

## Preventing and Treating Missing Data in Longitudinal Clinical Trials

### A Practical Guide

---

Recent decades have brought advances in statistical theory for missing data, which, combined with advances in computing ability, have allowed implementation of a wide array of analyses. In fact, so many methods are available that it can be difficult to ascertain when to use which method. This book focuses on the prevention and treatment of missing data in longitudinal clinical trials. Based on his extensive experience with missing data, the author offers advice on choosing analysis methods and on ways to prevent missing data through appropriate trial design and conduct. He offers a practical guide to key principles and explains analytic methods for the non-statistician using limited statistical notation and jargon. The book's goal is to present a comprehensive strategy for preventing and treating missing data, and to make available the programs used to conduct the analyses of the example dataset.

**Craig H. Mallinckrodt** is Research Fellow in the Decision Sciences and Strategy Group at Eli Lilly and Company. Dr. Mallinckrodt has supported drug development in all four clinical phases and in several therapeutic areas. He currently leads Lilly's Advanced Analytics hub for missing data and their Placebo Response Task Force, and is a member of a number of other scientific work groups. He has authored more than 170 papers, book chapters, and texts, including extensive works on missing data and longitudinal data analysis in journals such as *Statistics in Medicine*, *Pharmaceutical Statistics*, the *Journal of Biopharmaceutical Statistics*, the *Journal of Psychiatric Research*, the *Archives of General Psychiatry*, and *Nature*. He currently chairs the Drug Information Association's Scientific Working Group on Missing Data.

## Practical Guides to Biostatistics and Epidemiology

### Series advisors

Susan Ellenberg, *University of Pennsylvania School of Medicine*

Robert C. Elston, *Case Western Reserve University School of Medicine*

Brian Everitt, *Institute for Psychiatry, King's College London*

Frank Harrell, *Vanderbilt University Medical Center, Tennessee*

Jos W. R. Twisk, *VU University Medical Center, Amsterdam*

This series of short and practical but authoritative books is for biomedical researchers, clinical investigators, public health researchers, epidemiologists, and non-academic and consulting biostatisticians who work with data from biomedical, epidemiological, and genetic studies. Some books explore a modern statistical method and its applications, others may focus on a particular disease or condition and the statistical techniques most commonly used in studying it.

The series is for people who use statistics to answer specific research questions. Books will explain the application of techniques, specifically the use of computational tools, and emphasize the interpretation of results, not the underlying mathematical and statistical theory.

### Published in the series

*Applied Multilevel Analysis*, by **Jos W. R. Twisk**

*Secondary Data Sources for Public Health*, by **Sarah Boslaugh**

*Survival Analysis for Epidemiologic and Medical Research*, by **Steve Selvin**

*Statistical Learning for Biomedical Data*, by **James D. Malley, Karen G. Malley,**  
and **Sinisa Pajevic**

*Measurement in Medicine*, by **Henrica C.W. deVet, Caroline B. Terwee,**  
**Lidwine B. Mokkink, and Dirk L. Knol**

*Genomic Clinical Trials and Predictive Medicine*, by **Richard M. Simon**

# Preventing and Treating Missing Data in Longitudinal Clinical Trials

A Practical Guide



Craig H. Mallinckrodt



CAMBRIDGE  
UNIVERSITY PRESS



Shaftesbury Road, Cambridge CB2 8EA, United Kingdom  
One Liberty Plaza, 20th Floor, New York, NY 10006, USA  
477 Williamstown Road, Port Melbourne, VIC 3207, Australia  
314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India  
103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment, a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of education, learning and research at the highest international levels of excellence.

[www.cambridge.org](http://www.cambridge.org)

Information on this title: [www.cambridge.org/9781107031388](http://www.cambridge.org/9781107031388)

© Craig H. Mallinckrodt 2013

This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press & Assessment.

First published 2013

*A catalogue record for this publication is available from the British Library*

*Library of Congress Cataloging-in-Publication data*

Mallinckrodt, Craig, 1958–

Preventing and treating missing data in longitudinal clinical trials:

A practical guide / Craig Mallinckrodt.

pages cm – (Practical guides to biostatistics and epidemiology)

Includes bibliographical references and index.

ISBN 978-1-107-03138-8 (hardback) – ISBN 978-1-107-67915-3 (paperback)

1. Clinical trials – Longitudinal studies. 2. Medical sciences – Statistical methods.

3. Regression analysis – Data processing. I. Title.

R853.C55M3374 2013

610.72'4–dc23 2012038442

ISBN 978-1-107-03138-8 Hardback

ISBN 978-1-107-67915-3 Paperback

Cambridge University Press & Assessment has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

## Contents

<i>List of Figures</i>	<i>page xi</i>
<i>List of Tables</i>	<i>xiii</i>
<i>Acknowledgments</i>	<i>xv</i>
<i>Preface</i>	<i>xvii</i>

### PART I: BACKGROUND AND SETTING

1	Why Missing Data Matter	3
2	Missing Data Mechanisms	9
	2.1 Introduction	9
	2.2 Missing Data Taxonomy	10
3	Estimands	15
	3.1 Introduction	15
	3.2 Hypotheses	16
	3.3 Considerations	19

### PART II: PREVENTING MISSING DATA

4	Trial Design Considerations	23
	4.1 Introduction	23
	4.2 Design Options to Reduce Missing Data	23
	4.3 Considerations	31

<b>viii</b>	<b>Contents</b>	
5	Trial Conduct Considerations	33
	5.1 Introduction	33
	5.2 Trial Conduct Options to Reduce Missing Data	33
	5.3 Considerations	35
<b>PART III: ANALYTIC CONSIDERATIONS</b>		
6	Methods of Estimation	39
	6.1 Introduction	39
	6.2 Least Squares	40
	6.3 Maximum Likelihood	44
	6.4 Generalized Estimating Equations	46
	6.5 Considerations	47
7	Models and Modeling Considerations	49
	7.1 Introduction	49
	7.2 Correlation between Repeated Measurements	49
	7.3 Time Trends	52
	7.4 Model Formulation	53
	7.5 Modeling Philosophies	56
8	Methods of Dealing with Missing Data	59
	8.1 Introduction	59
	8.2 Complete Case Analysis	59
	8.3 Simple Forms of Imputation	60
	8.4 Multiple Imputation	63
	8.5 Inverse Probability Weighting	65
	8.6 Modeling Approaches	67
	8.7 Considerations	69
<b>PART IV: ANALYSES AND THE ANALYTIC ROAD MAP</b>		
9	Analyses of Incomplete Data	73
	9.1 Introduction	73
	9.2 Simple Methods for Incomplete Data	76
	9.3 Likelihood-Based Analyses of Incomplete Data	78

<b>ix</b>	<b>Contents</b>	
	9.4 Multiple Imputation-Based Methods	81
	9.5 Weighted Generalized Estimating Equations	84
	9.6 Doubly Robust Methods	87
	9.7 Considerations	88
10	MNAR Analyses	91
	10.1 Introduction	91
	10.2 Selection Models	93
	10.3 Shared Parameter Models	94
	10.4 Pattern-Mixture Models	94
	10.5 Controlled Imputation Methods	96
	10.6 Considerations	101
11	Choosing Primary Estimands and Analyses	103
	11.1 Introduction	103
	11.2 Estimands, Estimators, and Choice of Data	103
	11.3 Considerations	109
12	The Analytic Road Map	111
	12.1 Introduction	111
	12.2 The Analytic Road Map	112
	12.3 Testable Assumptions	114
	12.4 Assessing Sensitivity to Missing Data Assumptions	115
	12.5 Considerations	118
13	Analyzing Incomplete Categorical Data	121
	13.1 Introduction	121
	13.2 Marginal and Conditional Inference	121
	13.3 Generalized Estimating Equations	124
	13.4 Random-Effects Models	125
	13.5 Multiple Imputation	127
	13.6 Considerations	128
14	Example	129
	14.1 Introduction	129
	14.2 Data and Setting	129
	14.3 Objectives and Estimands	130

<b>x</b>	<b>Contents</b>	
	14.4 Analysis Plan	131
	14.5 Results	136
	14.6 Considerations	145
15	Putting Principles into Practice	147
	15.1 Introduction	147
	15.2 Prevention	148
	15.3 Analysis	150
	15.4 Software	151
	15.5 Concluding Thoughts	152
	<i>Bibliography</i>	153
	<i>Index</i>	161



## List of Figures

1.1.	Power for the contrast between drug and control from 10,000 simulated clinical trials with low, medium, and high rates of dropout	<i>page 7</i>
6.1.	Scatter plot and best fit regression line in a scenario with larger errors	40
6.2.	Scatter plot and best fit regression line in a scenario with smaller errors	41
12.1.	An analytic road map for continuous endpoints in longitudinal clinical trials	113
14.1.	Kaplan-Meier plot of days on treatment	138
14.2.	Visitwise means for completers and for patients who discontinued treatment early	138

## List of Tables

1.1.	Hypothetical Trial Results	<i>page</i> 5
1.2.	Hypothetical Trial Results Under Different Assumptions About the Missing Outcomes	5
3.1.	Proposed Estimands and Their Key Attributes	18
3.2.	An Additional Estimand and Its Key Attributes	19
6.1.	Calculating Variance and Standard Deviation	42
6.2.	Probabilities of Binomial Outcomes with True Proportions of 0.2 and 0.6	45
9.1.	Hypothetical Data Set Used to Illustrate Common Analyses for Incomplete Data	74
9.2.	Results from Analyses of Complete Data with and Without Baseline as a Covariate	75
9.3.	Predicted Values for Selected Subjects from Analyses of Complete Data with a Simple Model and a Model that Included Baseline Values as a Covariate	76
9.4.	Results from Least Squares Analysis of Incomplete Data Based on Complete Cases, Last Observation Carried Forward, and Baseline Observation Carried Forward	77
9.5.	Results from Likelihood-Based Analyses of Complete and Incomplete Data, with a Model Including Baseline as a Covariate	79
9.6.	Predicted Values for Selected Subjects from Analyses of Complete Data with a Simple Model and a Model that Included Baseline as a Covariate	80
9.7.	Results from Multiple Imputation–Based Analyses of Incomplete Data	82

**xiv**      **List of Tables**

9.8.	Observed and Imputed Values for Selected Subjects in Multiple Imputation Analysis of Incomplete Data	83
9.9.	Results from Weighted GEE Analyses of Incomplete Data	85
9.10.	Individual Subject Data and Weighting Factors for Weighted GEE Analyses	86
10.1.	Hypothetical Observed and Imputed Data to Illustrate Placebo Multiple Imputation	100
11.1.	Estimands and Their Key Attributes	104
11.2.	Hypothetical Data Used to Illustrate Estimation of Estimand 1	105
11.3.	Results from Likelihood-Based Analyses of Complete and Incomplete Data	105
14.1.	Visitwise Means for Completers and Patients that Discontinued Treatment Early	137
14.2.	Results from Varying Correlation Structures in Likelihood-Based Analyses of the Primary Outcome	139
14.3.	Results from the Primary Outcome with Data from Influential Sites Excluded	139
14.4.	Results from the Primary Outcome with Data from Influential Patients Excluded	140
14.5.	Results from the Primary Outcome with Data from Patients with Aberrant Residuals Excluded	140
14.6.	Results from Inclusive Modeling wGEE and MI Analyses of the Primary Outcome	141
14.7.	Results from Selection Model Analyses of the Primary Outcome	141
14.8.	Results from Pattern-Mixture Model Analyses of the Primary Outcome	142
14.9.	Results from Shared-Parameter Model Analyses of the Primary Outcome	143
14.10.	Results from Placebo Multiple Imputation Analyses of the Primary Outcome	143
14.11.	Summary of Results from Influence, Correlation, and Residual Diagnostics	144
14.12.	Summary of Missing Data Sensitivity Analysis Results	144

## Acknowledgments

It has been my good fortune to collaborate with many excellent researchers in the field of missing data. These collaborations were of great benefit to this book. First, many thanks to those who were forced to read early versions of this book and provided valuable feedback: Geert Molenberghs (Universiteit Hasselt, Diepenbeek), Lei Xu (Eli Lilly, Indianapolis), and Adam Meyers (BioGen Idec, Boston).

For assistance in developing the programs used to analyze example data: Ilya Lipkovich (Quintiles, Indianapolis), Hank Wei (Eli Lilly, Indianapolis), Qun Lin (Eli Lilly, Indianapolis), and Dustin Ruff (Eli Lilly, Indianapolis).

For collaborations over the years that significantly influenced the content of this book: Caroline Beunckens (Universiteit Hasselt, Diepenbeek), James Carpenter (London School of Hygiene and Tropical Medicine), Raymond Carroll (Texas A&M University, College Station), Christy Chuang-Stein (Pfizer, New York), Scott Clark (Eli Lilly, Indianapolis), Mike Detke (MedAvante, Hamilton), Ivy Jansen (Universiteit Hasselt, Diepenbeek), Chris Kaiser (Eli Lilly, Indianapolis), Mike Kenward (London School of Hygiene and Tropical Medicine), Peter Lane (Glaxosmithkline, Harlow), Andy Leon (Weill Medical College, Cornell, New York), Stacy Lindborg (BioGen Idec, Boston), Rod Little (University of Michigan, Ann Arbor), James Roger (London School of Hygiene and Tropical Medicine), Steve Ruberg (Eli Lilly, Indianapolis), Shuyi Shen (Genentech, Oceanside), Cristina Sotto (Universiteit Hasselt, Diepenbeek), Birhanu Teshome (Universiteit Hasselt, Diepenbeek),

**xvi**      **Acknowledgments**

Herbert Thijs (Universiteit Hasselt, Diepenbeek), and Russ Wolfinger (SAS, Cary).

To Donna and Marissa, for your understanding and for helping make the time possible to work on this book, and for the encouragement and support needed to finish it!

## Preface

This book focuses on the prevention and treatment of missing data in longitudinal clinical trials with repeated measures, such as are common in later phases of medical research and drug development. Recent decades have brought advances in statistical theory, which, combined with advances in computing ability, have allowed implementation of a wide array of analyses. In fact, so many methods are available that it can be difficult to ascertain when to use which method. A danger in such circumstances is to blindly use newer methods without proper understanding of their strengths and limitations, or to disregard all newer methods in favor of familiar approaches.

Moreover, the complex discussions on how to analyze incomplete data have overshadowed discussions on ways to prevent missing data, which would of course be the preferred solution. Therefore, preventing missing data through appropriate trial design and conduct is given significant attention in this book. Nevertheless, despite all efforts at prevention, missing data will remain an ever-present problem and analytic approaches will continue to be an important consideration.

Recent research has fostered an emerging consensus regarding the analysis of incomplete longitudinal data. Key principles and analytic methods are explained in terms non-statisticians can understand. Although the use of equations, symbols, and Greek letters to describe the analyses is largely avoided, sufficient technical detail is provided so readers can take away more than a peripheral understanding of the methods and issues. For those with in-depth statistical interests, reference to more technical literature is provided.

Part I begins with illustrations of how missing data erode the reliability and credibility of medical research. Subsequent chapters discuss missing

**xviii**      **Preface**

data mechanisms and the estimands (what is to be estimated) of interest in longitudinal trials with incomplete data. Part II covers trial design and conduct features that can help prevent missing data. Part III includes chapters on common methods of estimation, data and modeling considerations, and means of dealing with missing data (e.g., imputation). Part IV ties together the topics covered in Part III to illustrate various analyses applicable to incomplete longitudinal data. Small example data sets are used to illustrate and explain key analyses. An actual clinical trial data set is the focal point for proposing and implementing an overall analytic strategy that includes sensitivity analyses for assessing the impact of missing data.

This strategy is referred to as the analytic road map. A road map is different from driving instructions. Unlike driving directions, a road map does not chart a specific course, with instructions on exactly how far to go and when to turn. Instead, the road map lays out the alternatives so that the best route for a particular situation can be chosen.

The concluding chapter refocuses on the key issues covered throughout the book, presents a comprehensive strategy for preventing and treating missing data, and makes available the programs used to conduct the analyses of the example dataset.