JAMES R CARPENTER, JONATHAN W BARTLETT,
TIM P MORRIS, ANGELA M WOOD,
MATTEO QUARTAGNO, MICHAEL G KENWARD

Multiple Imputation and its Application

Second Edition

STATISTICS IN PRACTICE

WILEY

Multiple Imputation and its Application

Statistics in Practice

Founding Editor

Vic Barnett
Nottingham Trent University, UK

Statistics in Practice is an important international series of texts which provide detailed coverage of statistical concepts, methods and worked case studies in specific fields of investigation and study.

With sound motivation and many worked practical examples, the books show in down-to-earth terms how to select and use an appropriate range of statistical techniques in a particular practical field within each title's special topic area.

The books provide statistical support for professionals and research workers across a range of employment fields and research environments. Subject areas covered include medicine and pharmaceutics; industry, finance and commerce; public services; the earth and environmental sciences, and so on.

The books also provide support to students studying statistical courses applied to the above areas. The demand for graduates to be equipped for the work environment has led to such courses becoming increasingly prevalent at universities and colleges.

It is our aim to present judiciously chosen and well-written workbooks to meet everyday practical needs. Feedback of views from readers will be most valuable to monitor the success of this aim.

Multiple Imputation and its Application

Second Edition

James R. Carpenter
London School of Hygiene & Tropical Medicine, London, UK
MRC Clinical Trials Unit, University College London, UK

Jonathan W. Bartlett
London School of Hygiene & Tropical Medicine, London, UK

Tim P. Morris
MRC Clinical Trials Unit, University College London, UK

Angela M. Wood
University of Cambridge, Cambridge, UK

Matteo Quartagno
MRC Clinical Trials Unit at University College London, UK

Michael G. Kenward
Ashkirk, UK

WILEY

This edition first published 2023

Edition History: John Wiley & Sons Ltd (1e, 2013)

Registered Offices
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Limit of Liability/Disclaimer of Warranty
While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials or promotional statements for this work. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

Library of Congress Cataloging-in-Publication Data applied for

ISBN: 9781119756088 (Hardback); 9781119756095 (Adobe PDF); 9781119756101 (ePub)

Cover design: Wiley
Cover image: Courtesy of Harvey Goldstein

Set in 10/12pt TimesLTStd by Straive, Chennai, India

Contents

Preface to the second edition

No study of any complexity manages to collect all the intended data. Analysis of the resulting partially collected data must therefore address the issues raised by the missing data. Beyond simply estimating the proportion of missing values, the interplay between the substantive questions and the reasons for the missing data is crucial. There is no simple, universal, solution.

Suppose, for a substantive question at hand, the consequences of missing data in terms of bias and loss of precision are non-trivial. Then the analyst must make a set of assumptions about the reasons, or mechanisms, causing data to be missing, and perform an inferentially valid analysis under these assumptions. In this regard, analysis of a partially observed dataset is the same as any statistical analysis; the difference is that when data are missing we cannot assess the validity of these assumptions in the way we might do in a regression analysis, for example. Hence, sensitivity analysis, where we explore the robustness of inference to different assumptions about the reasons for missing data, is important.

Given a set of assumptions about the reasons data are missing, there are a number of statistical methods for carrying out the analysis. These include the expectation-maximization (EM) algorithm, inverse probability (of non-missingness) weighting, a full Bayesian analysis and, depending on the setting, a direct application of maximum likelihood. These methods, and those derived from them, each have their own advantages in particular settings. We focus on multiple imputation for its practical utility, broad applicability, and relatively straightforward application. Since the first edition was published ten years ago, new applications of multiple imputation have continued to emerge and we have had to be selective in what we cover. The topics included are those we have found most relevant for our research and teaching.

Like the first edition, the book is divided into three parts. Part I lays the foundations, with an introductory chapter outlining the issues raised by missing data, followed by a chapter describing the theoretical foundations of multiple imputation. Part II describes the application of multiple imputation for standard regression analyses, explaining how MI can be used for continuous, categorical, and ordinal data. Part III describes how to apply MI in a range of practical settings, specifically analysis with non-linear relationships, analysis of survival data, development and validation of prognostic models, analysis with multilevel data structures, sensitivity analysis, handling measurement error, analysis involving weights, and causal inference. We conclude with a chapter outlining some broad practical points on the application of multiple imputation. While readers may wish to read only specific relevant chapters in Part III, Chapter 14 is intended to be relevant to all readers. We illustrate ideas with a range of examples from the medical and social sciences, reflecting the wide application that MI has seen in recent years.

Each chapter concludes with a range of exercises, designed to consolidate and deepen understanding of the material. The computer-based exercises have been designed with `R` and `Stata` users in mind. The book's home page at https://missingdata.lshtm.ac.uk contains both (i) hints for the exercises (including suggestions for `R` and `Stata` code) and (ii) full solutions where applicable.

We welcome feedback from readers. Please email `james.carpenter@lshtm.ac.uk` in the first instance.

James R. Carpenter, Jonathan W. Bartlett, Tim P. Morris,
Angela M. Wood, Matteo Quartagno and Mike G. Kenward
September 2022

Data acknowledgements

We are grateful to the following:

1. AstraZeneca for permission to use data from the 5-arm asthma study in examples in Chapters 1, 3, 6, and 10;
2. GlaxoSmithKline for permission to use data from the dental pain study in Chapter 4;
3. Mike English (Director, Child and Newborn Health Group, Nairobi, Kenya) for permission to data from a multi-faceted intervention to implement guidelines and improve admission paediatric care in Kenyan district hospitals in Chapter 9;
4. Peter Blatchford for permission to use data from the Class Size Study (Blatchford *et al.*, 2002) in Chapter 9;
5. Sara Schroter for permission to use data from the study to improve the quality of peer review in Chapter 10.

In Chapter 12, we used data from the Millennium Cohort Study available through the UK Data Service (ukdataservice.ac.uk), study number 2000031.

In Chapters 1, 5, 9, and 10, we have analysed data from the Youth Cohort Time Series for England, Wales, and Scotland, 1984–2002 First Edition, Colchester, Essex, published by and freely available from the UK Data Archive, Study Number SN 5765. Thanks to Vernon Gayle for introducing us to these data.

In Chapter 6, we have analysed data from theAlzheimer's Disease Neuro-imaging Initiative (ADNI) database (adni.loni.ucla.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this book. A complete listing of ADNI investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

The ADNI was launched in 2003 as a public–private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org.

Data collection and sharing for this project was funded by the ADNI (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organisation is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

In Chapter 11, we have analysed data from the US National Health and Nutrition Examination Survey (NHANES), which is freely available from the US Centers for Disease Control and Prevention. We thank the participants and survey organisers of NHANES.

Acknowledgements

No book of this kind is written in a vacuum, and we are grateful to many friends and colleagues.

In particular, we would like to thank the following for discussions and/or inputs that have contributed to this book or shaped our thinking through the years: John Carlin, Orlagh Carroll, Stephen Cole, Suzie Cro, Rhian Daniel, Dan Jackson, Ruth Keogh, Katherine Lee, Clemence Leyrat, Geert Molenberghs, Tra My Pham, Freya Roberts, James Roger, Jeremy Saxe, Shaun Seaman, Jonathan Sterne, Kate Tilling, and Ian White.

Specific thanks to Tetiana Gorbach for checking the exercises in Chapter 2.

James would like to acknowledge many years' collaboration with Harvey Goldstein, who died in 2020.

James would like to thank Mike Elliott, Rod Little, Trivellore Raghunathan, and Jeremy Taylor for facilitating a visit to the Institute for Social Research and Department of Biostatistics at the University of Michigan, Ann Arbor, in Summer 2011, when the first draft of the first edition was written.

Thanks to Tim Collier for the anecdote in Section 1.3.

We also gratefully acknowledge funding support from the ESRC (three-year fellowship for James Carpenter, RES-063-27-0257, and follow-on funding RES-189-25-0103) and MRC (grants G0900724, G0900701 and G0600599, MC_UU_12023/21, MC_UU_12023/29, MC_UU_00004/07, MC_UU_00004/09 and MC_UU_00004/06, MR/T023953/1).

We would also like to thank Richard Davies and Kathryn Sharples at Wiley for their encouragement and support with the first edition, and Kimberly Monroe-Hill and Alison Oliver for their support and patient encouragement with the second edition.

Lastly, thanks to our families for their forbearance and understanding over the course of this project.

Despite the encouragement and support of those listed above, the text inevitably contains errors and shortcomings, for which we take full responsibility.

James R. Carpenter, Jonathan W. Bartlett, Tim P. Morris,
Angela M. Wood, Matteo Quartagno, and Mike G. Kenward
September 2022

Glossary

Indices and symbols

i	indexes units, often individuals
j	indexes variables in the dataset
n	total number of units in the dataset
p	depending on context, number of variables in the dataset or number of parameters in a statistical model
W, X, Y, Z	random variables
$Y_{i,j}$	ith observation on jth variable, $i = 1, \dots, n; j = 1, \dots, p$
R	response indicator, where $R_i = 1$ denotes observed data for unit i and $R_i = 0$ indicates missing data
θ	generic parameter
$\boldsymbol{\theta}$	generic parameter column vector, typically p by 1
$\alpha, \beta, \gamma, \delta$	regression coefficients
$\boldsymbol{\beta}$	column vector of regression coefficients, typically p by 1.
$*$	a random draw, typically from a probability distribution but sometimes data/a datum
$U(.)$	a generic score statistic

Matrices

$\mathbf{\Omega}$	Matrix, typically of dimension $p \times p$
$\mathbf{\Omega}_{i,j}$	i, jth element of $\mathbf{\Omega}$
$\mathbf{\Omega}^T$	Transpose of $\mathbf{\Omega}$, so that $\mathbf{\Omega}^T_{i,j} = \mathbf{\Omega}_{i,j}$
$\mathbf{Y}_j = (Y_{1,j}, \dots, Y_{n,j})^T$	n by 1 column vector of observations on variable j
$\mathrm{tr}(\mathbf{\Omega})$	Sum of diagonal elements of $\mathbf{\Omega}$, i.e. $\sum \mathbf{\Omega}_{i,i}$ known as the trace of the matrix

Abbreviations

CAR	censoring at random
CNAR	censoring not at random
EM	expectation maximisation
FCS	full conditional specification

FEV_1	forced expiratory volume in one second (measured in litres)
FMI	fraction of missing information
IPW	inverse probability weighting
IV	instrumental variable
MAR	missing at random
MCAR	missing completely at random
MI	multiple imputation
MNAR	missing not at random
PMM	predictive mean matching

Probability distributions

$f(\, . \,)$	Probability distribution function
$F(\, . \,)$	Cumulative distribution function
$\mid$	to be verbalised 'given' or 'conditional on', as in $f(Y \mid X)$ 'the probability distribution function of Y given [conditional on] X'

Miscellaneous

We use the terms *complete records* and *incomplete records* rather than *complete cases* and *incomplete cases*, respectively.

For generic regression of Y on $\mathbf{X}$, we use the terms *outcome* or *dependent variable* for Y and *covariates* or *independent variables* for X.

PART I

FOUNDATIONS

1

Introduction

Collecting, analysing, and drawing inferences from data are central to research in the medical and social sciences. Unfortunately, for any number of reasons, it is rarely possible to collect all the intended data. The ubiquity of missing data, and the problems this poses for both analysis and inference, has spawned a substantial statistical literature dating from 1950s. At that time, when statistical computing was in its infancy, many analyses were only feasible because of the carefully planned balance in the dataset (for example the same number of observations on each unit). Missing data meant the available data for analysis were unbalanced, thus complicating the planned analysis and in some instances rendering it infeasible. Early work on the problem was therefore largely computational (e.g. Healy and Westmacott, 1956, Afifi and Elashoff, 1966, Orchard and Woodbury, 1972, Dempster *et al.*, 1977).

The wider question of the consequences of non-trivial proportions of missing data for inference was neglected until the seminal paper by Rubin (1976). This set out a typology for assumptions about the reasons for missing data and sketched their implications for analysis and inference. It marked the beginning of a broad stream of research about the analysis of partially observed data. The literature is now huge and continues to grow, both as methods are developed for large and complex data structures, and as increasing computer power and suitable software enables researchers to apply these methods.

For a broad overview of the literature, a good place to start for applied statisticians is Little and Rubin (2019). They give a good overview of likelihood methods and an introduction to multiple imputation. Allison (2002) presents a less technical overview. Schafer (1997) is more algorithmic, focusing on the expectation maximisation (EM) algorithm and imputation using the multivariate normal and general location model. Molenberghs and Kenward (2007) focus on clinical studies, while Daniels and Hogan (2008) focus on longitudinal studies with a Bayesian emphasis.

Multiple Imputation and its Application, Second Edition.
James R. Carpenter, Jonathan W. Bartlett, Tim P. Morris, Angela M. Wood, Matteo Quartagno and Michael G. Kenward.

The above books concentrate on the parametric approaches. However, there is also a growing literature based around using inverse probability weighting, in the spirit of Horvitz and Thompson (1952), and associated doubly robust methods. In particular, we refer to the work of Robins and colleagues (e.g. Robins and Rotnitzky, 1995, Scharfstein *et al.*, 1999). Vansteelandt *et al.* (2009) give an accessible introduction to these developments. A comparison with multiple imputation in a simple setting is given by Carpenter *et al.* (2006). The pros and cons are debated in Kang and Schafer (2007) and the theory is brought together by Tsiatis (2006).

This book is concerned with a particular statistical method for analysing and drawing inferences from incomplete data called *multiple imputation* (MI). Initially proposed by Rubin (1987) in the context of surveys, increasing awareness among researchers about the possible effects of missing data (e.g. Klebanoff and Cole, 2008) has led to an upsurge of interest (e.g. Sterne *et al.* (2009), Kenward and Carpenter (2007), Schafer (1999a), Rubin (1996)), fuelled by the increasing availability of software and computing power.

MI is attractive because it is both practical and widely applicable. Well-developed statistical software (see, for example, issue 45 of the Journal of Statistical Software) has placed MI within the reach of most researchers in the medical and social sciences, whether or not they have undertaken advanced training in statistics. However, the increasing use of MI in a range of settings beyond that originally envisaged has led to a bewildering proliferation of algorithms and software. Further, the implications of the underlying assumptions in the context of the data at hand are often unclear.

We are writing for researchers in the medical and social sciences with the aim of clarifying the issues raised by missing data, outlining the rationale for MI, explaining the motivation and relationship between the various imputation algorithms and describing and illustrating its application in various settings and to some complex data structures.

Throughout most of the book (with the partial exception of Chapter 8), we will assume that a key aim of analysis with incomplete data is to recover the information lost due to missing data. More specifically, we will take the 'substantive model' as the model that would be used with complete data. We can then define certain desirable properties of our estimator with incomplete data. First, it should be unbiased for the value of the parameter we would see with complete data. Second, it should have low variance. Third, we should have a reliable variance formula and a means of constructing confidence intervals with the advertised coverage.

In the context of multiple imputation, it is worth noting that these remain our aims; the aim of multiple imputation is not to accurately predict the missing values. Rubin (1996) describes it as follows:

> 'Judging the quality of missing data procedures by their ability to recreate the individual missing values […] does not lead to choosing procedures that result in valid inference, which is our objective'.

An objection may be that the ability to perfectly predict missing values *would* result in valid inference; however, in our view, this hypothetical scenario would be one in which data are not really 'missing'.

Central to the analysis of partially observed data is an understanding of why the data are missing and the implications of this for the analysis. This is the focus of the remainder of this chapter. Introducing some of the examples that run through the book, we show how Rubin's typology (Rubin, 1976) provides the foundational framework for understanding the implications of missing data.

1.1 Reasons for missing data

In this section, we consider possible reasons for missing data, illustrate these with examples, and note some preliminary implications for inference. We use the word 'possible' advisedly, since we can rarely be sure of the mechanism giving rise to missing data. Instead, a range of possible mechanisms are consistent with the observed data. In practice, we therefore wish to analyse the data under different mechanisms to establish the robustness of our inference in the face of uncertainty about the missingness mechanism.

All datasets consist of a series of *units* each of which provides information on a series of *items*. For example, in a cross-sectional questionnaire survey, the units would be individuals, and the items their answers to the questions. In a household survey, the units would be households, and the items information about the household and members of the household. In longitudinal studies, units would typically be individuals, while items would be longitudinal data from those individuals. In this book, units

Figure 1.1 Detail from a senior mandarin's house front in New Territories, Hong Kong.

therefore correspond to the highest level in multi-level (i.e. hierarchical) data, and unless stated otherwise, data from different units are statistically independent.

Within this framework, it is useful to distinguish between units where all the information is missing, termed *unit non-response* and units who contribute partial information, termed *item non-response*. The statistical issues are the same in both cases and both can in principle be handled by MI. However, the main focus of this book is the latter.

Example 1.1 Mandarin tableau

Figure 1.1, which is also shown on the book's cover, shows part of the frontage of a senior mandarin's house in the New Territories, Hong Kong. We suppose interest focuses on characteristics of the figurines, for example their number, height, facial characteristics, and dress. Unit non-response then corresponds to missing figurines, and item non-response to damaged – hence, partially observed – figurines. □

1.2 Examples

We now introduce two key examples, which we return to throughout the book.

Example 1.2 Youth Cohort Study (YCS)

The Youth Cohort Study of England and Wales (YCS) is an ongoing UK government-funded representative survey of pupils in England and Wales at school-leaving age (School year 11, age 16–17) (UK Data Archive, 2007). Each year that a new cohort is surveyed, detailed information is collected on each young person's experience of education, and their qualifications, as well as information on employment and training. A limited amount of information is collected on their personal characteristics, family, home circumstances, and aspirations.

Over the life cycle of the YCS, different organisations have had responsibility for the structure and timings of data collection. Unfortunately, the documentation of older cohorts is poor. Croxford *et al.* (2007) have deposited a harmonised dataset that comprises YCS cohorts from 1984 to 2002 (UK Data Archive Study Number 5765 dataset). We consider data from pupils attending comprehensive schools from five YCS cohorts; these pupils reached the end of Year 11 in 1990, 1993, 1995, 1997, and 1999.

We explore relationships between Year 11 educational attainment (the General Certificate of Secondary Education) and key measures of social stratification. The units are pupils, and the items are measurements on these pupils, and a non-trivial number of items are partially observed. □

Example 1.3 Randomised controlled trial of patients with chronic asthma

We consider data from a five-arm asthma clinical trial to assess the efficacy and safety of budesonide, a second-generation glucocorticosteroid, on patients with chronic asthma. Four hundred and seventy-three patients with chronic asthma were enrolled in the 12-week randomised, double-blind, multi-centre parallel-group trial, which compared the effect of a daily dose of 200, 400, 800, or 1600 mcg of budesonide with placebo.

Key outcomes of clinical interest include patients' peak expiratory flow rate (their maximum speed of expiration in litres/minute) and their forced expiratory volume, FEV_1 (the volume of air, in litres, the patient with fully inflated lungs can breathe out in one second). In summary, the trial found a statistically significant dose–response effect for the mean change from baseline over the study for both morning peak expiratory flow, evening peak expiratory flow, and FEV_1 at the 5% level.

Budesonide-treated patients also showed reduced asthma symptoms and bronchodilator use compared with placebo, while there were no clinically significant differences in treatment-related adverse experiences between the treatment groups. Further details about the conduct of the trial, its conclusions, and the variables collected can be found elsewhere (Busse *et al.*, 1998). Here, we focus on FEV_1 and confine our attention to the placebo and lowest active dose arms. FEV_1 was collected at baseline, then 2, 4, 8, and 12 weeks after randomisation. The intention was to compare FEV_1 across treatment arms at 12 weeks. However (excluding three patients whose participation in the study was intermittent), only 37 out of 90 patients in the placebo arm, and 71 out of 90 patients in the lowest active dose arm, still remained in the trial at 12 weeks. □

1.3 Patterns of missing data

It is very important to investigate the patterns of missing data before embarking on a formal analysis. This can throw up vital information that might otherwise be overlooked and may even allow the missing data to be traced. For example, when analysing the new wave of a longitudinal survey, a colleague's careful examination of missing data patterns established that many of the missing questionnaires could be traced to a set of cardboard boxes. These turned out to have been left behind in a move. They were recovered, and the data entered.

Most statistical software now has tools for describing the pattern of missing data. Key questions concern the extent and patterns of missing values, and whether the pattern is *monotone* (as described in the next paragraph), as if it is, this can considerably speed up and simplify the analysis.

Missing data in a set of p variables are said to follow a *monotone missingness pattern* if the variables can be re-ordered such that, for every unit i and variable j,

1. if unit i is observed on variable j, where $j = 2, \dots, p$ it is observed on all variables $j' < j$, and
2. if unit i is missing on variable j, where $j = 2, \dots, p$ it is missing on all variables $j' > j$.

A natural setting for the occurrence of monotone missing data is a longitudinal study, where units are observed either until they are lost to follow up, or the study concludes. A monotone pattern is thus inconsistent with patterns of interim missing data, where some units are observed for a period, missing for the subsequent period, but then observed. Questionnaires may also give rise to monotone missing data patterns when individuals systematically answer each question in turn from the beginning till they either stop or complete the questionnaire. In other settings, it may be possible to re-order items to achieve a monotone pattern.

Example 1.2 Youth Cohort Study *(ctd)*

Table 1.1 shows the covariates we consider from the YCS. There are no missing data in the variables *cohort* and *boy*. The missingness pattern for General Certificate of Secondary Education (GCSE) score and the remaining two variables is shown in Table 1.2. In this example, it is not possible to re-order the variables (items) to obtain a monotone pattern due, for example to pattern 3 ($N = 697$). □

Example 1.3 Asthma study *(ctd)*

Table 1.3 shows the withdrawal pattern for the placebo and lowest active dose arms (all the patients are receiving their randomised medication). We have removed three patients with unusual interim missing data from Table 1.3 and all our analyses. The remaining missingness pattern is monotone in both treatment arms. □

Table 1.1 YCS variables for exploring the relationship between Year 11 attainment and social stratification.

Variable name	Description
Cohort	Year of data collection: 1990, ’93, ’95, ’97, ’99
Boy	Indicator variable for boys
Occupation	Parental occupation, categorised as managerial, intermediate, or working
Ethnicity	Categorised as Bangladeshi, Black, Indian, other Asian, Other, Pakistani, or White

Table 1.2 Pattern of missing values in the YCS data.

Pattern	GCSE score	Occupation	Ethnicity	No.	% of total
1	✓	✓	✓	55145	87%
2	✓	·	✓	6821	11%
3	·	✓	✓	697	1%
4	✓	·	·	592	1%