (4) design of clinical trials, in particular plans for follow-up after treatment discontinuation (degree of sampling, how many attempts are made, etc.), and (5) doable robust methods, to more clearly understand their strengths and vulnerabilities in practical settings. The development of software that supports coherent missing data analyses is also a high priority.**

# References


Amemiya, T. (1984). Tobit models: A survey. *Journal of Econometrics, 24*, 3-61.

Anderson, T.W. (1957). Maximum likelihood estimation for the multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association, 52*, 200-203.

Angrist, J.D., Imbens, G.W., and Rubin, D.B. (1996). Identification of causal effects using instrumental variables (with discussion and rejoinder). *Journal of the American Statistical Association, 91*, 444-472.

Baker, S.G. (1997). Compliance, all-or-none. In *The Encyclopedia of Statistical Science, Volume I,* S. Kotz, C.R. Read, and D.L. Banks (eds.), pp. 134-138. New York: John Wiley and Sons.

Baker, S.G., and Lindeman, K.S. (1994). The paired availability design: A proposal for evaluating epidural analgesia during labor. *Statistics in Medicine, 13*(21), 2,269-2,278.

Bang, H., and Robins, J.M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics, 61*, 962-972.

Barnard, J., and Rubin, D.B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika, 86*, 949-955.

Birmingham, J., Rotnitzky, A., and Fitzmaurice, G. (2003). Pattern mixture and selection models for analyzing monotone missing data. *Journal of the Royal Statistical Society, Series B, 65*, 275-297.

Carlin, B.P., and Louis, T.A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis, 2nd edition*. Boca Raton, FL: Chapman and Hall, CRC Press.

Carpenter, J. (2009). *Statistical Methods for Clinical Studies with Missing Data: What's Hot, What's Cool, and What's Useful.* Unpublished paper. Available: http://www.iscb2009.info/RSystem/Soubory/Prez%20Tuesday/S18.1%20Carpenter.pdf.

Daniels, M.J., and Hogan, J.W. (2008). *Missing Data in Longitudinal Studies*. Boca Raton, FL: Chapman and Hall, CRC Press.

DeGruttola, V., and Tu, X.M. (1994). Modeling progression of CD4-lymphocyte count and its relationship to survival time. *Biometrics, 50*, 1,003-1,014.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B, 39*, 1-38.

Diggle, P., and Kenward, M.G. (1994). Informative dropout in longitudinal data analysis (with discussion). *Applied Statistics, 43*, 49-94.

Diggle, P.J., Heagerty, P., Liang K.Y., and Zeger, S.L. (2002). *The Analysis of Longitudinal Data*, *2nd Edition.* Oxford, England: Oxford University Press.

Drennan, K. (2003). *Pharma Wants You: Clinical Trials Are Agencies' New Proving Ground.* Pharmaceutical Executive. Available: http://www.corbettaccel.com/press/pdf/200304_Pharma_Exec.pdf.

Emmanuel, E.J. (2005). Undue inducement—Nonsense on stilts. *American Journal of Bioethics, 5*(5), 9-13.

European Medicines Evaluation Agency. (1998). *Statistical Principles for Clinical Trials; Step 5: Note for Guidance on Statistical Principles for Clinical Trials*. International Conference on Harmonisation (ICH) Topic E9. Available: http://www.ich.org/LOB/media/MEDIA485.pdf.

European Medicines Evaluation Agency. (2009). *Guideline on Missing Data in Confirmatory Clinical Trials*. Committee for Medical Products for Human Use (April). Available: http://www.ema.europa.eu/pdfs/human/ewp/177699endraft.pdf.

Finkelstein, D.M., and Wolfe, R.A. (1985). A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics, 41*, 933-945.

Fitzmaurice, G.M., Laird, N.M., and Rotnitzky, A.G. (1993). Regression models for discrete longitudinal responses. *Statistical Science, 8*, 284-309.

Fitzmaurice, G.M., Laird, N.M., and Ware, J.H. (2004). *Applied Longitudinal Analysis*. Hoboken, NJ: Wiley Interscience.

Follmann, D. (2006). Augmented designs to assess immune response in vaccine trials. *Biometrics, 62*, 1,162-1,169.

Frangakis, C.E., and Rubin, D.B. (2002). Principal stratification in causal inference. *Biometrics, 58*, 21-29.

Gelfand, A.E., and Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association, 85*, 398-409.

Gelfand, A.E., Hills, S.E., Racine-Poon, A., and Smith, A.F.M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association, 85*, 972-985.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis, 1*(3), 515-533.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2003). *Bayesian Data Analysis*, *2nd edition*. London, England: CRC Press.

Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs' distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6*, 721-741.

Gilbert, P.B., Bosch, R.J., and Hudgens, M.G. (2003). Sensitivity analysis for the assessment of causal vaccine effects on viral load in HIV vaccine trials. *Biometrics, 59*, 531-541.

Gilks, W.R., Wang, C.C., Yvonnet, B., and Coursaget, P. (1993). Random effects models for longitudinal data using Gibbs' sampling. *Biometrics, 49*, 441-453.

Glynn, R., Laird, N.M., and Rubin, D.B. (1986). Selection modeling versus mixture modeling with nonignorable nonresponse. In *Drawing Inferences from Self-Selected Samples,* H. Wainer (ed.), pp. 119-146. New York: Springer-Verlag.

Glynn, R.J., Laird, N.M., and Rubin, D.B. (1993). Multiple imputation in mixture models for nonignorable nonresponse with follow-ups. *Journal of the American Statistical Association, 88*, 984-993.

Greenlees, W.S., Reece, J.S., and Zieschang, K.D. (1982). Imputation of missing values when the probability of nonresponse depends on the variable being imputed. *Journal of the American Statistical Association, 77*, 251-261.

Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems (with discussion). *Journal of the American Statistical Association, 72*, 320-340.

Heagerty, P.J. (1999). Marginally specified logistic-normal models for longitudinal binary data. *Biometrics, 55*(3), 688-698.

Heagerty, P.J. (2002). Marginalized transition models and likelihood inference for categorical longitudinal data. *Biometrics, 58*, 342-351.

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables, and a simple estimator for such models. *Annals of Economic and Social Measurement, 5*, 475-492.

Heitjan, D.F. (1993). Ignorability and coarse data: Some biomedical examples. *Biometrics, 49*, 1,099-1,109.

Heitjan, D.F. (1994). Ignorability in general complete-data models. *Biometrika, 81*, 701-708.

Heitjan, D.F. (1997). Annotation: What can be done about missing data? Approaches to imputation. *American Journal of Public Health, 87*, 548-550.

Heitjan, D., and Rubin, D.B. (1991). Ignorability and coarse data. *Annals of Statistics, 19*, 2,244-2,253.

Hirano, K., Imbens, G., Rubin, D.B., and Zhao, X.H. (2000). Estimating the effect of an influenza vaccine in an encouragement design. *Biostatistics, 1*, 69-88.

Hogan, J.W., Roy, J., and Korkontzelou, C. (2004). Biostatistics tutorial: Handling dropout in longitudinal data. *Statistics in Medicine, 23*, 1,455-1,497.

Imbens, G.W., and Rubin, D.B. (1997a). Bayesian inference for causal effects in randomized experiments with noncompliance. *Annals of Statistics, 25*, 305-327.

Imbens, G.W., and Rubin, D.B. (1997b). Estimating outcome distributions for compliers in instrumental variables models. *Review of Economic Studies, 64*, 555-574.

International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. (1994). *ICH Harmonised Tripartite Guideline: Dose-Response Information to Support Drug Registration: E4*. Available: http://www.ema.europa.eu/pdfs/human/ich/037895en.pdf.

International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. (2001). *ICH Harmonised Tripartite Guideline: Choice of Control Group and Related Issues in Clinical Trials: E10*. Available: http://www.ema.europa.eu/pdfs/human/ich/036496en.pdf.

Jansen, I., Beunckens, C., Molenberghs, G., Verbeke, G., and Mallinckrodt, C. (2006). Analyzing incomplete discrete longitudinal clinical trial data. *Statistical Science, 21*, 52-69.

Jennrich, R.I., and Schluchter, M.D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics, 42*, 805-820.

Joffe, M.M., Small, D., and Hsu, C.Y. (2007). Defining and estimating intervention effects for groups that will develop an auxiliary outcome. *Statistical Science, 22*, 74-97.

Kang, J.D.Y., and Schaffer, J.L. (2007). A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science, 22*(4), 523-539.

Kenward, M.G. (1998) Selection models for repeated measurements with non-random dropout: An illustration of sensitivity. *Statistics in Medicine, 17*(23), 2,723-2,732.

Kenward, M.G., and Carpenter, J.R. (2008). Multiple imputation. In *Longitudinal Data Analysis*, G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs (eds.). New York: CRC Press.

Kenward, M.G., and Molenberghs, G. (2009). Last observation carried forward: A crystal ball? *Journal of Biopharmaceutical Statistics, 9*(5), 872-888.

Kenward, M.G., Lesaffre, E., and Molenberghs, G. (1994). An application of maximum likelihood and estimating equations to the analysis of ordinal data from a longitudinal study with cases missing at random. *Biometrics, 50*, 945-953.

Kenward, M.G., Molenberghs, G., and Thijs, H. (2003). Pattern mixture models with proper time dependence. *Biometrika, 90*(1), 53-71.

Laird, N.M., and Ware, J.H. (1982). Random effects models for longitudinal data. *Biometrics, 38*, 963-974.

Lakatos, E. (1988). Sample sizes based on the logrank statistic in complex clinical trials. *Biometrics, 44*, 229-241.

Lavori, P.W., Brown, C.H., Duan, N., Gibbons, R.D., and Greenhouse, J. (2008). Missing data in longitudinal clinical trials part A: Design and conceptual issues. *Psychiatric Annals, 38*(12), 784-792.

Leon, A.C., Hakan, D., and Hedeken, D. (2007). Bias reduction with an adjustment for participants' intent to dropout of a randomized controlled clinical trial. *Clinical Trials, 4*, 540-547.

Liang, K-Y., and Zeger, S.L. (1986). **Longitudinal data analysis using generalized linear** models. *Biometrika, 73*, 13-22.

Liang, K-Y., Zeger, S.L., and Qaqish, B. (1992). Multivariate regression analyses for categorical data (with discussion). *Journal of the Royal Statistical Society, Series B, 54*, 3-40.

Little, R.J.A. (1985). A note about models for selectivity bias. *Econometrica, 53*, 1,469-1,474.

Little, R.J.A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association, 88*, 125-134.

Little, R.J.A. (1994). A class of pattern-mixture models for normal missing data. *Biometrika, 81*, 471-483.

Little, R.J.A. (1995). Modeling the drop-out mechanism in longitudinal studies. *Journal of the American Statistical Association, 90*, 1,112-1,121.

Little, R.J.A. (2008). Selection and pattern mixture models. In *Advances in Longitudinal Data Analysis*, G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs (eds.), pp. 409-431. London, England: CRC Press.

Little, R.J.A., and Rubin, D.B. (2000). Causal effects in clinical and epidemiological. studies via potential outcomes: Concepts and analytical approaches. *Annual Review of Public Health, 21*, 121-145.

Little, R.J.A., and Rubin, D.B. (2002). *Statistical Analysis with Missing Data, 2nd edition*. New York: Wiley.

Little, R.J.A., and Wang, Y.-X. (1996). Pattern mixture models for multivariate incomplete data with covariates. *Biometrics, 52*, 98-111.

Little, R.J.A., and Yau, L. (1998). Statistical techniques for analyzing data from prevention trials: Treatment of no-shows using Rubin's causal model. *Psychological Methods, 3*, 147-159.

Little, R.J.A., Long, Q., and Lin, X. (2009). A comparison of methods for estimating the causal effect of a treatment in randomized clinical trials subject to noncompliance. *Biometrics, 65*(2), 640-649.

Liu, M., Taylor, J.M.G., and Belin, T.R. (2000). Multiple imputation and posterior simulation for multivariate missing data in longitudinal studies. *Biometrics, 56*, 1,157-1,163.

Long, Q., Little, R.J.A., and Lin, X. (in press). Estimating the CACE in trials involving multi-treatment arms subject to noncompliance: A Bayesian framework. To appear in the *Journal of the Royal Statistical Society, Series C.*

Marcus, B.H., Lewis, B.A., Hogan, J., King, T.K., Albrecht, A.E., Bock, B., Parisi, A.F., Niaura, R., and Abrams, D.B. (2005). The efficacy of moderate-intensity exercise as an aid for smoking cessation in women: A randomized controlled trial. *Nicotine and Tobacco Research, 7*(6), 871-880.

Mehrotra, D., Li, X., and Gilbert, P.B. (2006). A comparison of eight methods for the dual endpoint evaluation of efficacy in a proof of concept HIV vaccine. *Biometrics, 62*, 893-900.

Meng, X.L., and van Dyk, D. (1997). The EM algorithm—An old folk song sung to a fast new tune. *Journal of the Royal Statistical Society, Series B, 59*, 511-567.

Molenberghs, G., and Kenward, M.G. (2007). *Missing Data in Clinical Studies*. Chichester, UK: Wiley.

Molenberghs, G., Kenward, M.G., and Lesaffre, E. (1997). The analysis of longitudinal ordinal data with informative dropout. *Biometrika, 84*(1), 33-44.

Molenberghs, G., Michiels, B., Kenward, M.G., and Diggle, P.J. (1998). Monotone missing data and pattern mixture models. *Statistica Neerlandica, 52*(2), 153-161.

Molenberghs, G., Verbeke, G., Thijs, H., Lesaffre, E., and Kenward, M.G. (2001). Influence analysis to assess sensitivity of the dropout process. *Computational Statistics and Data Analysis, 37*(1), 93-113.

Mori, M., Woolson, R.F., and Woodsworth, G.G. (1994). Slope estimation in the presence of informative censoring: Modeling the number of observations as a geometric random variable. *Biometrics, 50*, 39-50.

Murray, G.D., and Findlay, J.G. (1988). Correcting for the bias caused by dropouts in hypertension trials. *Statistics in Medicine, 7*, 941-946.

Nathan, D.G., and Wilson, J.D. (2003). Clinical research and the NIH: A report card. *New England Journal of Medicine, 349*, 1,860-1,865.

Normand, S.L., Rector, T.S., Neaton, J.D., Pina, I.L., Lazar, R.M., Proestel, S.E., Fleischer, D.J., Cohn, J.N., and Spertus, J.A. (2005). Clinical and analytical considerations in the study of health status in device trials for heart failure. *Journal of Cardiac Failure, 11*, 396-403.

Office for Human Research Protections. (2008). *Draft Guidance on Important Considerations for When Participation of Human Subjects in Research Is Discontinued*. Department of Health and Human Services. Available: http://www.primr.org/uploadedFiles/PRIMR_Site_Home/Public_Policy/Recently_Files_Comments/Draft_Guidance_on_Discontinued_Participation.pdf.

Oleske, D.M., Kwasny, M.M., Lavender, S.A., and Andersson, G.B. (2007). Participation in occupational health longitudinal studies: Predictors of missed visits and dropouts. *Annals of Epidemiology, 17*, 9-18.

O'Neill, R.T. (2009). *Missing Data in Clinical Trials Intended to Support Efficacy and Safety of Medical Products: The Need for Consensus*. 30th Annual Conference of the International Society for Clinical Biostatistics, Prague, Czech Republic. Available: http://www.iscb2009.info/RSystem/Soubory/Prez%20Tuesday/S18.3%20O'Neill.pdf.

Park, T. (1993). A comparison of the generalized estimating equation approach with the maximum likelihood approach for repeated measurements. *Statistics in Medicine, 12*, 1,723-1,732.

Pocock, S.J. (1983). *Clinical Trials: A Practical Approach*. New York: Wiley.

Qin, L., Gilbert, P.B., Follmann, D, and Li, D. (2008). Assessing surrogate endpoints in vaccine trials with case-cohort sampling and the Cox model. *Annals of Applied Statistics, 2*,386-2,407.

Raghunathan, T., Lepkowski, J. VanHoewyk, M., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology, 27*(1), 85-95.

Robins, J.M. (1989). The analysis of randomized and non-randomized treatment trials using a new approach to causal inference in longitudinal studies. In *Health Service Research Methodology: A Focus on AIDS,* L. Sechrest, H. Freeman, and A. Mulley (eds.), pp. 113-159. Washington, DC: U.S. Public Health Service, National Center for Health Services Research.

Robins, J.M. (1993). Analytic methods for estimating HIV treatment and cofactor effects. In *Methodological Issues of AIDS Mental Health Research,* D.G. Ostrow and R. Kessler (eds.), pp. 213-290. New York: Plenum.

Robins, J.M., and Finkelstein, D (2000). Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics, 56*(3), 779-788.

Robins, J.M., and Greenland, S. (1996). Discussion of identification of causal effects using instrumental variables. *Journal of the American Statistical Association, 91*, 456-458.

Robins, J.M., and Rotnitsky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology-Methodological Issues,* N. Jewell and D.K. Farewell (eds.), pp. 297-331. Boston, MA: Birkhäuser.

Robins, J.M., and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association, 90*, 122-129.

Robins J.M., and Rotnitzky, A. (2001). Comment on Inference for semiparametric models: Some questions and an answer by Bickel and Kwon. *Statistica Sinica, 11*, 920-936.

Robins, J.M., Rotnitzky, A., and Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association, 89*, 846-866.

Robins, J.M., Rotnitzky, A., and Zhao, L.P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association, 90*, 106-121.

Robinson, K.A., Dennison, C.R., Wayman, D.M., Pronovost, P.J., and Needham, D.M. (2007). Systematic review identifies a number of strategies important for retaining study participants. *Journal of Clinical Epidemiology, 60*(8), 757-765.

Rose, E.A., Gelijns, A.C., Moskowitz, A.J., Heitjan, D.F., Stevenson, L.W., Debistky, W., Long, J.W., Ascheim, D., Tierney, A.R., Levitan, R.G., Watson, J.T., and Meier, P. (2001). Long-term use of a left ventricular assist device for end-stage heart failure. *New England Journal of Medicine, 345*, 1,435-1,443.

Rosenbaum, P.R., and Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41-55.

Rotnitzky, A., Holcroft, C.A., and Robins, J.M. (1997). Efficiency comparisons in multivariate multiple regression with missing outcomes. *Journal of Multivariate Analysis, 61*(1), 102-128.

Rotnitzky, A., Robins, J.M., and Scharfstein, D.O. (1998). Semiparametric regression for repeated measures outcomes with nonignorable nonresponse. *Journal of the American Statistical Association, 93*, 1,321-1,339.

Rotnitzky, A., Farall, A., Bergesion, A., and Scharfstein, D. (2007). Analysis of failure time data under competing censoring mechanisms. *Journal of the Royal Statistical Society, Series B, 69*, 307-327.

Rotnitzky, A., Bergesio, A., and Farall, A. (2009). Analysis of quality of life adjusted failure time data in the presence of competing possibly informative censoring mechanisms. *Lifetime Data Analysis, 15*, 1-23.

Rubin, D.B. (1976). Inference and missing data. *Biometrika, 63*, 581-592.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.

Rubin, D.B. (1996). Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association, 91*, 473-489.

SAS Institute, Inc. (2008a). *SAS/STAT®9.2, User's Guide, The MIXED Procedure*. Cary, NC: Author.

SAS Institute, Inc. (2008b). *SAS/STAT®9.2, User's Guide, The NLMIXED Procedure*. Cary, NC: Author.

Schafer, J.L. (1996). *Analysis of Incomplete Multivariate Data*. London, England: Chapman and Hall.

Scharfstein, D.O., and Robins, J.M. (2002). Estimation of the failure time distribution in the presence of informative censoring. *Biometrika, 89*, 617-634.

Scharfstein, D.O., Rotnitzky, A., and Robins, J.M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion). *Journal of the American Statistical Association, 94*, 1,096-1,146.

Schluchter, M.D. (1992). Methods for the analysis of informatively censored longitudinal data. *Statistics in Medicine, 11*, 1,861-1,870.

Schneider, M.M., Hoepelman, A., Schattenkerk, **D.K., Eeftinck, T.L., Nielsen, van def Graaf,** Y., Frissen, J.P., van def Ende, I.M., Kolsters, A.F., and Borleffs, J.C. (1992). **A controlled** trial of aerosolized pentamidine or trimethoprim-sulfamethoxazole as primary prophylaxis against Pneumocystis carinii pneumonia in patients with human immunodeficiency virus infection. *New England Journal of Medicine, 327*(26), 1,836-1,841.

Shepherd, B.E., Redman, M.W., and Ankerst, D.P. (2008). Does Finasteride affect the severity of prostate cancer? A causal sensitivity analysis. *Journal of the American Statistical Association, 103*, 1,392-1,404.

Shih, J.H. (1995). Sample size calculation for complex clinical trials with survival endpoints. *Controlled Clinical Trials, 16*(6), 395-407

Slaughter M.S., Rogers, J.G., and Milano, C.A., (2009). Advanced heart failure treated with continuous-flow left ventricular assist device. *New England Journal of Medicine, 361*, 2,241-2,251.

Snow, W.M., Connett, J.E., Sharma, S., and Murray, R.P. (2007), Predictors of attendance and dropout at the Lung Health Study 11-year follow-up. *Contemporary Clinical Trials, 28*(1), 25-32.

Sprague, S., Leece, P., Bhandari, M., Tornetta, P., 3rd, Schemitsch, E., and Swiontkowksi, M.F. (2003). Limiting loss to follow-up in a multicenter randomized trial in orthopedic surgery. *Controlled Clinical Trials, 24*(6), 719-725.

Stolzenberg, R.M., and Relles, D.A. (1990). Theory testing in a world of constrained research design—The significance of Heckman's censored sampling bias correction for nonexperimental research. *Sociological Methods and Research, 18*, 395-415.

Tanner, M.A. (1991). *Tools for Statistical Inference: Observed Data and Data Augmentation Methods*. New York: Springer-Verlag.

Tanner, M.A., and Wong, W.H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association, 82*, 528-550.

Thijs, H., Molenberghs, G., and Verbeke, G. (2000). The milk protein trial: Influence analysis of the dropout process. *Biometrical Journal, 42*(5), 617-646.

U.S. Food and Drug Administration. (2002). *Guidance for Industry: Antiretroviral Drugs Using Plasma HIV; RNA Measurements—Clinical Considerations for Accelerated and Traditional Approval.* Available: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM070968.pdf.

U.S. Food and Drug Administration. (2008). *Guidance for Sponsors, Clinical Investigators, and IRBs: Data Retention When Subjects Withdraw from FDA-Regulated Clinical Trials*. Office of the Commissioner, Good Clinical Practice Program. Available: http://www.fda.gov/downloads/RegulatoryInformation/Guidances/UCM126489.pdf.

Van Buuren, S., and Oudshoorn, C.G.M. (1999). *Flexible Multivariate Imputation by MICE.* (TNO/VGZ/PG 99.054.) Leiden, The Netherlands: TNO Preventie en Gezondheid. Available: http://www.multiple-imputation.com.

Van Steen, K., Molenberghs, G., Verbeke, G., and Thijs, H. (2001). A local influence approach to sensitivity analysis of incomplete longitudinal ordinal data. *Statistical Modeling, 1*, 125-142.

Verbeke, G., and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer-Verlag.

Verbeke, G., and Molenberghs, G. (in press). Arbitrariness of models for augmented and coarse data, with emphasis on incomplete-data and random-effects models. *Statistical Modeling, 9*.

Verbeke, G., Molenberghs, G., Thijs, H., Lesaffre, E., and Kenward, M.G. (2001), Sensitivity analysis for nonrandom dropout: A local influence approach. *Biometrics, 57*(1), 7-14.

Wang-Clow, F., Lange, M., Laird, N.M., and Ware, J.H. (1995). Simulation study of estimators for rate of change in longitudinal studies with attrition. *Statistics in Medicine, 14*, 283-297.

Warden, D., Trivedi, M.H., Wisniewski, S.R., Davis, L., Nierenberg, A.A., Gaynes, B.N., Zisook, S., Hollon, S.D., Balasubramani, G.K., Howland, R., Fava, M., Steward, J.W., and Rush, A.J. (2007). Predictors of attrition during initial (citalopram) treatment for depression: A STAR*D report. *American Journal of Psychiatry, 164*, 1,189-1,197.

White, I.R. (2005). Uses and limitations of randomization-based efficacy estimators. *Statistical Methods in Medical Research, 14*, 327-347.

Williams, P.L., Van Dyke, R., Eagle, M., Smith, D., Vincent, C., Cuipak, G, Oleske, J., and Seage, G.R. III. (2008). Association of site-specific and participant-specific factors with retention of children in a long-term pediatric HIV cohort. *American Journal of Epidemiology, 167*, 1,375-1,386.

Wu, M.C., and Bailey, K.R. (1989). Estimation and comparison of changes in the presence of informative right censoring: conditional linear model. *Biometrics, 45*, 939-955.

Wu, M.C., and Carroll, R.J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics, 44*, 175-188.

Zeger, S.L., and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics, 42*(1), 121-130.

Zhang, G., and Little, R.J.A. (2009). Extensions of the penalized spline of propensity prediction method of imputation. *Biometrics, 65*, 911-918.

# Appendix A

# Clinical Trials: Overview and Terminology

Prior to the adoption of a new treatment for use in a population, it is important to assess the impact that the use of the treatment will have on the general health of the population. That is, one wants to know how the general health of the population after adoption of the treatment compares with what it would have been if the treatment had not been adopted. In practice, this can never be known exactly (since it is a counterfactual). But the governmental agencies that regulate approval of new treatments are charged with judging the treatment's impact to the extent possible. This appendix presents an overview of the purposes and various aspects of clinical trials and definitions of some of the key terms used in our study.

## BASIC TERMS

An *effective* treatment is one that provides improvement in the general health of the population viewed as a whole. An *efficacious* treatment is one that in some identifiable subpopulation results in an outcome judged more beneficial than that which would exist without treatment. An efficacious treatment may not be effective owing either to its inability to be administered safely in a broad population or to its effect on other aspects of patient treatment and behaviors beyond the outcome used to evaluate efficacy.

It is also useful to differentiate among the concepts of a simple *treatment*, which would usually consist of a prescribed dose of given frequency and duration; a *treatment regimen,* which would usually involve rules for dose escalation or reduction in order to obtain greater effect while avoiding intolerable adverse experiences; and a *treatment strategy*, which would

include plans for auxiliary treatments and progression to other treatments in the face of disease progression.

In a phase III confirmatory study (see below), the ideal is typically an effectiveness study of a treatment strategy: effectiveness because it is the impact of a treatment on the population and a treatment strategy because the initial prescribed treatment may greatly affect the concomitant treatments and follow-on treatments administered to patients. However, true effectiveness can never be tested in an unbiased fashion because the trial setting itself is artificial and because observational studies are always subject to unmeasured bias. Phase III studies should be much closer to an effectiveness study than would be the phase II studies that might use surrogate biomarkers as a primary outcome in a subpopulation of the patients that might ultimately receive an approved treatment.

Whether the primary goal of a clinical trial is effectiveness or efficacy, the scientific validity of the comparison of the new treatment to some standard depends on the comparability of the groups that receive the experimental and control treatments. Randomization of patients to two or more treatment groups is the primary tool to ensure the comparability of samples, at least on average. Hence, it is of utmost importance that the data from each clinical trial be analyzed consistent with the *intent-to-treat* principle, which dictates that each subject's data be included in the treatment group to which he or she is randomized. This approach is clearly in keeping with an evaluation of the effectiveness of a treatment strategy, but even when evaluation of efficacy is the goal, the clinical trial should ideally be designed in such a way that all randomized patients will contribute to the estimate of treatment efficacy. However, in limited situations, it might be judged acceptable to evaluate efficacy in a *modified intent-to-treat* subgroup of randomized patients defined on the basis of measurements made prior to randomization and ascertained in an unbiased fashion for each treatment group. In this setting, safety would still be evaluated in patients who are not in the subgroup.

In neither effectiveness nor efficacy studies would an analysis based on a *compliant* or *per-protocol* analysis population (defined as patients who adhered strictly to the prescribed dose, frequency, and duration of the assigned treatments) be considered a scientifically rigorous assessment of the treatment. Instead, when the efficacy of the treatment in a compliant population is of interest, one needs to find a way to randomize only those patients who can tolerate the treatment and who will adhere to the protocol (see below).

## GOAL: INDICATION FOR A NEW TREATMENT

Ultimately, a new treatment is characterized by its "indication." An ideal *treatment indication* will consist of a disease, a patient population, an intervention, and an outcome.

***The Disease*** The exact medical definition of "disease" can range from primarily signs and symptoms (e.g., headache, pneumonia) to presumed causative agents (e.g., pneumococcal pneumonia, gram negative pneumonia, fungal pneumonia, or carcinomatous pneumonia). The definition of a disease is frequently not refined further than necessary to decide on an appropriate treatment strategy. In this way, the identification of a beneficial treatment often becomes the definition of the disease (e.g., all gram negative pneumonias are considered together due to the common treatment chosen in those settings). Other times, the lack of efficacy of the usual treatment is incorporated into the definition of the disease (multidrug resistant *Staph aureus*). It is common that particular diseases are diagnosed through a series of tests and procedures. The nomenclature for the disease may include the method of diagnosis (e.g., culture positive gram negative septicemia). However, it is rare that any sign or symptom be "pathognomonic"—uniquely identifying—for the disease.

***The Population of Patients*** Because of concurrent medical conditions, a treatment might be indicated only for a subpopulation of patients who satisfy the diagnostic criteria for the disease. There might be known safe and effective therapies that are regarded as the first-line treatment of the disease. In such a case, an indication for a new treatment might indicate the treatment's use only in patients for whom the standard therapy is *a priori* judged inadvisable due to concurrent medical conditions (e.g., pediatrics, pregnancy, poor renal function in a drug cleared by the kidneys) or who cannot take the standard therapy (e.g., due to lack of tolerance with respect to side effects or lack of efficacy).

***The Intervention*** An intervention consists of a formulation of the drug(s) or device(s), a mode of administration, the dosing strategy, auxiliary treatments, and the duration of treatment. Some treatments are combinations of drugs, either in a common formulation or administered separately. A mode of administration can include topical, oral, subcutaneous, intramuscular, or intravenous. In some circumstances the mode of administration may even stipulate special training for the person administering the treatment. The dose may be specified as a common level to be used by all individuals or as a dose specific to patient body size or body surface area. The dosing strategy might include a gradual increase in dose as treatment is initiated, a tapering

of dose as the patient is weaned from the therapy, a regimen for increasing or decreasing the dose in response to observed patient conditions at the time of dosing (e.g., serum glucose in insulin therapy) or observed patient response to therapy (e.g., increasing the dose if the effect is not optimal or decreasing the dose in the presence of treatment toxicity).

Auxiliary treatments might be administered as prophylaxis against known toxic effects (e.g., G-CSF with cancer treatments, antihistamines with drugs that tend to trigger an immune response) or for rescue from toxic effects (e.g., methotrexate followed by leucovorin rescue). One also has to indicate the frequency of administering the treatment. Finally, there is the duration of treatment, which might include "drug holidays."

***The Desired Outcome*** The intended outcome of a treatment is typically characterized clinically, as outcomes that materially affect the clinical manifestations of the disease (e.g., lower risk of mortality, relief of symptoms, improvement in quality of life). In some settings, a strong risk factor thought to represent a surrogate outcome measure of subclinical disease or disease risk will be used (e.g., hypertension). The distinction between surrogate and clinical outcomes depends on the degree to which a patient's sense of well-being is directly related to the outcome or the degree to which it is known that any modifications in the biomarker might possibly not be associated with an improvement in the clinical outcome (i.e., treating the symptom but not the disease). The precise definition of the outcome might explicitly include the time frame of measurement (e.g., postprandial serum glucose levels) and the method of measurement (e.g., decreasing serum glucose levels as reflected in Hemoglobin A1c), or the time frame might only be implicitly defined.

## Treatment Discovery Process

The treatment discovery process is an iterative process of studying a disease, hypothesizing and developing treatments, evaluating those treatments, and, for successful treatments, further refining the indication to account for lack of efficacy or toxicities (or both) in particular subgroups of patients. As a rule, the scientific development of a particular treatment indication is often connected with that of other treatments, and thus it may be difficult to identify the exact process that led to the adoption of some treatment. Nevertheless, the following describes a general chronology of events.

Initially, some targeted disease is characterized from observational studies (including epidemiologic studies of risk factors for the disease), clinical observation of typical disease progression and predictors of outcomes, and laboratory studies of biochemical and histologic changes in the diseased

patients' tissues. Often, this characterization of a disease starts with a constellation of symptoms and signs, and much of the ensuing observational research is directed toward finding a causative agent. Observational studies of diseased patients are then often augmented by laboratory experiments that try to further illuminate the causation of the disease and the cellular and physiologic mechanism that cause its major complications. These experiments might involve *in vitro* studies of cell lines and animal models of the disease.

Based on the understanding of the disease gained from the above types of studies, scientists might propose a potential treatment or preventive strategy. The proposed treatment is then evaluated and further refined in a series of preclinical laboratory and animal experiments. Such experiments might focus on two general approaches: *in vitro* characterization of the chemical and biochemical interactions of new drugs with cellular and extracellular constituents of the human body, and *in vitro* characterization of the effects of the new therapies on cellular mechanisms using cell lines or animal experiments in suitable species. The goal of this work is to characterize:

- pharmacokinetics, measuring the effect of dose on rates of absorption and excretion of drugs from various body compartments;
- pharmacodynamics, measuring the intended or unintended effects of dose on physiologic measures;
- toxicology, measuring the effect of dose on histopathologic lesions in major organ systems;
- reproductive and embryologic effects as a function of dose; and
- *in vivo* drug-drug interactions that might lead to attenuation or potentiation of intended or unintended effects of the treatment or that might affect the pharmacokinetics of the drugs.

When sufficient preclinical studies have been performed to conclude that the treatment is basically safe, work moves to experiments in human volunteers. In order to sequentially investigate safety and then efficacy and effectiveness issues in a manner that protects human subjects from harm, the process of investigating new treatments typically goes through a phased series of clinical trials. The considerations during each phase depend on whether the investigational product is targeting disease prevention, diagnosis, or treatment, as well as the severity of disease, the type of intervention (e.g., drug, biologic agent, device, or behavior), and the prior knowledge of treatment risks. The following thus describes only the general principles behind the phased investigation.

## Clinical Trial Phases

*Phase I* clinical trials provide initial safety data to support further testing with larger samples. As the focus of these studies is primarily safety of investigation rather than efficacy or effectiveness of treatment, the study subjects are frequently a small number of healthy volunteers. A notable exception occurs when a treatment that is designed to be administered in life-threatening disease is known to have severe toxicity. For instance, in phase I cancer clinical trials, the treatment might be first tested in patients whose disease has proven resistant to all other therapies.

"First in human" clinical trials might start with a single administration of the treatment at an extremely low dose in a few subjects. In the absence of unacceptable toxicity, subsequent patients might receive increasing doses. Owing to a desire to slowly increase exposure to the treatment, patients may not be randomized across all doses. In cancer chemotherapy trials, in particular, there may be no control group. Pharmacokinetic data might be gathered on both single doses and repeated dosing to assess the rates of absorption and elimination in humans. These kinds of studies might also consider pharmacokinetics in the presence of renal or hepatic impairment, as well as pharmacokinetics in the presence of meals and other drugs. When phase I trial results show unexpected severe toxicities, further consideration of the treatment might be curtailed.

*Phase II* clinical trials seek further safety data and preliminary evidence in support of biological effect. A slightly larger sample of subjects are administered the treatment at a dose or doses that were preliminarily judged safe in the phase I studies. Safety data are collected in a systematic fashion, including specified monitoring of any potential side effects that were identified previously. Phase II studies also serve to screen for treatments that show some sign of biological effect, such as a biological marker that is a surrogate for the clinical outcome that is of interest. Products that fail to demonstrate a certain level of biological activity might be abandoned. Such a screening process is more efficient than other approaches in finding effective treatments from a large population of ideas.

Even when the phase II clinical trials demonstrate a desired effect on the biologic outcome, it is common for investigators to use the results of the clinical trial to identify more specific factors:

- a more precise definition of the disease characteristics that would indicate the types of patients likely to benefit most from the treatment,
- a more refined definition of the population to be treated in order to eliminate subjects who might experience greater toxicity,
- a single treatment regimen (dose or dosing strategy, frequency, duration, auxiliary prophylactic, or rescue therapies), or

- a clinical measure to serve as the primary outcome, as well as a statistical measure to summarize the distribution of that clinical outcome across subjects.

The selection of this primary outcome (and summary measure) is based on consideration of (in order of importance): (1) the clinical measure that is most indicative of an improved clinical outcome for the patient, (2) a measure that the treatment might plausibly affect, and (3) an outcome that can be compared across treatment groups with good statistical precision.

*Phase III* clinical trials, which are the main focus of the panel's report, are large confirmatory studies meant to establish an acceptable benefit/safety profile in order to gain regulatory approval for a precisely defined indication ("registrational" clinical trials). Phase III trials are well-controlled trials that provide scientifically credible and statistically strong evidence about the treatment indication hypothesized at the end of phase II investigation.

In order for a phase III trial to be regarded as confirmatory, it is crucial that the hypotheses being tested be specified before the start of the trial. Sample sizes are typically chosen to have a high probability of ruling out the possibility of ineffective therapies and to estimate the treatment effect with high statistical precision. Collection of safety data continues to play a major role, as the larger sample sizes in the phase III study afford a better opportunity to identify relatively rare serious toxicities. As a general rule, the approval process does not demand statistically proven increased rates of toxicities prior to providing warnings to patients and physicians. Depending on the disease and patient population, anecdotal occurrence of unexpected extremely serious adverse events will often dictate further study of a proposed treatment.

Evidence from phase III studies that strongly support the proposed indication will generally lead to adoption of the therapy. Sometimes, however, even when a proposed treatment has "met its outcome" in the overall study population, the indication (treatment) actually adopted might be more restrictive than was initially proposed due to lessened efficacy or heightened toxicity observed in a subgroup of patients.

Suppose, for example, that there are two subpopulations, A and B, and that the proposed therapy "met its outcome" in the combined sample. But suppose that when analyzed alone, subpopulation B did not appear to have an acceptable benefit/risk ratio (which indicates that subpopulation A exhibited a strong benefit of the treatment). Because it is not uncommon for proposed treatments to present safety issues, more focus is often placed on making sure that harmful treatments do not get adopted. In this example, subpopulation A might be approved to receive the new treatment, while regulators require additional data in support of the benefits of the treatment for subpopulation B.

There are two potential drawbacks to this "data-driven" restriction of indication. One is that if the observed difference in treatment benefit/risk is spurious, subpopulation B is deprived of a useful therapy until additional data is gathered. The other is that if the observed difference in treatment benefit/risk is spurious, the commercial sponsor will have lost income from sales to B as well as having the added expense of further studies in that subgroup.

*Phase IV* clinical trials are postmarketing trials that are meant to evaluate rare but serious effects that cannot be assessed in the smaller Phase III studies.

The above description is most applicable to the evaluation of new therapies. In disease prevention, some authors have suggested that phase III trials focus on efficacy by demonstrating the treatment benefit of prevention through some surrogate biomarker of the disease (e.g., colon polyps as a precursor lesion for colon cancer) and phase IV trials focus on effectiveness, using the clinically relevant outcome in a population-based sample of the types of individuals likely to receive an adopted treatment. In doing so, these studies may also aim to evaluate changes in individual behavior that might mitigate the efficacy of the treatment (e.g., increased risk-taking behavior when vaccinated for HIV or treated for peanut allergies).

## GOAL OF EFFECTIVENESS VERSUS EFFICACY

Evidence-based medicine often involves a stepwise process that closely parallels the parts of a treatment indication described above. These steps consider (1) patient population (the definition of the disease and any restrictions on patient characteristics apart from disease manifestations), (2) intervention, (3) comparison (alternatives to the intervention that might be considered), and (4) outcome (the clinical condition that is desired), and they are referred to as PICO. Distinctions between effectiveness and efficacy are given below in the context of PICO.

### Effectiveness: Phase III Trials

An effective treatment is one that provides improvement in the general health of the population viewed as a whole: "general health" in some sense considers the average state of health of the population. An "effectiveness trial" enrolls a representative sample from the population of patients who would eventually receive the treatment. The effectiveness study should strive for an inclusive setting, which might include both independently living as well as institutionalized patients, and it should, insofar as safety permits, not restrict patients based on concomitant disease unless such a restriction will be in the ultimate indication.

Ideally, the eligibility criteria would consist of inclusion criteria that define the population of patients that would ultimately be in the indication for the treatment, the criteria would also delineate patients who might be inappropriate for a randomized controlled trial evaluating an unproven therapy. The intervention would then be administered as it will be given to a patient, which might include (a) decreased dosage due to lack of tolerability, (b) lack of compliance on the part of the patient, (c) auxiliary treatments used to prevent or treat unintended effects of the treatment, and (d) other changes of patient or treating clinician behavior.

The trial would then compare the new intervention to the treatment the population of patients would otherwise (in the absence of the new intervention) have received and evaluate an outcome that is the best summary of "general health" for the patient population, which might be affected by (a) other changes in behavior that are associated with receiving the treatment, and (b) the timing of the intervention and the timing and methods of measurement for the outcome.

### Efficacy: Phase II Trials

Treatment efficacy can be defined in a subset of the patients who would eventually be treated, and it can be based on an outcome that is merely thought to be an indicator of eventual clinical benefit. An "efficacy trial" might enroll patients from a defined subset of diseased patients who are most likely to show evidence of treatment effect. This could be because (a) they have been previously (prior to randomization) found to be able to tolerate the treatment (e.g., during a screening "run-in" phase), (b) they have been previously (prior to randomization) found to be compliant with randomized controlled trial procedures (e.g., during a screening "run-in" phase), or (c) there is reason to believe prior to randomization that they are relatively more likely to have a beneficial treatment effect than the population of all patients with the disease. A key point is that all randomized patients will be analyzed for their outcomes. Any "enrichment" of the sample to maximize response needs to be done prior to randomization.

The intervention must be clearly defined and may differ from the eventual ("effectiveness") intervention for several reasons: (a) the care providers administering the treatment are more highly trained, (b) the treatment protocol is more rigidly enforced, (c) inducements for high compliance are used, or (d) auxiliary treatments and additional treatments in the presence of lack of efficacy are restricted or proscribed. In addition, the care that the comparison groups receive might not be the standard of care the patient would have received in the absence of the randomized controlled trial. Moreover, the primary outcome may not be the clinical outcome of greatest public health relevance because it may be measured using techniques or

schedules that do not coincide with usual clinical practice (e.g., heightened radiographic surveillance for subclinical progression of cancer rather than examinations based on clinical events), or it may be an intermediate marker that is believed, but not known, to be a necessary and sufficient indicator of the true clinical outcome (e.g., tumor response in a cancer clinical trial, arrhythmias in studies of survival following myocardial infarction).

### Notes on Efficacy Versus Effectiveness

True effectiveness can never be tested in an unbiased fashion because the randomized controlled trial setting itself is artificial and because observational studies are always subject to unmeasured bias. Nonetheless, it is important that in phase III trials, the effectiveness of a therapy be assessed as accurately and precisely as possible.

An efficacious treatment may not be effective for at least four major reasons. First, the kind of patients who were not represented in the efficacy trial have worse clinical outcomes that overwhelm any benefit seen in the efficacy trial sample. This may happen because (a) they have a heightened susceptibility to serious adverse events leading to poor clinical outcomes; (b) they cannot tolerate the treatment, and the therapeutic window for administering alternative therapies has passed; (c) the broader population of patients includes individuals whose disease is so mild or so severe that the intervention provides no benefit, but those patients do experience toxicities; or (d) off-label use of the therapy confers risk but no benefit at a level that outweighs the benefit in the more restrictive population in the indication.

Second, the intervention tested in the efficacy trial differs from the intervention that would be realized in the more inclusive population of patients with the disease or condition, because (a) the skill of the investigators administering the intervention is necessary for the treatment benefit, but that skill is not present in the general setting; (b) the efficacy trial restricted use of auxiliary treatments that interact negatively with the experimental treatment; (c) the efficacy trial restricted use of auxiliary treatments that are in wide use in the population and provide the same benefit as the treatment (perhaps with fewer toxicities); or (d) the compliance of patients with the experimental therapy is markedly worse than was achieved in the efficacy trial, and the toxic effects of the therapy are manifested with lesser exposure than the beneficial effects.

Third, the comparison group in the efficacy trial does not encompass the true standard of care that patients would receive in the absence of the experimental treatment, and the experimental treatment does not provide added benefit over that standard of care.

Fourth, the primary outcome used in the efficacy trial is not predictive of the true clinical outcome, because (a) the predictive value of an interme-

diate marker is affected by the treatment (i.e., "treating a symptom, not the disease"); (b) the schedule of outcome assessment in the efficacy trial led to additional beneficial auxiliary treatments that are not realized in standard medical care; or (c) the population of patients changed their behavior (e.g., risk taking) when taking a treatment that is or is thought to be protective.

## DEFINITIONS

### Treatments

As noted above, it is useful to differentiate between the concepts of a simple *treatment,* a *treatment regimen,* and a *treatment strategy.*

- Treatment (sometimes referred to as nominal experimental treatment) includes formulation, administration, dose (fixed, per weight, per body surface area, adaptive), frequency (including drug holidays), and duration.
- Treatment regimen includes nominal experimental treatment as above, prescribed prophylactic treatments to prevent adverse events, dose modifications in the presence of adverse events or demonstrated efficacy, and prescribed auxiliary treatments for known adverse events.
- Treatment strategy includes treatment regimen as above, patient compliance, auxiliary treatments according to the usual standards of care, and rescue treatments for lack of effect following the usual standards of care with prior characterization of potential rescue treatments. Rescue treatment may represent (a) a second-line (less effective) treatment used in failure of primary therapy, (b) a crossover to an established standard of care that is used as the control treatment, (c) a crossover to the experimental treatment, or (d) a progression to a treatment known to be more effective, but avoided for other reasons (e.g., opiates in pain relief). The treatment strategy is what is truly tested when randomized controlled trial data is analyzed.

### Study Design

The following are some common study designs for randomized clinical trials:

- Randomized cohort design: eligible patients are randomized to therapy and followed for outcomes.
- Prerandomization run-in: patients are started on a placebo to measure compliance with treatment and study procedures or are started on an experimental treatment to ensure tolerance.
- Randomized withdrawal: All subjects start on experimental treat-

ment, and proof of efficacy is based on worsened clinical status following randomized withdrawal.

### Time Frame of Measurement

The following are some common time frames for randomized clinical trials:

- Single fixed study time: outcomes are assessed at some fixed time defined postrandomization.
- Interval study time: outcome is averaged over a specified interval of time postrandomization, or outcome is contrasted over a specified interval of time postrandomization.
- Single fixed event time: outcome is assessed at some time defined by a particular event (e.g., childbirth).
- Interval based on event: outcome is assessed over the interval up to a particular event (e.g., time until liver transplant).
- Administratively censored time to event: outcome is time to some defined event, length of follow-up may vary by individual, and censoring occurs only due to time from randomization to data analysis.
- Time to event subject to competing risks: outcome is time to some defined event providing it occurs prior to another (nuisance) event that would preclude ability to measure, length of follow-up may vary by individual, and (scientific relevance depends on whether competing risk is noninformative and whether other processes will alter risk of the (nuisance) event.

### Scientific Outcomes

Two scientific outcomes for randomized clinical trials are common. One is clinical outcomes, which include survival, specific quality-of-life factors (e.g., serious events leading to hospitalization, diminished functioning such as nonfatal myocardial infarctions, resolution of a chief complaint such as headache), and general quality of life. The other is surrogate outcomes, which include modification of risk factors for clinical outcomes (e.g., blood pressure, HbA1c), intermediate subclinical outcomes (e.g., tumor progression), and biomarkers (e.g., PSA).

There are also studies with multiple outcomes. They include

(1) Coprimary outcomes: the treatment must demonstrate effect on each of several outcomes separately, though there are situations in which the individuals do not need to meet each of the outcomes, which would include cases when safety and efficacy are evaluated separately.

(2) An outcome index: an index for each individual is defined as the sum or average of measurements made on several different outcomes.

(3) Composite outcomes, which include: (a) good outcome is defined by an individual's meeting all outcomes, (b) bad outcome is defined by an individual's failing to meet any of the outcomes, and (c) time of bad outcome is defined as the earliest occurrence of any undesirable event.

### Subject Participation

Eight levels of participation for subject participants can be specified.

1. *Screening only*: an individual is considered for inclusion, may have protocol-specified measurements, and there are no protocol-specified interventions or treatments.
2. *Run-in*: individuals receive protocol specified interventions, namely, placebo (for general compliance behavior) or experimental (for tolerance to treatment), and are nonrandomized (for individual specific measures) or randomized (for investigator training). Evaluation of outcomes will not be included in evaluation of efficacy or effectiveness, but if an individual receives an experimental treatment, safety outcomes likely will be evaluated.
3. *Enrolled*: patients are included who were ever assigned (by the criterion specified in the protocol) to receive the study intervention, and, in a randomized study, they are subjects who receive a randomization code.
4. *Active participation with treatment*: subjects adhere to some part of nominal treatment or treatment regimen, or subjects adhere to monitoring schedule.
5. *Active follow-up after study treatment discontinuation*: a subject has stopped nominal treatment or treatment regimen but is adhering to full monitoring schedule.
6. *Reduced follow-up after study treatment discontinuation*: a subject has stopped nominal treatment or treatment regimen, as well as most invasive or inconvenient monitoring schedule, and but is still followed for passively observable major clinical outcomes (e.g., survival).
7. *Loss to follow-up*: clinical investigators cannot contact the participant, though participation may resume as "active participation with treatment," "active follow-up after study treatment discontinuation," or "reduced follow-up after study treatment discontinuation" in the event the participant is later found.
8. *Withdrawn consent*: the subject has withdrawn consent for further participation of any kind.

## Analysis Populations

The scientific and statistical validity of a randomized controlled trial depends on the comparability of the treatment groups. That comparability is achieved (on average) at the start of a study by randomizing patients to treatments. All events that occur postrandomization are then, plausibly, the result of the treatment assignment. Several terms are used to describe the analysis populations often discussed in clinical trials:

- *Intent to treat*: covers all patients who were ever enrolled. Patients are included in their assigned treatment group. In a randomized study, this population guarantees comparability of treatment groups, and only this analysis population allows generalizability to specified eligibility criteria.
- *Modified intent to treat*: covers a subsample of enrolled patients for which the comparability of randomized groups and generalization of specified eligibility criteria is valid. Subjects are included in their assigned treatment group regardless of treatment actually received. Exclusion of enrolled patients is based on criteria defined prior to enrollment, though the reporting of the measurements used as the basis might be delayed for logistical reasons. No postrandomization events can be allowed to influence eligibility. The purpose is generally to focus an efficacy (not effectiveness) analysis for a subset of patients for whom the treatment is hypothesized to work best, but logistics precludes identification of that group in real time. As an example, an efficacy trial of a treatment for gram negative sepsis may use only those patients whose blood cultures obtained prior to enrollment are found to be positive for gram negative organisms on the laboratory reading performed 48 hours postspecimen collection.
- *Experimental treatment population (per protocol)*: Covers the subset of enrolled patients who received any amount of the study drug (or other treatment). Patients are included in the assigned treatment group. This group does not include patients who were randomized but never received any treatment. Comparability of treatment groups is compromised in an unblinded study because the reasons for not administering the assigned treatment might be based on investigators' or subjects' biases.
- *Safety population*: covers the patients included for the experimental treatment population, but any patient receiving the experimental intervention is analyzed with the experimental group.

## Types of Clinical Trial Data

The types of data collected in a clinical trial can be characterized by their ultimate use:

- *Prerandomization*: data include determination of eligibility, indication of stage or severity of disease, indication of concomitant risk factors, and indication of important subgroups for specified or exploratory analyses.
- *Postrandomization primary treatment compliance*: data include information on compliance (dose reduction, delay, and termination, protocol-specified adaptations versus patient/provider choice) and on realized treatment, which includes duration of treatment, cumulative dosing, and dose intensity.
- *Postrandomization concomitant or auxiliary treatments*: data include safety and efficacy outcomes, including intermediate measures and surrogate measures, as well as measures of secondary outcomes.

### Mechanisms of Missing Data

There are a variety of ways that data that are intended to be collected in a clinical trial can be missing. A patient can fail to be included in the denominator for which measurement is scientifically relevant for at least two major reasons. One is that the patient was never included for scientific reasons (e.g., pregnancy test in men) or for efficiency reasons. The second is that the patient is no longer included due to end of protocol-specified time frame due to scientific reasons (e.g., death), efficacy or efficiency reasons (e.g., symptom relief in a trial separating efficacy and safety analyses), or ethical reasons (e.g., crossover to a known, more effective rescue therapy).

There is also item nonresponse, which can be due to: (1) clinical inadvisability for specific invasive procedures (e.g., liver biopsy in patient with bleeding disorder), (2) patient refusal for specific invasive procedures or measurements (e.g., refused biopsies), (3) patient refusal to answer specific questions (e.g., sexual behavior, income), or (4) patient's missing clinic visits on time-sensitive measurements.

There is administrative missingness, when the protocol allows study termination prior to complete data collection on each subject, leading to missing repeated measures or censored time to event. There is also missingness from competing risks (e.g., censoring by death from other causes in a cancer clinical trial), missingness due to treatment noncompliance (which is relevant when trying to evaluate a treatment or treatment regimen, rather than a treatment strategy), missingness due to loss to follow-up, missingness due to withdrawal of consent, and missingness due to data editing (values out of range).

**Sensitivity Analysis**

In the body of the report, we focus our discussion of sensitivity analyses on sensitivity to the assumption about the underlying mechanism producing the missing values. There are other aspects of a statistical model for which sensitivity should be assessed. Here is an outline of the steps leading to a comprehensive sensitivity analysis for such models:

(1) *Presumed mechanisms of missing data*: steps would include identification of data likely to be missing, speculation on mechanisms leading to that missing data, and specification of analyses of missing data patterns.

(2) *Planned analyses to deal with missing data*: presumed model assumes either missing completely at random, missing at random, or missing not at random (as defined in Chapter 3); the population with available data that will be used (e.g., complete cases, all available data, etc.); the variables that will be used; how variables will be modeled; distributional assumptions; the statistical model; and the statistical paradigm (Bayesian, frequentist, likelihood).

(3) *Sensitivity analyses*: one will need (a) a framework for exploring effect of distributional assumptions, (b) a framework for exploring effect of variable modeling (e.g., linear, dichotomized, interactions), (c) a framework for exploring effect of considering other variables, (d) a framework for exploring effect of changing population used for modeling, (e) a framework for exploring effect of assumptions of missing at random or missing not at random, and, finally, (f) possible augmented data collection that can shed light on assumptions.

# Appendix B

# Biographical Sketches of Panel Members and Staff

**RODERICK J.A. LITTLE** (*Chair*) is Richard D. Remington collegiate professor of biostatistics in the School of Public Health at the University of Michigan. Previously, he held positions at the World Fertility Survey, as an American Statistical Association (ASA)/National Science Foundation fellow at the U.S. Census Bureau, and in the Department of Biomathematics at the University of California at Los Angeles. His areas of research focus on the analysis of data with missing values in many areas of application, including clinical contexts. He received the ASA's Wilks' Memorial Award in 2004, and he gave the president's invited address at the Joint Statistical Meetings in 2005. He is an elected member of the International Statistical Institute and a fellow of ASA. He received a B.A. with honors in mathematics from Cambridge University and an M.S. in statistics and operational research and a Ph.D. in statistics from the Imperial College of Science and Technology of London University.

**MICHAEL L. COHEN** (*Study Director*) is a senior program officer for the Committee on National Statistics where he directs studies involving statistical methodology, in particular on defense system testing and decennial census methodology. Formerly, he was a mathematical statistician at the Energy Information Administration, an assistant professor in the School of Public Affairs at the University of Maryland, and a visiting lecturer in the Department of Statistics at Princeton University. A fellow of ASA, he has a B.S. in mathematics from the University of Michigan, and M.S. and Ph.D. degrees in statistics from Stanford University.

**RALPH D'AGOSTINO** is chair of the Mathematics and Statistics Department, professor of mathematics/statistics and public health, and director of the Statistics and Consulting Unit and the executive director of the M.A./Ph.D. program in biostatistics, all at Boston University. He has been affiliated with the Framingham Study since 1982, and is coprincipal investigator of the core contract and director of data management and statistical analysis for the study. His major fields of research are clinical trials, epidemiology, prognostics models, longitudinal analysis, multivariate analysis, robustness, and outcomes/effectiveness research. He is a fellow of ASA and the Cardiovascular Epidemiology Council of the American Heart Association. He has twice received the special citation of the commissioner of the U.S. Food and Drug Administration (FDA), and he was named statistician of the year by the Boston Chapter of ASA. He received A.B. and A.M. degrees in mathematics from Boston University and a Ph.D. in mathematical statistics from Harvard University.

**KAY DICKERSIN** is director of the Center for Clinical Trials at the Bloomberg School of Public Health and professor in the Department of Epidemiology, both at Johns Hopkins School of Public Health. Previously, she served as the director of the Center for Clinical Trials and Evidence-based Health Care at Brown University and held faculty positions in the Department of Epidemiology and Preventive Medicine and the Department of Ophthalmology at the University of Maryland School of Medicine. Her areas of research include randomized clinical trials, trials registers, systematic reviews and meta-analysis, publication bias, peer review, and evidence-based health care. She has received a Howard Hughes Fellowship in medical research from Harvard Medical School, and she is an elected member of the Institute of Medicine. She received B.A. and M.A. degrees in zoology from the University of California at Berkeley and a Ph.D. in epidemiology from the School of Hygiene and Public Health of Johns Hopkins University.

**SCOTT S. EMERSON** is professor of biostatistics in the Department of Biostatistics at the University of Washington. Previously, he held faculty positions at the Fred Hutchinson Cancer Research Center and the University of Arizona. His areas of research are clinical trials, sequential testing, survival analysis, categorical data, nonparametric Bayesian statistics, classification and regression trees, statistical consulting, and computer-intensive methods in statistics. He is a fellow of ASA. He received a B.A. in physics, an M.S. in computer science, and an M.D. from the University of Virginia, as well as a Ph.D. in biostatistics from the University of Washington.

**JOHN T. FARRAR** is assistant professor of epidemiology in the Department of Biostatistics and Epidemiology at the University of Pennsylvania School of Medicine. Previously, he held positions at the Children's Hospital of San Francisco, at the New York Hospital of Cornell Medical Center in New York, and in the Department of Neurology of Memorial Sloan Kettering Cancer Center. His areas of research include studies of pain and symptom management. Other interests include pharmaco-epidemiological studies using large databases, functional neuroimaging studies of the neurological manifestations of pain related disease, and novel methodologies in the design and execution of clinical trials. He received a Sc.B. from Brown University, an M.S. in clinical epidemiology from the University of Pennsylvania School of Medicine, an M.D. from the University of Rochester School of Medicine, and a Ph.D. in epidemiology and biostatistics from the University of Pennsylvania School of Medicine.

**CONSTANTINE FRANGAKIS** is associate professor in the Department of Biostatistics in the Bloomberg School of Public Health at Johns Hopkins University. His areas of research include the development of designs and methods of analyses to evaluate treatments in medicine, public health and policy (causal inference), as well as new methods for studies that explore the factors that can be controlled. He is an elected fellow of the Center for Advanced Studies in the Behavioral Sciences, and the recipient of the H.C. Yang Memorial Faculty Award in Cancer Prevention from Johns Hopkins University. He received a B.Sc. in mathematics with statistics from Imperial College of the University of London and his A.M. and Ph.D. degrees in statistics from Harvard University.

**JOSEPH W. HOGAN** is professor in the biostatistics section of the Program in Public Health and a faculty member in the Center for Statistical Sciences, both at Brown University. His research focuses on statistical methods for missing data, causal inference, and sensitivity analysis, including work on informative dropout and noncompliance. Recent topics of investigation include formulation of coherent sensitivity analyses for understanding the effects of missing data assumptions on statistical inferences, use of informative prior distributions to characterize assumptions about missing data mechanisms, and use of flexible models such as regression splines for analyzing incomplete longitudinal data. He is a fellow of ASA. He received a B.A. in mathematics from the University of Connecticut, an M.S. in statistics from the University of Southern California, and a Ph.D. in biostatistics from Harvard University.

**GEERT MOLENBERGHS** is professor of biostatistics at Universiteit Hasselt and Katholieke Universiteit Leuven in Belgium. He was the founding director of the Center for Statistics at Universiteit Hasselt, and he is also the director of the Interuniversity Institute for Biostatistics and statistical bioinformatics. His research interests focus on surrogate markers in clinical trials, and on categorical, longitudinal, and incomplete data. He has served as president of the International Biometric Society, and he received the Guy Medal in Bronze from the Royal Statistical Society and the Myrto Lefkopoulou Award from the Harvard School of Public Health. He received a B.S. degree in mathematics and a Ph.D. in biostatistics (1993) from Universiteit Antwerpen.

**SUSAN A. MURPHY** is H.E. Robbins professor of statistics in the Department of Statistics, research professor at the Institute for Social Research, and professor of psychiatry, all at the University of Michigan. Her primary interest is in causal inference and multistage decisions, sometimes called dynamic treatment regimes or adaptive treatment strategies. Prior to her current position, she held faculty positions in the Department of Statistics at the Pennsylvania State University. She is a fellow of ASA and the Institute of Mathematical Statistics, and she is an invited fellow at the Center for Advanced Study in the Behavioral Sciences. She received a B.S. in mathematics from Louisiana State University and a Ph.D. in statistics from the University of North Carolina at Chapel Hill.

**JAMES D. NEATON** is professor of biostatistics in the Division of Biostatistics in the School of Public Health at the University of Minnesota. His areas of research include the design and conduct of clinical trials and the application of statistical models to the analysis of data arising from intervention studies. He has served as president of the Society for Clinical Trials, as editor in chief of *Controlled Clinical Trials*, and on numerous data and safety monitoring committees. He is a fellow of ASA. He received B.A. and M.S. degrees in biometry from the University of Minnesota and a Ph.D. in biometry from the University of Minnesota.

**ANDREA ROTNITZKY** is a professor in the Departmento de Economia at the Universidad Torcuato Di Tella in Buenos Aires, Argentina. She is also a visiting professor in the Department of Biostatistics at the Harvard School of Public Health, where she previously held faculty positions. Her areas of research include inference with missing data, causal inference from observational studies with time dependent treatment and confounders, analysis of clinical trials with noncompliance, recovery of information from surrogate marker data in clinical trials, analysis of informatively censored data, and semiparametric efficiency theory. She received a B.S. in mathematics from

the Universidad de Buenos Aires in Argentina, and M.A. and Ph.D. degrees in statistics from the University of California at Berkeley.

**DANIEL SCHARFSTEIN** is a professor and director of the graduate program in the Department of Biostatistics in the Bloomberg School of Public Health at Johns Hopkins University. His research interests focus on inferences about population parameters when they are not estimable from observed data without the imposition of strong, untestable assumptions. He focuses on both frequentist and Bayesian approaches to evaluating the robustness of results to such assumptions as missing at random and no unmeasured confounding. He received a B.S. in economics and in applied science from the University of Pennsylvania, an M.S. in operations research from Georgia Tech, and M.S. and Ph.D. degrees in biostatistics from the Harvard School of Public Health.

**WEICHUNG (JOE) SHIH** is professor and chair of the Department of Biostatistics at the University of Medicine and Dentistry of New Jersey, where he also holds appointments in the Cancer Institute of New Jersey and the Environmental and Occupational Health Sciences Institute. Previously, he was a senior investigator and director of scientific staff in the Department of Clinical Biostatistics and Research Data Systems at Merck. His research interests include statistical methods for handling missing data and adaptive designs of clinical trials. He is a fellow of ASA and is an elected member of the International Statistics Institute. He received the Excellence in Service Award for participating in the advisory board of FDA of the U.S. Department of Health and Human Services. He received a Ph.D. degree in statistics from the University of Minnesota.

**JAY P. SIEGEL** is group president for biotechnology, immunology, and oncology research and development and worldwide regulatory affairs, quality assurance, and benefit risk management at Johnson & Johnson. He also serves as the president of Centocor Research and Development. In addition to previous positions at Johnson & Johnson, he served as the director of Office of Therapeutics Research and Review in the Center for Biologics Evaluation and Research at FDA, where he was also the founding director of the Division of Clinical Trial Design and Analysis. His areas of research include the development of new biotechnology pharmaceutical products, new uses for approved products, and new technologies for the efficient manufacture of such products. He has received numerous awards from the U.S. Department of Health and Human Services for his government work, including the Distinguished Service Medal, the highest honor awarded by the Public Health Service. He is a fellow of the American College of Physicians and of the Infectious Disease Society of America. He received a B.S. in biology from

California Institute of Technology and an M.D. from the Stanford University School of Medicine.

**HAL STERN** is professor and chair of the Department of Statistics at the University of California at Irvine. Previously, he held faculty positions at Harvard University and at Iowa State University, where he directed graduate studies and held the Laurence H. Baker chair in biological statistics. His interests are in the areas of Bayesian methods, model diagnostics, and statistical applications to biological and social sciences. He received a B.S. in mathematics from the Massachusetts Institute of Technology and M.S. and Ph.D. degrees in statistics from Stanford University.

## COMMITTEE ON NATIONAL STATISTICS

The Committee on National Statistics (CNSTAT) was established in 1972 at the National Academies to improve the statistical methods and information on which public policy decisions are based. The committee carries out studies, workshops, and other activities to foster better measures and fuller understanding of the economy, the environment, public health, crime education, immigration, poverty, welfare, and other public policy issues. It also evaluates ongoing statistical programs and tracks the statistical policy and coordinating activities of the federal government, serving a unique role at the intersection of statistics and public policy. The committee's work is supported by a consortium of federal agencies through a National Science Foundation grant.