Yu-Kang Tu
Darren C. Greenwood *Editors*

Modern Methods for Epidemiology

Springer

Modern Methods for Epidemiology

Yu-Kang Tu • Darren C. Greenwood
Editors

Modern Methods for Epidemiology

Springer

Editors
Yu-Kang Tu
Division of Biostatistics
Leeds Institute of Genetics
Health and Therapeutics
University of Leeds
Leeds, UK

Darren C. Greenwood
Division of Biostatistics
Leeds Institute of Genetics
Health and Therapeutics
University of Leeds
Leeds, UK

ISBN 978-94-007-3023-6 ISBN 978-94-007-3024-3 (eBook)
DOI 10.1007/978-94-007-3024-3
Springer Dordrecht Heidelberg New York London

Library of Congress Control Number: 2012934174

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Statistical methods are important tools for scientific research to extract information from data. Some statistical methods are simple whilst others are more complex, but without such methods our data are just numbers and useless to our understanding of the world we are living in. In epidemiology, researchers use more advanced and complex statistical methods than colleagues who work with experimental data, often under more controlled conditions than can be achieved with the larger datasets and more "real-life" conditions required by observational data. The issues of observational data are not just about the amount of data but also the quality of data. Epidemiological data usually contains missing values in some variables for some patients, and the instruments used for data collection may be less accurate or precise than those used for experimental data. Therefore, textbooks of epidemiology often contain much discussion of statistical methods for dealing with those problems in analysis and interpretation of data, and very often they also contain some discussion of the philosophy of science. This is because elaborating causes and their consequences from observational data usually requires certain epistemological theories about what constitutes "causes" and "effects".

Routine applications of advanced statistical methods on real data have become possible in the last 10 years because desktop computers have become much more powerful and cheaper. However, proper understanding of the challenging statistical theory behind those methods remains essential for correct application and interpretation, and rarely seen in the medical literature. This textbook contains a general introduction to those modern statistical methods that are becoming more important in epidemiological research, to provide a starting point for those who are new to epidemiology, and for those looking for guidance in more modern statistical approaches. For those who wish to pursue these methods in greater depth, we provide annotated lists of further reading material, which we hope are useful for epidemiological researchers who wish to overcome the mathematical barrier of applying those methods to their research.

The Centre for Epidemiology and Biostatistics at the University of Leeds, United Kingdom, where we have been working for many years, has a masters

degree programme in the field of Statistical Epidemiology, aiming to provide a unique opportunity for researchers to obtain further training in both epidemiology and statistics. Several modules in the programme teach statistical methods that are not discussed in standard textbooks of epidemiology or biostatistics. For example, very few textbooks of epidemiology discuss multilevel modelling, whilst very few textbooks of biostatistics discuss confounding using Directed Acyclic Graphs (DAGs). Here we bring these two important topics in modern epidemiology together in the same book. For topics such as G-estimation, latent class analysis, regression trees, or generalised additive modelling, students have previously had to dig into monographs or journal articles for those methods, which are usually aimed at more advanced readers. We feel that there is a need for a textbook that can be used for teaching modern, advanced statistical methods to postgraduate students studying epidemiology and biostatistics and also a good source of self-learning for researchers in epidemiology and medicine. We therefore invited colleagues from Leeds, Bristol, Cambridge and London in the United Kingdom, and colleagues in Denmark and South Africa, all leading experts in their respective fields, to contribute to writing this book.

This volume contains 17 chapters dedicated to modern statistical methods for epidemiology. The opening chapter starts with the most important, but also the most controversial concept in epidemiology: confounding. Before the introduction of DAGs into epidemiology, the definition of confounding was sometimes confusing and deficient. Graham Law and his co-authors provide an overview of DAGs and show why DAGs are so useful in statistical reasoning surrounding the potentially causal relationships in observational research. Chapter 2 discusses another troubling issue in observational research: incomplete data or missing data. James Carpenter and his colleagues provide an overview of incomplete data problems in biomedical research and various strategies for imputing missing values. At the heart of all epidemiology is an appropriate assessment of exposure. Chapter 3 discusses this problem of measurement error in epidemiological exposures. Darren Greenwood provides a concise introduction to the problems caused by measurement error and outlines some potential solutions that have been suggested. Chapter 4 discusses the issue of selection bias in epidemiology, a particular problem in the context of case-control studies. Graham Law and his co-authors use DAGs as a tool to explain how this problem affects the results of observational studies and how it may be resolved.

Chapter 5 discusses multilevel modelling for clustered data, a methodology also widely used in social sciences research. Andrew Blance provides an overview of the basic principles of multilevel models where random effects are assumed to follow a normal distribution. In Chap. 6, Mark Gilthorpe and his co-authors discuss the issues of outcomes formed from a mixture of distributions and use zero-inflated models as an example. Chapter 7 can be seen as an extension of Chaps. 5 and 6. Wendy Harrison and her co-authors discuss scenarios where the assumption that random effects follow a normal distribution is not appropriate, instead assuming a discrete distribution, describing discrete components that can be viewed as latent classes. Chapters 8 and 9 both discuss Bayesian approach for sparse data, where

observations of events are scattered in space or time, Chap. 8 discussing bivariate disease mapping and Chap. 9 discussing multivariate survival mapping models. Samuel Manda and Richard Feltbower use data from the Yorkshire region in the United Kingdom and from the South Africa to illustrate these approaches. In Chap. 10, Darren Greenwood discusses meta-analysis of observational data. This is more complex than meta-analysis of randomised controlled trials because of greater heterogeneity in design, analysis, and reporting of outcome and exposure variables. Methods and software packages available to deal with those issues are discussed.

Chapter 11 returns to the concepts introduced in the opening chapters, focusing on the resemblance between DAGs and path diagrams. Yu-Kang Tu explains how to translate regression models into both DAGs and path diagrams and how those graphical presentations can inform us the causal relations in the data. Chapter 12 discusses latent growth curve modelling, which is equivalent to multilevel modelling for longitudinal data analysis. Yu-Kang Tu and Francesco D'Auito use a dataset from Periodontology to illustrate the flexibility of latent growth curve modelling in accommodating nonlinear growth trajectories. These ideas are extended in Chap. 13 by allowing random effects to follow a discrete distribution. Darren Dahly shows how growth mixture modelling can be used to uncover distinctive early growth trajectories, which may be associated with increased disease risk in later life. Chapter 14 focuses on the problem of time-varying confounding, and Kate Tilling and her colleagues explain how G-estimation may be used to overcome it.

Chapter 15 discusses generalised additive modelling for exploring non-linear associations between variables. Robert West gives a concise introduction to this complex method and shows how it can be extended to multivariable models. He then continues to explain regression trees and other advanced methods for classification of variables in Chap. 16. These methods have become popular in biomedical research for modelling decision-making. In the final chapter, Mark Gilthorpe and David Clayton discuss the intricate issues surrounding statistical and biological interaction. They use the example of gene-environment interaction to show that statistical interactions and biological interactions are different concepts and much confusion arises where the former is used to describe the latter.

Editing this book has been an exciting experience, and we would like to thank all the authors for their excellent contributions. We also want to thank Dr Brian Cattle for his help with the preparation of the book and our editors in Springer for their patience with this project.

Leeds, UK

Yu-Kang Tu
Darren C. Greenwood

Contents

Contributors

Vibeke Baelum School of Dentistry, Faculty of Health Sciences, University of Aarhus, Aarhus, Denmark

Paul D. Baxter Division of Biostatistics, Centre for Epidemiology and Biostatistics, Leeds Institute of Genetics, Health & Therapeutics, University of Leeds, Leeds, UK

Andrew Blance Division of Biostatistics, Centre for Epidemiology and Biostatistics, Leeds Institute of Genetics, Health & Therapeutics, University of Leeds, Leeds, UK

James R. Carpenter Department of Medical Statistics Unit, London School of Hygiene and Tropical Medicine, London, UK

Yaping Cheng Division of Biostatistics, Centre for Epidemiology and Biostatistics, Leeds Institute of Genetics, Health & Therapeutics, University of Leeds, Leeds, UK

David G. Clayton Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory, Cambridge University, Cambridge, UK

Francesco D'Auito Department of Periodontology, Eastman Dental Institute, University College London, London, UK

Darren L. Dahly Division of Biostatistics, Centre for Epidemiology and Biostatistics, Leeds Institute of Genetics, Health & Therapeutics, University of Leeds, Leeds, UK

Vanessa Didelez School of Mathematics, University of Bristol, Bristol, UK

Amy Downing Cancer Epidemiology Group, Centre for Epidemiology and Biostatistics, University of Leeds, Leeds, UK

George T.H. Ellison Division of Biostatistics, Centre for Epidemiology and Biostatistics, Leeds Institute of Genetics, Health & Therapeutics, University of Leeds, Leeds, UK

Richard G. Feltbower Division of Epidemiology, Centre for Epidemiology and Biostatistics, Leeds Institute of Genetics, Health & Therapeutics, University of Leeds, Leeds, UK

Morten Frydenberg Department of Biostatistics, Faculty of Health Sciences, Institute of Public Health, University of Aarhus, Aarhus, Denmark

Mark S. Gilthorpe Division of Biostatistics, Centre for Epidemiology and Biostatistics, Leeds Institute of Genetics, Health & Therapeutics, University of Leeds, Leeds, UK

Harvey Goldstein Medical Statistics Unit, London School of Hygiene and Tropical Medicine, London, UK

Graduate School of Education, University of Bristol, Bristol, UK

Rosie Green Department of Nutrition and Public Health Intervention Research, London School of Hygiene and Tropical Medicine, London, UK

Darren C. Greenwood Division of Biostatistics, Centre for Epidemiology and Biostatistics, Leeds Institute of Genetics, Health & Therapeutics, University of Leeds, Leeds, UK

Wendy Harrison Division of Biostatistics, Centre for Epidemiology and Biostatistics, Leeds Institute of Genetics, Health & Therapeutics, University of Leeds, Leeds, UK

Michael G. Kenward Department of Medical Statistics Unit, London School of Hygiene and Tropical Medicine, London, UK

Graham R. Law Division of Biostatistics, Centre for Epidemiology and Biostatistics, Leeds Institute of Genetics, Health & Therapeutics, University of Leeds, Leeds, UK

Samuel O.M. Manda Biostatistics Unit, South Africa Medical Research Council, Pretoria, South Africa

Jonathan A.C. Sterne School of Social and Community Medicine, University of Bristol, Bristol, UK

Kate Tilling School of Social and Community Medicine, University of Bristol, Bristol, UK

Yu-Kang Tu Division of Biostatistics, Centre for Epidemiology and Biostatistics, Faculty of Medicine and Health, Leeds Institute of Genetics, Health & Therapeutics, University of Leeds, Leeds, UK

Robert M. West Division of Biostatistics, Centre for Epidemiology and Biostatistics, Leeds Institute of Genetics, Health & Therapeutics, University of Leeds, Leeds, UK

Chapter 1
Confounding and Causal Path Diagrams

Graham R. Law, Rosie Green, and George T.H. Ellison

1.1 Causal Models

The issue of causation is a challenging one for epidemiologists. Politicians and the public want to know whether something of concern causes a disease or influences the effectiveness of healthcare services. However, the training provided to statisticians, and to scientists more generally, tends to stress that non-experimental research will only ever offer evidence for association and that suitably designed experimental studies are required to offer robust evidence of causation. In the real world, where experimental data are rare, difficult or impossible to produce, the extent to which associations between variables can and should be interpreted as evidence of causality is less a technical question than a philosophical, moral, cultural or political one. These issues have been discussed at some length elsewhere (see for example Susser 1973; and Pearl 1998, 2000), and although these influence the extent to which associational evidence from non-experimental studies is (and should be) used in real-world settings, the following Chapter will focus on the more technical issue of strengthening the causal inferences drawn from non-experimental data by using causal path diagrams when designing and describing the analysis of data from non-experimental studies. In this chapter we will introduce causal path diagrams (specifically Directed Acyclic Graphs; DAGs) and explore the issue of confounding.

G.R. Law (✉) • G.T.H. Ellison
Division of Biostatistics, Centre for Epidemiology and Biostatistics, Leeds Institute of Genetics, Health & Therapeutics, University of Leeds, Room 8.01, Worsley Building, LS2 9LN Leeds, UK
e-mail: g.r.law@leeds.ac.uk

R. Green
Department of Nutrition and Public Health Intervention Research, London School of Hygiene and Tropical Medicine, London, UK

Y.-K. Tu and D.C. Greenwood (eds.), *Modern Methods for Epidemiology*,
DOI 10.1007/978-94-007-3024-3_1,

1.1.1 Directed Acyclic Diagrams (DAGs), Nomenclature and Notation

A causal path diagram is a visual summary of the likely (and, where relevant, the speculative) causal links between variables. Constructing these diagrams is based on *a priori* knowledge and, in the case of speculative and hypothesised relationships being explored in the analysis, on conjecture. Causal path diagrams have been used informally for many years in causal analysis and in recent years have been formally developed for use in expert-systems research (Greenland et al. 1999). Although such diagrams are beginning to be adopted by the epidemiological community (Hoggart et al. 2003; Hernandez-Dìaz et al. 2006; Shrier and Platt 2008; Head et al. 2008, 2009; Geneletti et al. 2011; Tu and Gilthorpe 2012), a causal diagram is still a novel epidemiological tool which can be used in a variety of ways: to think clearly about how exposure, disease and potential confounder variables, relevant to the research hypothesis, are related to each other; to communicate these inter-relationships to academic and professional audiences; to indicate which variables were important to measure; and to inform the statistical modelling process – particularly the identification of confounding, confounders and competing exposures.

In this Chapter we discuss the use of causal path diagrams (Pearl 2000), specifically Directed Acyclic Graphs (DAGs), to develop models that can inform the analysis of one variable (the 'exposure') as a potential cause of another (the 'outcome'). Within epidemiology, such analyses include exploring: the potential role of risk factors (as 'exposures') in the aetiology of disease (where the 'outcome' is the prevalence, incidence or severity of disease); and the role of specific characteristics of healthcare systems (where these characteristics are the 'exposures') in the effective and efficient delivery of health services (where this constitutes the 'outcome').

1.1.1.1 Nomenclature and the Construction of DAGs

The nomenclature of DAGs is still evolving, and can be off-putting to the uninitiated, particularly when accompanied by statistical notation (such as that developed by Geneletti et al. (2009)). However, the terminology that is developing helps to specify each of the components of DAGs in a way that facilitates their consistent application and further utility. And, with this in mind, we have provided a comprehensive glossary of terms in Table 1.1, and a more detailed explanation of these below.

Nodes, Arcs and Directed Arcs

In statistical parlance, each variable in a DAG is represented by a *node* (also known as a *vertex*), and relationships between two variables are depicted by a line connecting the nodes, called an *arc* (or alternatively an *edge* or a *line*). A *directed arc* indicates

Table 1.1 Glossary of terms for causal diagrams

Term	Description
Ancestor	A variable that causes another variable in a *causal path* in which there are intermediary variables situated along the *causal*/*direct path* between them
Arc	A line with one arrow that connects two *nodes* (synonymous with *edge* and *line*)
Backdoor path	A path that goes against the direction of the *arc* on the path, but can then follow or oppose the direction of any subsequent *arc*
Blocked path	A path that contains at least one *collider*
Causal path	A path that follows the direction of the *arcs* (synonymous with *direct path*)
Child	A variable that is directly affected by another variable, with no intermediary variables situated along the *causal path* between them
Collider	A variable that a *path* both enters and exits via *arcs*
Descendant	A variable that is caused by one or more preceding variables in a direct *causal path* in which there is one or more intermediary variables situated along the *causal path* between them
Direct path	A path that follows the direction of the *arcs* (synonymous with *causal path*)
Directed arc	An arrow between two variables that indicates a known, likely or speculative causal relationship between them
Edge	A line with one arrow that connects two *nodes* (synonymous with *arc* and *line*)
Line	A line with one arrow that connects two *nodes* (synonymous with *arc* and *edge*)
Node	A point within the diagram which denotes a variable, such as the (key) exposure variable of interest, the (key) outcome (of interest), and another covariates (synonymous with *vertex*)
Parent	A variable that directly affects another variable, with no intermediary variables situated along the *causal path* between them
Path	An unbroken route between two variables, in either direction (synonymous with *route*)
Route	An unbroken route between two variables, in either direction (synonymous with *path*)
Unblocked path	A *path* that does not contain a *collider*
Vertex	A point within the diagram which denotes a variable, such as the (key) exposure variable of interest, the (key) outcome (of interest), and another covariates (synonymous with *vertex*)

known (i.e. from a firm grasp of established functional biological, social or clinical relationships between variables); *likely* (i.e. from previous robust empirical studies); or *speculative* (i.e. hypothesised) relationships between any two variables, with an arrow representing causality – the direction of causality following in the direction of the arrow. For example, 'X causes Y' would be represented as $X \rightarrow Y$, where X and Y are nodes (or vertices) and the arrow between them is an arc (or edge or line).

Parents, Children, Ancestors and Descendants

DAGs are usually depicted with the nodes arranged in a temporal and thus causal sequence, with the preceding variables to the left of the diagram and subsequent

variables to the right. This is not mandatory, but can help when deciding which of two closely related variables precedes the other and acts as its cause. A node immediately preceding another node to which it is connected (i.e. a node at the non-arrow end of an arc) is known as a *parent* of the node at the arrow end of the arc, which is in turn known as a *child*. Thus, in the example $X \rightarrow Y$, X is the parent node and Y is the child. Similarly, a node 'preceding' another node but connected to another node via at least one other node is known as an *ancestor,* whereas the preceding node from which it is separated is known as a *descendent*. Therefore, in the example $X \rightarrow Y \rightarrow Z$, X is the ancestor of Z, and Z is the descendent of X; while Y (which is a child of X and a parent of Z) lies on the causal pathway between X and Z.

Directed Paths, Backdoor Paths, Colliders and Blocked Paths

A *path* is the sequence of arcs connecting two or more nodes, thus $X \rightarrow Y \rightarrow Z$ is the path (or route) connecting the nodes X and Z. A *direct* (or *causal*) *path* is one where the arcs all follow in the direction of causality. In contrast, a *backdoor path* is where one exits a node along an arc pointing into it, against the causal direction, to another node across any number of arcs pointing in either direction. For example, when $X \leftarrow Z \rightarrow Y$ backdoor path exists between X and Y via Z. A node becomes a *collider* where both arcs of the path entering and leaving the node have arrows pointing into it. For example, Y is a collider when $X \rightarrow Y \leftarrow Z$ and a path is *blocked* if it contains at least one collider. A *directed acyclic graph* occurs if no directed path forms a closed loop, reflecting the assumption that that no variable can cause itself (an assumption that may limit the utility of DAGs for modelling functional processes containing positive or negative feedback loops).

Identification of Arcs

All arcs in a DAG reflect *a priori* presumptions about cause and effect in a specific context. Some of these presumptions will be based on *known* causal relationships between variables (drawing on established functional biological, social and clinical processes); others on *likely* causal relationships (drawing, for example, on the statistical findings of previous robust empirical studies); as well as speculative relationships (drawing on unsubstantiated hypotheses – including the specific hypotheses being tested in the analyses). These arc-related presumptions cannot (and should not) be inferred empirically from data on which the analyses will be conducted, but must be drawn from established mechanisms or strong research evidence, both of which are crucial for developing an accurate DAG as the basis on which suitable statistical analyses can then be designed (Tu et al. 2004; Weinberg 2005; Tu and Githorpe 2012).

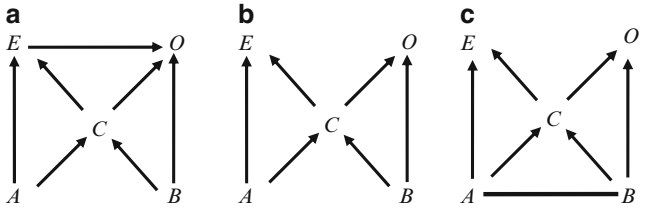

Fig. 1.1 An example of Directed Acyclic Graphs. Key to variables: *E* exposure, *O* outcome, *A*, *B*, *C* additional

1.1.1.2 Notation

An additional technical approach to represent the statistical relationships between variables (as nodes) in causal path diagrams is to use the notation developed by Geneletti et al. (2009). For example, the notation $A \perp\!\!\!\perp B|C$ signifies A as being independent of B given C, where *A*, *B* and *C* are known variables. For example the DAG represented in Fig. 1.1 consists of 5 variables: *E* the exposure of interest, *O* the outcome of interest and 3 other additional variables *A*, *B*, and *C*.

In Fig. 1.1a the exposure, *E*, causes the outcome, *O*. This can be represented as

$$O \not\perp\!\!\!\perp E$$

1.1.2 The Speculative Nature of DAGs and Their Limitations

In most research studies the causal pathways described and summarised within causal path diagrams are not established (i.e. 'proven') causal relationships, but are in the main based on evidence from whatever previous studies are available. Proof in this context is essentially more of a philosophical than a scientific concept, and can be subject to intense debate. The pathways included in the diagrams are therefore often based on: (i) incomplete or predominantly theoretical understanding (rather than established knowledge) of the functional relationships between the variables involved; (ii) the statistical findings of empirical research which may not themselves be definitive; and (iii) hypotheses based on putative, tentative or speculative beliefs about the sorts of relationships that exist – not least the one between the exposure(s) and the outcomes that the study set out to address. These three very different ingredients involved in the conceptualisation of causal pathways are important to recognise as they influence both: the extent to which different causal path diagrams can be drawn for the same variables (reflecting different views of what is *known*, *likely* or *speculative*) and the extent

to which these different diagrams might be more (or less) useful for generating robust evidence of causality between two or more specific variables. Despite this DAGs are useful because they force researchers to make explicit their presumptions about the relationships between pairs of variables, whether or not these presumptions prove to be correct. Other analysts are then able to critique, (re)interpret and (where necessary) repeat and improve on the analyses conducted, based on different presumptions or firmer knowledge of the causal relationships involved.

However, alongside their assumption that no variable can be its own cause (which, as mentioned earlier, reduces the utility of DAGs for modelling systems containing feedback loops), a key limitation of DAGs is that they will only ever be able to include variables (as nodes) that are (as Donald Rumsfeld would have it) 'knowns' (i.e. are recognised as conceptual entities within the epistemological context concerned). Likewise, analyses based on DAGs will only ever be able to include those variables for which data are available (i.e. that have been measured – in Donald Rumsfeld's parlance, 'known knowns'). This is a fundamental limitation of all analyses of data from non-randomised non-experimental studies, not least because unknown or unmeasured confounders cannot be taken into account when modelling or analysing potential causal relationships. Nonetheless, using DAGs to identify the most appropriate statistical analyses for any given set of measured variables will reduce the likelihood that these are subject to confounding (from known and measured confounders) and help others to critique, (re)interpret and (where necessary and possible) repeat and improve on the analyses conducted. These then are the core strengths of using DAGs to design the analysis of data from non-experimental studies – strengths we explore in greater detail in Sect. 1.2.4, below.

Meanwhile, another potential limitation of DAGs is that, despite the potential for visual complexity (particularly for those DAGs with more than a handful of nodes), they are essentially an oversimplification of the causal relationships between variables. For example, a causal diagram does not indicate whether an effect is harmful or protective or whether effect modification is actually occurring (Hernan et al. 2004 – although Weinberg 2007 recently suggested how DAGs might be modified to include this), nor does a causal diagram identify whether a cause is sufficient or necessary to elicit the outcome(s) involved (Rothman 1976). Nonetheless, it bears restating that one of the key strengths of such diagrams is that they enable researchers to think clearly and logically about the research question at hand, and to make explicit any presumptions that are being made about the (presumed) relationships between the pairs of variables involved. This visual summary can then be used as an aid to communicate these inter-relationships to academic and professional audiences and to explicitly identify, for example, if important variables or relationships are missing from or misrepresented in the diagram or, indeed, whether any of the presumed relationships are contentious.

1.1.3 Notation

One way to represent the statistical relationships between variables (as nodes) in causal diagrams is to use the notation developed by Geneletti and colleagues (2009). For example, the notation $A \perp\!\!\!\perp B|C$ signifies A is independent of B given C, where A, B and C are known variables. For example the DAG represented in Fig. 1.1 consists of 5 variables: E the exposure of interest, O the outcome of interest and 3 other additional variables A, B, and C.

In Fig. 1.1a the exposure, E, causes the outcome, O. This can be represented as

$$O \not\perp\!\!\!\perp E$$

A common practice in epidemiology is to consider other covariates at the same time as the exposure. For example, these might include a measure of socio-economic status, age or sex.

1.2 Confounding and Confounders

Confounding is a central concept in epidemiological research. It is a process that can generate biased results when examining the association between exposure and outcome. Historically there have been many definitions of confounding, but they may be divided broadly into two main types: "comparability-based" and "collapsibility-based" (Greenland and Robins 1986):

- In terms of the "comparability-based" definition, confounding is said to occur when there are differences in outcome in the unexposed and exposed populations that are not due to the exposure, but are due to other variables that may be referred to as 'confounders'. This results in bias in the estimate of the effect of a particular exposure on a particular outcome (McNamee 2003).
- In terms of the "collapsibility-based" definition, confounding may be: (i) reduced by adjusting the data by the potential confounder; or (ii) eliminated by stratifying the data by the potential confounder (McNamee 2003). This second definition is therefore based solely on statistical considerations and confounding is said to occur if there is a difference between unadjusted or "collapsed" estimates of the effect of exposure on outcome and estimates that have been adjusted or stratified by the potential confounder.

Although these two definitions of confounding have often been considered indistinguishable, focusing on confounding as a causal rather than a statistical issue leads one to adopt the "comparability-based" definition over the "collapsibility-based" definition (Greenland and Morgenstern 2001). The "comparability-based" definition of confounding can then be used to establish which epidemiological criteria can and should be used to establish whether a variable should be classified as a confounder or not. First, the variable concerned must be a cause of the outcome (or a proxy for a

cause) in unexposed subjects (i.e. a 'risk factor'). Second, the variable concerned must be correlated with the exposure variable within the study population concerned. Finally, the variable concerned must not be situated on any causal pathway between exposure and outcome (Hennekens and Buring 1987). More recently, the last of these three conditions has been replaced with an even stricter one: the variable concerned must not be an effect of the exposure (McNamee 2003).

Confounding can exist at the level of the population, or as a consequence of a biased sample. This is an important point; the consideration of confounding should not be solely based on a study sample, indeed it may be the case that apparent confounding in a study is due to sampling and is not true confounding in the population as a whole. Many studies are often able to identify more than one relevant confounder in their analyses, and we will discuss later how one might establish whether the analyses have accounted for a sufficient set of confounders (or whether too few/too many have been included in the analyses: see Sect. 1.2.3, below).

We may have a situation where $E \rightarrow O$ and $A \rightarrow O$, but there is no association between E and A. This happens in a successfully randomised controlled trial (RCT) where baseline variables (A) are balanced between groups – so A is independent of E (due to the success of randomisation for treatments). Nonetheless, because A is a competing exposure for O, the precision with which the relationship between E and O is characterised improves after adjusting for A.

1.2.1 Confounding and DAGs

The use of causal path diagrams to identify confounding and confounders in epidemiological research was introduced by Greenland et al. (1999). The use of DAGs represents a rigorous approach to assessing confounding and identifying confounders, and DAGs are particularly useful given the absence of any objective criteria or test for establishing the presence (or absence) of confounding. Compared with the use of traditional epidemiological criteria to identify confounders, the key additional insight that DAGs provide is the extent to which adjustment for a confounding variable may create further confounding which in turn requires adjustment (Greenland et al. 1999). DAGs also allow analysts to select a subset of potential confounders (i.e. a subset selected from all identified potential confounders) that is sufficient to adjust for potential confounding. Indeed, DAGs can be used to identify the full range of such subsets and thereby test and select the most appropriate one to use (Greenland et al. 1999).

1.2.2 Identifying Confounding

In order to explain how DAGs can be used to determine whether there is potential for confounding in the apparent relationship between an exposure and an outcome let us first use a simple DAG as an example (see Fig. 1.1a). To determine

if confounding is present the following algorithm is applied to the DAG (Greenland et al. 1999):

(i) delete all single headed arrows that exit from the exposure variable (i.e. remove all exposure effects); and
(ii) check if there are any unblocked backdoor paths from exposure to outcome (i.e. examine whether exposure and outcome have a common cause).

If there are no unblocked backdoor paths the relationship between exposure and outcome should not be subject to potential confounding (albeit from those variables that have been measured precisely and are available for inclusion in the model and its related statistical analyses). For example, in order to check if the relationship of E on O in Fig. 1.1a is subject to potential confounding:

(i) the arrow between E and O is deleted; and
check if there are any unblocked backdoor paths from E to O (there are three: $E \leftarrow C \rightarrow O$; $E \leftarrow A \rightarrow C \rightarrow O$; and $E \leftarrow C \leftarrow B \rightarrow O$)

Because there are three unblocked backdoor paths from E to O, there is the potential for confounding of the effect of E on O, because we can identify three potential confounders – A and B, and C (which lies on the pathway between A and O, and between B and E). When confounding is present an additional algorithm can be applied to identify where adjustment is required and of which variables (see Sect. 1.2.3 below).

However, before we address this it is important to point out that two variables that are not associated with each other, and that share a child (or descendent) that is a confounder, may also become associated within at least one stratum of the confounder. This is a well-established observation in epidemiological research (Weinberg 1993). Adjusting for one confounder may also alter the associations between other variables. In a DAG, this is equivalent to creating a non-directed arc between the two variables and therefore a new backdoor path that has to be dealt with when adjusting for confounding. This can be illustrated using the example in Fig. 1.1a, where controlling only for C in the relationship between E and O may create an association between A and B, because both A and B are parents of C. If this is the case, then A and B must also be included as confounders, otherwise additional confounding will have been introduced by adjustment for C alone.

1.2.3 Sufficient Set of Confounders

Where confounding is present it is usually possible and desirable to identify a subset of variables (S) using a DAG that is sufficient to address confounding through adjustment. In other words, S constitutes the subset of variables with which it is possible to address all confounding through adjustment. In order to

assesswhether *S* removes all confounding another algorithm is applied to the DAG (Pearl 1993):

(i) delete all single headed arrows that exit from the exposure variable;
(ii) draw non-directed arcs that connect each pair of variables that share a child that is either in *S* or has a descendant in *S* (i.e. account for any associations between variables that are generated by controlling for *S*); and
(iii) check if there are any unblocked backdoor paths from exposure to outcome that do not pass through *S* – if there is no unblocked backdoor path then *S* is sufficient for control of confounding.

If we apply this algorithm to the five-variable DAG described earlier in Fig. 1.1a to check whether a tentative set of variables (*S'*) that contains *A, B* and *C* would be sufficient for controlling for any potential confounding control would involve:

(i) deleting the arrow between *E* and *O*;
(ii) drawing a nondirected arc between *A* and *B* (since *C* is a child of *A* and *B*; see Fig. 1.1c); and
(iii) assessing whether there are no unblocked backdoor paths from *E* to *O* that do not pass through *A*, *B* and *C* (there are none).

Using this approach in this example would therefore lead us to conclude that adjusting for *A*, *B* and *C* would be sufficient to address potential for confounding in the relationship between *E* and *O*.

However, in order to check whether there might be an even smaller *sub*set of the tentative subset of confounders (*S'*; *A*, *B*, and *C*) it is worth exploring the consequences of deleting each of these variables in turn:

Deleting A would still mean that:

- the backdoor path $E \leftarrow A - B \rightarrow O$ in Fig. 1.1c would be blocked at *B*;
- $E \leftarrow A - B \rightarrow C \rightarrow O$ would be blocked at *B*;
- $E \leftarrow A \rightarrow C \rightarrow O$ would be blocked at *C*; and
- $E \leftarrow C \rightarrow O$ would be blocked at *C*.

Therefore *B* and C are minimally sufficient. In other words, it is not necessary to adjust for *A* in addition to *B* and *C*.

Deleting B would mean that:

- the backdoor path $E \leftarrow A - B \rightarrow O$ in Fig. 1.1c would be blocked at *A*;
- $E \leftarrow A - B \rightarrow C \rightarrow O$ would be blocked at *A*;
- $E \leftarrow A \rightarrow C \rightarrow O$ would be blocked at *A*; and
- $E \leftarrow C \rightarrow O$ would be blocked at *C*.

Therefore *A* and *C* (like *B* and *C*, above) would also be minimally sufficient. *Deleting C* would mean that:

- the backdoor path $E \leftarrow A \rightarrow C \rightarrow O$ in Fig. 1.1c would be blocked at *A*;
- $E \leftarrow C \rightarrow B \rightarrow O$ would be blocked at *B*; and
- $E \leftarrow C \rightarrow O$ would be blocked

Therefore *A* and *B* would not be minimally sufficient.

As this example shows, there can be more than one minimally sufficient set (*S*). However, these sets may also vary in size and may not necessarily overlap (Greenland et al. 1999). It can therefore be helpful to identify *all* minimally sufficient sets so that the best one can be chosen for dealing with confounding through adjustment. For example, some sets may need to be rejected if they contain variables that were not measured in the study. Others may be rejected due to concerns about measurement error, or because they contain many more variables than other sets and would thereby generate less precise estimates from multivariable statistical analyses on the sample sizes available. As such an important advantage of using DAGs over traditional approaches to identifying potential confounding is that the latter are usually unable to identify any of the potential sufficient subsets of potential confounders, and all potential confounders would therefore need to be included in the analysis (at cost to the precision of the estimates produced).

1.2.4 Strengths and Weaknesses of Causal Path Diagrams

As we have shown in this chapter, DAGs can be used to identify confounding and confounders in a systematic way, and by helping researchers to identify these objectively and explicitly, DAGs can help to reduce bias and advance debate. Moreover, despite the various limitation mentioned earlier in this Chapter (see Sect. 1.1.2, above), one of the main strengths of using causal path diagrams in epidemiological analyses of data from non-experimental studies is that it enables researchers to think clearly and logically about the *known*, *likely* and *speculative* causal relationships between variables that are relevant to the research hypothesis and related analytical questions. Causal path diagrams thereby facilitate the communication of any causal presumptions that have been made during data analysis to academic and professional audiences using a structured approach that is explicit and easy to critique or re-model.

DAGs also enable the identification of variables that are important to measure in a prospective research study, and thereby improve the efficiency of both data collection and statistical analyses by avoiding the unnecessary measurement or inclusion of variables that are irrelevant to the study and its analysis.

Nonetheless, a somewhat surprising feature of tackling confounding using DAGs is that incorrect specification of the model itself can itself create more problems than it solves. For example, bias may be introduced by including variables that are consequences of the exposure, while additional confounding may be created by including variables that are common descendents of other confounders. Likewise, as we saw earlier, stratification may lead to key changes to some of the paths within the DAG, and these changes may lead to previously blocked paths becoming unblocked and causing further confounding. However, both of these potential flaws can be put to good use in identifying whether adjustment for specific confounders might create new associations between variables that may generate

further confounding that will also need to be addressed. As such, these features are arguably an additional strength of using DAGs in analytical design.

One important weakness of DAGs is that with increasing numbers of highly inter-related variables they can rapidly become visually complex to read. DAGs also represent an inherent oversimplification of causal relationships between variables as they do not indicate whether: any relationships are positive or negative (e.g. harmful or protective); effect modification might occur; each causal relationship is weak or strong; and some of the variables might only be able to cause an effect in combination with other variables.

Moreover, as with all causal models, DAGs are only as good as the functional and empirical knowledge and speculative hypotheses on which they are based. In particular, DAGs may be based on a set of presumptions that are wrong (either as a result of incorrect knowledge, weak empirical evidence or fallacious hypotheses). However, because DAGs ensure that these presumptions are explicitly stated, the key benefit of DAGs is that they facilitate criticism, (re) interpretation and (where necessary) modification of the model to assess whether different conclusions would be reached about: which variables are true confounders (see Chap. 11 on structural equation modelling); and which subset of variables are best to adjust for in order to address confounding while taking into account the availability and quality of data on each of the variables involved.

1.3 Conclusions

Directed acyclic graphs (DAGs) have great potential utility in epidemiological analyses of data from non-experimental studies; not least because they encourage researchers to formally structure presumed and predicted causal pathways. These causal path diagrams are essentially intuitive to construct but nonetheless require considered thought. As with all models, careful interpretation remains imperative. Following established algorithms, they can nonetheless be used to identify sufficient sets of confounders which will greatly advance analytical modelling strategies and their subsequent interpretation, critique, testing and re-modelling by other researchers.

References

Geneletti, S., Richardson, S., & Best, N. (2009). Adjusting for selection bias in retrospective, case-control studies. *Biostatistics, 10*, 17–31.

Geneletti, S., Gallo, V., Porta, M., Khoury, M. J., & Vineis, P. (2011). Assessing causal relationships in genomics: From Bradford-Hill criteria to complex gene-environment interactions and directed acyclic graphs. *Emerging Themes in Epidemiology, 8*, 5.

Greenland, S., & Morgenstern, H. (2001). Confounding in health research. *Annual Review of Public Health, 22*, 189–212.

Greenland, S., & Robins, J. M. (1986). Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology, 15*, 413–419.

Greenland, S., Pearl, J., & Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology, 10*, 37–48.

Head, R. F., Gilthorpe, M. S., Byrom, A., & Ellison, G. T. H. (2008). Cardiovascular disease in a cohort exposed to the 1940–45 Channel Islands occupation. *BMC Public Health, 8*, 303.

Head, R. F., Gilthorpe, M. S., & Ellison, G. T. H. (2009). Cholesterol levels in later life amongst UK Channel Islanders exposed to the 1940–45 German occupation as children, adolescents and young adults. *Nutrition and Health, 20*, 91–105.

Hennekens, C. H., & Buring, J. E. (1987). *Epidemiology in medicine* (1st ed.). Boston/Toronto: Little Brown and Company.

Hernan, M. A., Hernandez-Dìaz, S., & Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology, 15*, 615–625.

Hernandez-Dìaz, S., Schisterman, E. F., & Hernan, M. A. (2006). The birth weight "paradox" uncovered? *American Journal of Epidemiology, 146*, 1115–1120.

Hoggart, C. J., Parra, E. J., Shriver, M. D., Bonilla, C., Kittles, R. A., Clayton, D. G., & McKeigue, P. M. (2003). Control of confounding of genetic associations in stratified populations. *American Journal of Human Genetics, 72*, 1492–1504.

McNamee, R. (2003). Confounding and confounders. *Occupational and Environmental Medicine, 60*, 227–234.

Pearl, J. (1993). Comment: Graphical models, causality and intervention. *Statistical Science, 8*, 266–269.

Pearl, J. (1998). Graphs, causality, and structural equation models. *Sociological Methods and Research, 27*, 226–284.

Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge: University Press.

Rothman, K. J. (1976). Causes. *American Journal of Epidemiology, 104*, 587–592.

Shrier, I., & Platt, R. W. (2008). Reducing bias through directed acyclic graphs. *BMC Medical Research Methodology, 8*, 70.

Susser, M. (1973). *Causal thinking in the health sciences*. New York: Oxford University Press.

Tu, Y.-K., & Gilthorpe, M. S. (2012). *Statistical thinking in epidemiology*. Boca Raton: CRC Press.

Tu, Y.-K., West, R. W., Ellison, G. T. H., & Gilthorpe, M. S. (2004). Why evidence for the fetal origins of adult disease can be statistical artifact: The reversal paradox examined for hypertension. *American Journal of Epidemiology, 161*, 27–32.

Weinberg, C. R. (1993). Toward a clearer definition of confounding. *American Journal of Epidemiology, 137*, 1–8.

Weinberg, C. R. (2005). Barker meets Simpson. *American Journal of Epidemiology, 161*, 33–35.

Weinberg, C. R. (2007). Can DAGs clarify effect modification? *Epidemiology, 18*, 569–572.

Chapter 2
Statistical Modelling of Partially Observed Data Using Multiple Imputation: Principles and Practice

James R. Carpenter, Harvey Goldstein, and Michael G. Kenward

2.1 Introduction

Missing data are inevitably ubiquitous in experimental and observational epidemiological research. Nevertheless, despite a steady flow of theoretical work in this area, from the mid-1970s onwards, recent studies have shown that the way partially observed data are reported and analysed in experimental research falls far short of best practice (Wood et al. 2004; Chan and Altman 2005; Sterne et al. 2009). The aim of this Chapter is thus to present an accessible review of the issues raised by missing data, together with the advantages and disadvantages of different approaches to the analysis.

Section 2.2 gives an overview of the issues raised by missing data, and Sect. 2.3 explores those situations in which a 'complete case' analysis, using those units with no missing data, will be appropriate. Section 2.4 describes the advantages and disadvantages of various methods for the analysis of partially observed data and argues that multiple imputation is the most practical approach currently available to applied researchers. Section 2.5 reviews some key issues that arise when using multiple imputation in practice. We conclude with a worked example in Sect. 2.6 and discussion in Sect. 2.7.

J.R. Carpenter (✉) • M.G. Kenward
Department of Medical Statistics, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK
e-mail: James.Carpenter@lshtm.ac.uk

H. Goldstein
Graduate School of Education, University of Bristol, 35 Berkeley Square, BS8 1JA, Bristol

Y.-K. Tu and D.C. Greenwood (eds.), *Modern Methods for Epidemiology*,
DOI 10.1007/978-94-007-3024-3_2,

2.2 Issues Raised by Missing Data

We illustrate the issues raised by missing data using Fig. 2.1, which shows the frontage of a high-level mandarin's house in the New Territories, Hong Kong.

First, we notice missing data can either take the form of completely missing figurines, or damaged— i.e. partially observed—figurines. The former is analogous to what is usually termed unit non-response, while the latter is analogous to item non-response. However, the statistical issues raised are the same in both cases. For simplicity, we therefore assume there are no completely missing figurines.

Next, we see that the effect of missing data on any inference depends crucially on the question at hand. For instance, if interest lies in the position of the figurines in the tableau shown in Fig. 2.1, then missing data are not a problem. If, instead, interest is in the height, or facial characteristics of the figurines, then missing data raises issues that have to be addressed. Thus, when assessing the impact of missing data it is not the number, or proportion of missing observations per se that is the key, rather the extent of the missing information about the question at hand. Changing the example, if we are interested in the prevalence of a rare disease, missing the disease status of two individuals—potentially non-randomly—out of 1,000 means we have lost a substantial amount of information.

Now suppose we are interested in estimating a facial characteristic—say average hair length—of the four figurines shown. Two are missing their heads, and we cannot be sure why. In order to estimate the average hair length we need to make an assumption about why the two heads are missing, and/or how their mean hair length relates to those whose heads are present. Our assumptions must take one of the following three forms:

1. the reason for the missing heads is random, or at any rate unconnected to any characteristics of the figurines;
2. the reason for the missing heads is not random; but within groups of 'similar' figurines (e.g. with similar neckties) heads are missing randomly, or
3. the reason for the missing heads is not random, and—even within groups of apparently similar figurines—depends on hair length (i.e. depends directly on what we want to measure).

In case 1, the 'data' (hair length) are said to be Missing Completely At Random (MCAR). What is usually termed the missingness mechanism may depend on the position of the figurines relative to missing tiles in the roof above, but is independent of information relevant to the question at hand. Under this assumption there is no difference in the distribution of hair-length between the figurines, so we can get a valid estimate using the complete cases (i.e. figurines with heads). In case 2, the data are said to be Missing At Random (MAR). The reason for the missing data (hair length) depends on the unseen value (hair length) but we can form groups based on observed data (e.g. necktie) within which the reason for the missing data does not depend on the unseen value (missing hair length). If we assume hair length is MAR given necktie, we can estimate hair length among figurines with straight

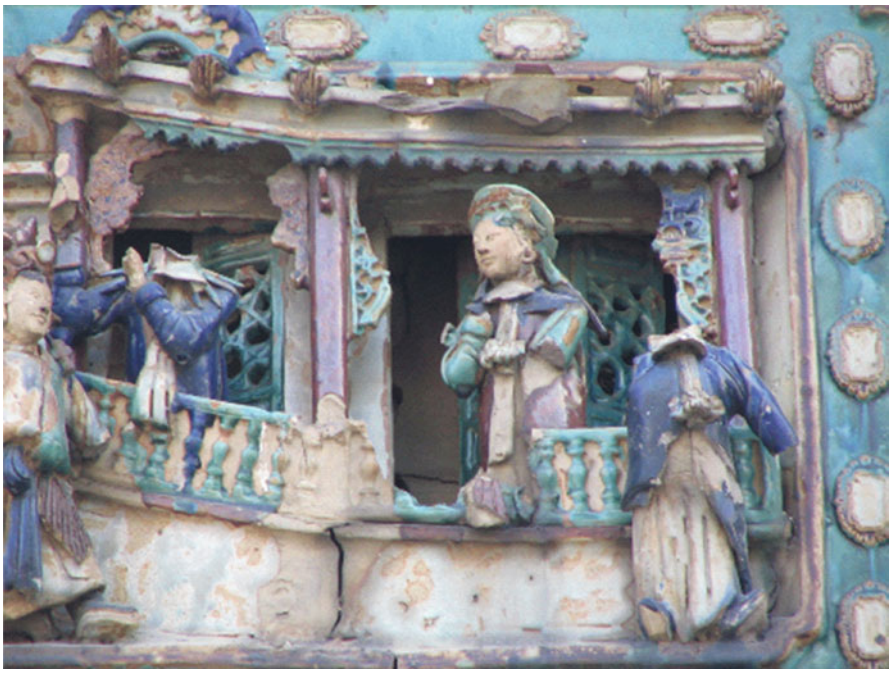

Fig. 2.1 Mandarin's house, New Territories, Hong Kong (Photo H. Goldstein)

neckties, and among those that end in a bobble. We can then calculate a weighted average of these—weighting by the number with each kind of necktie—to estimate mean hair length across the 'population' of figurines.

In case 3, the data are said to be Missing Not At Random (MNAR). In this case, we cannot estimate average hair length across the figurines without knowing either (i) the relationship between the chance of a headless figurine and hair length or (ii) the difference in mean hair length between figurines with, and without, heads.

This terminology was first proposed by Rubin (1976), and despite the slightly counter-intuitive meaning of 'Missing At Random' it is now almost universally used. We now highlight two things, implicit in the above discussion, which are universal in the analysis of partially observed data:

2.2.1 Ambiguity Caused by Missing Data

Given Fig. 2.1, we do not know which of the assumptions 1–3 above is correct; furthermore each has different implications for how we set about validly estimating mean hair length. Therefore, the best we can do is state our assumptions clearly, arrive at valid inference under those assumptions, and finally report how inference

varies with the assumptions. The latter is referred to as sensitivity analysis, and is fundamental to inference from partially observed data. We hope that our inference is pretty robust to different assumptions about the missing data, so that we can be fairly confident about our conclusions. However, as we cannot verify our assumptions using the data at hand, our readers can reasonably be expected to be informed if this is indeed the case.

2.2.2 *Duality of Missingness Mechanism and Distribution of Missing Data Given Observed Data*

Each of the assumptions 1–3 above makes a statement both about the probabilistic mechanism causing the missing data (which we refer to as the missingness mechanism) and the difference between the distribution of the missing data given the observed data. To see this, suppose that Y is hair length, X is characteristics of the body (observed on all figurines) and R = 1 if the head is present and 0 if absent.

Under MCAR, the chance of R = 1 given X, Y —for which we use the notation [R|X, Y]—does not depend on X or Y, that is [R|X, Y] = [R]. This means that the distribution of Y given X does not depend on R. More formally, using the definition of conditional probability,

$$\begin{aligned}[Y|X,R] &= \frac{[Y,X,R]}{[X,R]} = \frac{[R|X,Y][X,Y]}{[R|X][X]} \\ &= \frac{[R][Y,X]}{[R][X]} \text{ (because of MCAR assumption)} \\ &= [Y|X] \end{aligned} \tag{2.1}$$

Thus the missingness mechanism tells us about the distribution of the missing data given the observed, and vice versa.

A similar argument gives (2.1) if data are MAR, for then the chance of R = 1 does not depend on Y once we take X into account, so that [R |Y, X] = [R|X]. Thus, if data are MCAR or MAR, the distribution of the partially observed variables (hair length) given the fully observed ones (body characteristics) is the same across individuals, regardless of whether—for a particular individual—the partially observed variable (hair length) is seen or not.

However, this relationship does not hold if data are MNAR. In that case the chance of R = 1 depends on both X and Y, and this means that the distribution of [Y |X] is different depending on whether Y is observed or not (i.e. whether R = 1 or not). This makes MNAR analyses more difficult, as we either have to say (i) exactly how [R] depends on Y, X or (ii) exactly how [Y |X] differs according to R—i.e. whether Y is observed or not.