Investigating Solutions to Minimise Participation Bias in Case-Control Studies

Claire Michelle Keeble

Submitted in accordance with the requirements

for the degree of

Doctor of Philosophy

The University of Leeds

School of Medicine

May 2016

Contributions

The candidate confirms that the work submitted is her own, except where work which has formed part of jointly authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Chapter 2 contains a review which is published,[1] jointly authored by Claire M Keeble and her three supervisors; Dr Paul D Baxter, Dr Stuart Barber and Dr Graham R Law. The first draft of the review "C Keeble, PD Baxter, S Barber and GR Law. Participation rates in epidemiology studies and surveys: A review 2007–2015. *The Internet Journal of Epidemiology*, 2015" was completed by Claire M Keeble, with all other authors supervising her work. Claire M Keeble had the idea to conduct the review, collected all the data and completed all the written work.

Chapter 3 is based upon a published article,[2] jointly authored by Claire M Keeble and her three supervisors; Dr Stuart Barber, Dr Graham R Law and Dr Paul D Baxter. The first draft of the article "C Keeble, S Barber, GR Law, and PD Baxter. Participation bias assessment in three high impact journals. *Sage Open*, 3(4):1-5, 2013" was completed by Claire M Keeble, with all other authors supervising the work. Claire M Keeble was the researcher mentioned in §3.2 who conducted the data collection and who completed all the written work.

Chapter 4 contains a published article,[3] jointly authored by Claire M Keeble and her three supervisors; Dr Graham R Law, Dr Stuart Barber and Dr Paul D Baxter. The first draft of the review "C Keeble, GR Law, S Barber and PD Baxter. Choosing a method to reduce selection bias: A tool for researchers. *Open Journal of Epidemiology*, 5, 155–162, 2015" was completed by Claire M Keeble, with all other authors supervising her work. Claire M Keeble developed the

research tool for publication and completed all the written work.

Chapter 5 contains an article[4] which has been prepared for submission which is jointly authored by Claire M Keeble, her three supervisors, Dr Peter A Thwaites and Dr Roger C Parslow. Claire M Keeble applied chain event graphs to the diabetes dataset, and completed the first draft for "C Keeble, PA Thwaites, PD Baxter, S Barber, RC Parslow, GR Law. Learning through chain event graphs: The role of maternal factors in childhood type I diabetes". Dr Paul D Baxter, Dr Stuart Barber and Dr Graham R Law gave feedback on the draft. Dr Peter A Thwaites and Dr Roger C Parlsow provided guidance on the chain event graphs and childhood type I diabetes dataset respectively. All calculations and written work were completed by Claire M Keeble.

Chapter 6 contains an article[5] which has been prepared for submission which is jointly authored by Claire M Keeble, her three supervisors and Dr Peter A Thwaites. Claire M Keeble developed all the adaptations in the article and completed the first draft of "C Keeble, PA Thwaites, S Barber, GR Law, PD Baxter. Adaptation of chain event graphs for use with case-control studies". Dr Stuart Barber, Dr Graham R Law and Dr Paul D Baxter provided feedback as her supervisors. Dr Peter A Thwaites provided feedback as an additional author who had worked using chain event graphs. All calculations and written work were completed by Claire M Keeble.

Chapter 7 is based upon a published article,[6] jointly authored by Claire M Keeble, her three supervisors and Dr Roger C Parslow. The initial idea and first draft of the article "C Keeble, S Barber, PD Baxter, RC Parslow, and GR Law. Reducing participation bias in case-control studies: Type 1 diabetes in children and stroke in adults. *Open Journal of Epidemiology*, 4(3):129–134, 2014" were completed by Claire M Keeble. Dr Stuart Barber, Dr Paul D Baxter and Dr Graham R Law (supervisors) and Dr Roger C Parlsow (who was the principal investigator of the original childhood type I diabetes project[7]) provided feedback in a supervisory role. All calculations and written work were completed by Claire M Keeble.

Acknowledgements

Thank you to all three academic supervisors; Dr Graham R Law, Dr Paul D Baxter & Dr Stuart Barber, from the University of Leeds, who have given continued and invaluable support throughout the entire PhD, which is greatly appreciated.

Thank you to Dr Roger C Parslow & the Yorkshire Childhood Diabetes Register for allowing the use of the type I diabetes dataset as an example case-control study throughout the thesis, and for Roger's contributions to the publications on these data. The dataset has provided the opportunity to explore the methods to reduce participation bias beyond hypothetical examples and simulated data, and helped to identify any potential problems which may be encountered with real data. In return, the analyses which have used these data have been published to contribute the findings to the diabetes literature.

Thank you also to Dr Peter A Thwaites for his introduction to the chain event graphs used in Chapters 5 & 6 and his contribution to the original articles using chain event graphs. Chain event graphs have not been used with case-control study data before and hence the publication of these articles should contribute positively to both the case-control and the chain event graph literature.

Thank you to my examiners, Professor Kate Tilling (University of Bristol) & Dr Sarah Fleming (University of Leeds), for agreeing to undertake the examination, for travelling to the University, and for assessing my work.

On a more personal level, thank you to Mr Michael R Keeble, Mrs Margaret E Keeble, Miss Nicola D Keeble, Mr Jamie R Owen, Mr Andrew G Davies, Dr Karen E Porter & Professor Mark S Gilthorpe for their encouragement and support.

Finally, this thesis has been funded by a Medical Research Council (MRC) Capacity Building Studentship, which is very gratefully acknowledged.

Abstract

Case-control studies are used in epidemiology to try to determine variables associated with a disease, by comparing those with the disease (cases) against those without (controls). Participation rates in epidemiology studies have declined over recent years, particularly in the control group where there is less motivation to participate. Non-participation can lead to bias and this can result in the findings differing from the truth.

A literature review of the last nine years shows that non-participation occurred in published studies as recently as 2015, and an assessment of articles from three high impact factor epidemiology journals concludes that participation bias is a possibility which is not always controlled for. Methods to reduce bias resulting from non-participation are provided, which suit different data structures and purposes. A guidance tool is subsequently developed to aid the selection of a suitable approach. Many of these methods rely on the assumption that the data are missing at random. Therefore, a new solution is developed which utilises population data in place of the control data, which recovers the true odds ratio even when data are missing not at random.

Chain event graphs are a graphical representation of a statistical model which are used for the first time to draw conclusions about the missingness mechanisms resulting from non-participation in case-control data. These graphs are also adapted specifically to further investigate non-participation in case-control studies.

Throughout, in addition to hypothetical examples and simulated data, a diabetes dataset is used to demonstrate the methods. Critical comparisons are drawn between existing methods and the new methods developed here, and discussion provided for when each method is suitable. Identification of factors associated with a disease are crucial for improved patient care, and accurate analyses of case-control data, with minimal biases, are one way in which this can be achieved.

Contents

List of Figures

List of Tables

Abbreviations

Table 1: Table of notation and abbreviations.

Notation	Meaning
A	Auxiliary variable
AAPOR	The American Association for Public Opinion Research
AHC	(Bayesian) Agglomerative Hierarchical Clustering (algorithm)
AJE	American Journal of Epidemiology
Amnio	Amniocentesis
ARR	Absolute risk reduction
a	Number of exposed cases in the sample
a'	Number of exposed cases in the target population
B_i	Coefficients for independent, individual, variable values
BP	Blood pressure
b	Number of exposed controls in the sample
b'	Number of exposed controls in the target population
c	Number of non-exposed cases in the sample
c'	Number of non-exposed cases in the target population
C	A chain event graph
Ca	Case
CEG	Chain event graph
CI	Confidence interval
Co	Control

Notation and abbreviations table continues on the next page

Table 1 – Continued.

Notation	Meaning
Csec	Caesarean
CT	Computed tomography
d	Number of non-exposed controls in the sample
d'	Number of non-exposed controls in the target population
D	Disease of interest
DAG	Directed acyclic graph
dp	Decimal place
E	Exposure of interest
e.g.	From Latin ‘exempli gratia’ meaning ‘for example’
Epi.	Epidemiology (Journal)
etc	From Latin ‘et cetera’ meaning ‘and the others’
GP	General practitioner
Γ	Odds of exposure
H	Hayfever
HES	Hospital Episode Statistics
i	Index
i.e.	From Latin ‘id est’ meaning ‘that is’
I	Intermediate position
IDDM	Insulin-dependent diabetes mellitus
IJE	International Journal of Epidemiology
IPW	Inverse probability weighting
ITT	Intention to treat
j	Level
J	Number of strata
k	Edge index
l	Leaves
log	Logarithm

Notation and abbreviations table continues on the next page

Table 1 – Continued.

Notation	**Meaning**
m	Number of edges in a floret
MAP	Maximum a posteriori
MAR	Missing at random
MCAR	Missing completely at random
MNAR	Missing not at random
N/A	Not applicable
n	Number of random variables
NHS	National Health Service
O	Outcome variable
ONS	Office for National Statistics
OR	Odds ratio (in text)
OR	Odds ratio (in a calculation)
P	Number of individuals in the population
p	Participation probability
PhD	Doctor of Philosophy
Pop.	Population
π	Probability
π_d	Probability of disease
RR	Relative risk
RCT	Randomised controlled trial
ρ	Correlation
s	Situation
SES	Socio-economic status
SMS	Short Message Service
Ta	Tablet
THIN	The Health Improvement Network
u	Stage

Notation and abbreviations table continues on the next page

Table 1 – Continued.

Notation	Meaning
U	Set of all stages
UK	United Kingdom
UKCCS	United Kingdom Childhood Cancer Study
w	Position
W	Set of all positions
X, Y, Z	Random variables
X_i	Recorded independent, individual, variable values
$\leftarrow\rightarrow\uparrow\downarrow$	Arcs in a directed graph
✓	Requirement
§	Section number

Chapter 1

Introduction

Case-control studies are used in epidemiology to try to determine variables associated with a disease, by comparing those with the disease (cases) against those without (controls). Participation rates in epidemiology studies have declined over recent years, with efforts to improve participation proving unsuccessful. Case-control studies have been found to be susceptible to non-participation, particularly in the control group where there is less motivation to participate. Non-participation can lead to bias and this can result in the findings differing from the truth. Significant factors may be missed or insignificant variables may be reported, and in extreme instances, the effects of risk factors may be reversed.

Chapter 2 discusses background information relevant to case-control studies and participation bias in more detail. An assessment of the treatment of non-participation in a sample of published epidemiology literature is given in Chapter 3. The presence of participation bias and the action taken by researchers in all articles of an issue of each of the top three impact factor epidemiology journals is reported and summarised. This includes whether the study design was likely to be affected by participation bias, any methods applied to reduce the bias, and whether this method was appropriate.

The current methods proposed to reduce participation bias in case-control studies are summarised and demonstrated in Chapter 4. A childhood type I diabetes dataset is also introduced and used for the remainder of the thesis as a common theme by which to demonstrate methods to reduce participation bias. Further details of the data can be found in Appendix A. The three methods which are most commonly used according to Chapter 3 are each applied to a hypothetical example

and the real diabetes data, and critically assessed. Other methods are described and assessed, with references for further reading. The chapter ends with a flowchart tool, intended as a guide to assist researchers in choosing a suitable method to reduce participation bias.

Having assessed the treatment of participation bias in the literature, and with the limitations of the current methods identified, chain event graphs were applied to case-control studies in Chapter 5. This new approach is able to aid the identification of potential participation bias and assist with the selection of a suitable method to reduce the bias, by investigating the missingness mechanism. Chain event graphs are a relatively recent graphical methodology, developed in artificial intelligence and statistics. Although their use has been demonstrated in cohort studies they have not, according to the literature, been applied to case-control studies. Chapter 5 begins with an introduction to this methodology, followed by application to the diabetes data. This includes theories regarding the missingness mechanism and the likely categories of the missing data. To further increase the utility of chain event graphs with case-control studies, Chapter 6 then proposes adaptations to these graphs for scenarios often encountered with case-control data. Examples are provided for each adaptation and their use in investigating non-participation is explained.

In Chapter 7 a new method is developed and demonstrated, which uses population data in place of control data in a case-control study. This method to reduce participation bias has several advantages over the current methods, such as recovering the true odds ratio when data are missing not at random, and a critical evaluation of this method is provided.

Chapter 8 summarises and critically evaluates the findings from all chapters, including a comparison between the results of the diabetes data from all methods. Suggestions are made for future work, and both the strengths and limitations of the work in the thesis are discussed.

Participation bias affects a range of different study types, and the analysis of *all* of these types is beyond the scope of this thesis. The aim of this thesis is to investigate solutions to minimise participation bias in case-control studies, since they have been reported to be heavily affected by non-participation in the control group. However, findings from this thesis may be relevant to other research affected by participation bias. This includes not only other study designs used in epidemiology, but also areas with similar problems such as non-response in the survey literature.

Chapter 2

Background

This chapter provides definitions and background information relating to both case-control studies and participation bias, which will be required for the later chapters. First, general epidemiological definitions are provided, then case-control studies are described and participation bias is introduced.

2.1 Epidemiological Definitions

For clarity and to ensure accessibility for those from other disciplines, the definitions of some key epidemiological terms are given below.

- **Bias:** Systematic "deviation of results or inferences from the truth, or processes leading to such deviation" [11, p14]. This means the true value will be over- or under-estimated.
- **Case:** "In epidemiology, a person in the population or study group identified as having the particular disease, health disorder, or condition under investigation" [11, p21]. Therefore a case is an individual who has the disease of interest in a study.
- **Case-control study:** "The observational epidemiological study of persons with a disease (or other outcome variable) of interest and a suitable control (comparison, reference) group of persons without the disease. The relationship of an attribute to the disease is examined by comparing the diseased and non-diseased with regard to how frequently the attribute is present or, if quantitative, the levels of the attribute, in each of the groups" [11, p22].

Therefore case-control studies compare the characteristics of those with and without the disease of interest to try to establish differences between these groups of individuals. Case-control studies are described in further detail in §2.2.

- **Causal diagram:** A graphical display of the causal relations among variables, in which each variable is assigned a fixed location in the graph (called a *node* or *vertex*) and in which each direct causal effect of one variable on another is represented by an arrow (called an *edge*) with its tail at the cause and its head at the effect.[12]
- **Cohort study:** Is a longitudinal study design which follows individuals over a period of time (often years) to compare the occurrence of disease in a group of individuals who were and who were not exposed to a risk factor of interest.[13]
- **Confounding:** The effect of an extraneous variable which wholly or partially accounts for the apparent effect of the exposure, or which masks an underlying real association.[14] Therefore, an apparent association between the exposure and disease may actually be due to another variable. Alternatively, the apparent lack of an association may be the consequence from failing to control for the effect of another factor.[14] A confounding variable is a variable which is:
  - A cause (or surrogate) of the disease,
  - Correlated with the exposure,
  - Not caused by the exposure.[15]

  Confounding is a result of the complex relationships acting between the exposure and the disease, which can result in the relationship being over- or underestimated and it can even cause the direction of the effect to be reversed.[16]
- **Confounding bias:** "Distortion of the estimated effect of an exposure on an outcome, caused by the presence of an extraneous factor associated both with the exposure and the outcome" [11, p37]. Therefore confounding bias is the over- or under-estimation of the true effect, resulting from a confounder in the data. Further details are provided in §2.2.6.
- **Control:** "As used in...*case-control study*..., *control* means person(s) in a comparison group that differs, in disease experience...from the subjects of the study" [11, p40]. Hence a control is an individual who does not have the disease of interest, who is used for comparison with the case(s). The control may have other diseases which are not of interest in the given study.

- **Missing data:** Are where some data values in a dataset are unavailable for some variables or some individuals. This missingness could result in bias, including participation bias or selection bias. Further details are given in §2.3.7.
- **Participant:** "Person upon whom research is conducted" [11, p132]. Therefore the participant is an individual who is included in the study and providing data for analysis.
- **Participation bias:** "(Non)participation bias refers to the systematic errors introduced in the study when reasons for study participation are associated with the epidemiologic area of interest".[17] Hence it is a type of systematic error in observational studies when the participants are not representative of the population, which results when the participation rates vary between different groups of individuals. Participation bias is a subset of selection bias. Some authors use the terms 'selection bias' and 'participation bias' interchangeably. This may be due to uncertainty of the difference between the two forms of bias, since they are so similar, but it can cause confusion for the reader. The definitions also differ slightly between fields, for example econometricians frequently use 'selection bias' to refer to any form of bias resulting from selection or confounding.[18] Participation bias is described in further detail in §2.3.
- **Randomised controlled trial (RCT):** Is a study design in which individuals are randomly allocated to one of at least two groups to test a treatment of interest. One group receives this new treatment, while the other group(s) receive an alternative or no treatment.[19] The groups are followed through time and compared to try to determine the effectiveness of the treatment of interest.
- **Retrospective study:** "A research design that is used to test etiologic hypotheses in which inferences about exposure to the putative causal factor(s) are derived from data relating to characteristics of the persons under study or to events or experiences in their past" [11, p159]. Therefore a retrospective study uses data about the individual or their past, rather than data regarding future events.
- **Selection bias:** The error due to systematic differences in characteristics between those who take part in a study and those who do not. Selection bias invalidates conclusions and generalisations that might otherwise be drawn from studies. It is a frequent and commonly overlooked problem.[11]

- **Study population:** Also known as the *sample*, it is a selected subset of the population, which may be random or non-random and representative or non-representative.[11]
- **Target population:** The target population, also know as the *base population* or *study base*, is "the collection of individuals, items, measurements, etc., about which inferences are desired. The term is sometimes used to indicate the population from which a sample is drawn and sometimes to denote any "reference" population about which inferences are required" [11, p178]. In case-control studies, this is the group of individuals who would be cases if they were to develop the disease of interest.

2.1.1 Definitions for Graphical Models

Graphical models will be used through the thesis and have their own associated terminology. Definitions are provided below, which are applicable to most graphical models, including directed acyclic graphs (DAGs) which will be used in §2.3.4.1 and later in Chapters 4–6. The definitions below are taken from both the causal graph or DAG,[20] and chain event graph[21] literature (see Chapters 5 and 6).

- **Vertex:** (or node) A point used to represent a variable.
- **Edge:** (or arc or line) A line connecting vertices. May also be referred to as undirected.
- **Path:** A set of vertices with an edge between one vertex and the next.
- **Directed path:** Where the edges in a path have arrows leading from one vertex to the next, representing causality. The variable at the arrow head is caused by the variable at the tail. Bi-directed arrows have an arrow head at each end.
- **Collider:** A vertex on a path with both arrows pointing towards it.
- **Parent:** The vertex from which a directed arrow originates.
- **Child:** The vertex to which a directed arrow points.
- **Root vertex:** A vertex with no parents.
- **Leaf:** A vertex with no children.
- **Situation:** Not a leaf.
- **Directed graph:** A graph where all edges are directed.
- **Cycle:** A directed path which begins and ends with the same vertex.

- **Tree:** A vertex and edge set, and is a connected directed graph which has no cycles, one root vertex and all other vertices have exactly one parent.
- **Subtree:** A tree which is the child of a vertex, and a *floret* is a subtree consisting of a vertex set and its children, plus the associated edges.
- **Directed acyclic graph (DAG):** A graph where all edges are directed and there are no cycles.
- **Descendants (of a vertex):** The set of vertices with a path leaving the vertex.
- **Non-descendants (of a vertex):** All vertices of a graph which are not descendants of the vertex.
- **Predecessors (of a vertex):** The vertices which come before the vertex.
- A $\boxed{\text{box}}$ around a vertex means that the variable represented by the vertex has been conditioned on, which will be explained further in §4.3.

2.2 Case-Control Studies

A case-control study is an observational study design, which is primarily used to compare the personal characteristics and exposures of individuals with the disease of interest against those without the disease. They are retrospective studies meaning they use data from participants about their past, usually through means such as interviews, questionnaires and medical records. The disease and exposure of interest need to be carefully specified, for example, the exposure would need to be defined by intensity, duration and total dose, while the disease would need to be specified by symptoms, signs or laboratory findings.[14] Case-control studies are different from other types of study designs, as the sampling relies on the disease status rather than the exposure status.[22] Modern case-control studies evolved in around 1926 when Lane-Claypon investigated the role of reproductive experiences in the etiology of breast cancer.[23] Case-control studies offered a solution to studying diseases with long latency periods, saving both time and the number of required participants.[23] These studies are therefore time efficient, statistically efficient and also generally easy to analyse[24] (see §2.2.5). A key event for case-control studies was the demonstration that they can be used to estimate relative risks,[25] in the same way as cohort studies are able to. Since a range of different exposures can be investigated from the participants, case-control studies can be very useful when little is known about the disease being considered.

The general set up for a case-control study is as follows; there is a population, which consists of all those who are at risk and from this there are two groups recruited; those with the disease of interest (cases) and those without the disease of interest (controls). The entire sampling procedure should result in an "unbiased ascertainment of eligible cases and controls and a procedure for the selection of a sample from them in a manner that assures that each individual has an equal chance of appearing in the study" [14, p80]. The exposure of interest is then investigated by comparing the levels of exposure for individuals in each of the two study groups. This information can be used to assess whether the exposure may be associated with the disease. However, the validity and the generalisability of the results depends on aspects such as how the cases and controls are identified and how they are recruited. Since the study design is retrospective, the findings usually relate to associations rather than causal conclusions, since the ordering of events is often unclear.

Case-control studies are observational, usually reasonably low cost, quick and easy to conduct. The main difficulties with this study design arise through trying to minimise the various forms of bias which can occur, such as selection bias, misclassification and recall bias.[22] However, case-control studies are still widely used to study rare diseases as they require considerably smaller sample sizes than the corresponding longitudinal or cross-sectional studies.[26,27] The most controversial part of a case-control study and the main source of criticism is how the controls are recruited. Therefore they are most appropriate for detecting large effects, which appear genuine rather than a consequence of the study design.[22] Case-control studies are, like all study types, not suitable when the disease affects the exposure as well as being a result of it, for example when poor nutritional status is considered as an exposure for diarrhoea,[22] which is an example of reverse causality.[28]

2.2.1 Methods for Selecting the Participants

There are several different methods available for selecting cases and controls. *Random sampling* involves selecting individuals such that each sample has a fixed and determined probability of selection.[14] *EPSEM (Equal Probability of SElection Method)* is where each of the population individuals has an equal probability of selection; this is more commonly used than random sampling.[14] *Systematic sampling* refers to the sequential selection of individuals, who are conceptually separated on lists by an interval of selection. For example, the researcher would select

every kth eligible individual for the study.[14] In *stratified sampling* the individuals are selected at random from defined subgroups, or strata, of the target population. The samples from each subgroup need not be the same size.[14] Finally, *matched sampling* involves the pairing of at least one control to each case on the basis of certain characteristics, while retaining some randomness. The effects of the matched characteristics are consequently eliminated from the study.[14]

2.2.2 Selection of the Cases

Case-control studies are most valuable if the participants are selected in a suitable, unbiased way. Ideally, all the cases possessing the disease of interest within the target population should be included in the study, but this is unlikely to be possible in practice. However, to generate reliable results this would not be required, just a sufficient quantity of cases to be representative of the population.

The first requirement is the definition of the disease of interest. For example, the disease could be specified as 'cancer', 'liver cancer', or 'stage II liver cancer'; this must be decided before recruiting cases. The method of diagnosis should also be decided. For example, whether self-diagnosis of asthma would be sufficient, or whether a general practitioner (GP) diagnosis would be required. If the disease is difficult to diagnose, the study may contain false positives where individuals are mistakenly diagnosed, which should be considered during the analysis of the study.[14]

Next, the population of interest must be decided, which can be a hospital or geographical area, and the time point chosen, whether it be a period of time or a particular time point.[16] Bias can arise from using just one hospital due to similarities of individuals within the local area, but this is attractive due to its simplicity. The study population, are selected in such a way as to represent a larger target population which the study aims to draw conclusions about, hence the choice of data source is important. It is proposed that validity is more important that generalisability in case-control studies, since results which are not valid will "preclude any ability to generalise the results",[16] so it is suggested that only cases who have complete and reliable information available are chosen, rather than those which may be more representative of the general population.[16] However, this is an opinion which is highly controversial as it can introduce bias, since those with complete information may differ from those without.

The source of the cases is likely to depend on how, and where, the disease of interest is recorded. For example, if the disease requires hospital admission, then the cases may be recruited using hospital inpatient lists. Diseases requiring hospital admission could result in the sudden death of some individuals before treatment, so the study results must either be applicable only to treatable cases, or the analysis must consider those who died before treatment. Another source for recruiting cases is GP lists, where the diseases may be initially recorded. Other sources include NHS Digital,[29] Hospital Episode Statistics (HES),[30] or The Health Improvement Network (THIN) database.[31] Whatever the source for recruiting cases, it must be one through which the majority of the cases can be identified.

Generally, it is more useful to recruit incident cases, who are recently diagnosed, rather than prevalent cases, who have been diagnosed for a longer period of time,[32] since prevalent cases are survivors forming part of a larger previous group. Prevalent cases may therefore be receiving treatment and so not represent those currently without treatment, meaning their results may only generalise to a limited group of individuals. Other factors to consider include the stage of the disease or disease presentation, if relevant.

2.2.3 Selection of the Controls

Controls are used to determine the frequency with which the exposure occurs in individuals who do not have the disease of interest. This can then be used as a comparison to the case frequency, so the controls must be representative of the source (or base or target) population from which the cases were recruited.[16] To achieve this, the controls should be selected randomly and independent of their exposure status, from the same population as the cases.[27,33] The controls must not have the disease of interest when the study begins and should only be excluded during the study if they do not satisfy the required criteria. If a control develops the disease of interest during the study, they must be treated as a case. It is possible that controls may become cases at some point in the future after the study has been completed, but since the diseases considered in case-control studies are rare, this is unlikely and hence would not have a drastic effect on the results.

Some sources for control recruitment include;

- GP lists; the same GP lists from which the cases were recruited.

- Colleagues; if the disease is being investigated through a place of work.
- Electoral lists; the same electoral lists from which the cases were recruited.
- Hospital lists; patients from the same hospital as the cases, usually with a condition which is completely unrelated to the disease of interest. However, these controls may not be representative of the target population,[27] since there are groups of individuals who are more likely to be admitted to hospital, for example, those of a lower socio-economic status (SES), introducing bias relating to social class.
- Random selection from the same area as the cases; random digit dialing was a method used to find controls.
- Friends, neighbours, spouses, relatives.

Controls can be selected in two different ways; either as a group of controls or as a control to match a particular case. Matched case-control studies are when the control is chosen to be similar to the case with regard to as many of their characteristics as possible; usually considered confounders, including variables such as age, location and occupation. Alternatively, controls can be recruited for each case, but without matching any characteristics. If a matched case-control study is used, no information will be obtained on the matched variable(s), so it is sensible to only match on variables which are known to be unassociated with the disease of interest, or for confounders, if the effects are already well documented.

The number of control groups can vary. Ideally, there should be a single control group which is the most comparable group to the cases,[16] but it can be difficult to find this ideal control group. Instead, it is common practice to recruit more than one control group and compare the results. If the exposure of the cases differs to that of all the different control groups, the theory behind the exposure of interest is strengthened.[16] It is also possible to have more than one disease group, depending on the disease of interest. Another approach is to have a ratio of cases to controls.

The ratio of cases:controls can vary from study to study. When the number of available cases and controls is large, and the cost of recruiting both groups is comparable, the optimal ratio is 1:1. However, when the sample size of cases is limited, due to the disease being rare, or when the cost of recruiting cases is greater than controls, it is common for there to be a ratio of 2:1 for controls:cases.[16] As the number of controls per case increases, the power of the study increases, but it is not recommended to have more than a 4:1 ratio, since not much statistical power is gained

beyond this point.[34]

The selection approach used can vary depending on the nature of the study. For example, a recent publication[35] investigated the recruitment of controls into case-control studies as part of the study of infectious disease outbreaks, and found that neighbourhood controls were the most frequently used, and face-to-face interviews were the most common method of data collection. While these approaches are likely to be convenient and provide the timely responses required under the circumstances, they may not result in a randomly-selected representative sample which is required generally for a case-control study.

2.2.4 Data Collection

There options available for ways to collect data from cases and controls, including interviews, questionnaires, official data records or a combination of these sources. To prevent bias from occurring, it is strongly recommended that the same method is used for collecting case and control data, but this is often unfeasible due to differences in the data availability for each disease group.

Whichever source is chosen, the reliability and accuracy of the data should be assessed. Hospital or electoral records are common sources, but have the potential to be out of date or incomplete. Interviews or questionnaires ensure relevant data are sought, but the questions and questioning method should be as similar as possible between the cases and the controls to reduce bias, including the interview location, time and interviewer. It is also useful for the data collector to be blinded to the disease status of the individual and the hypothesis.[16] This interview approach may be less helpful if an individual is too ill to communicate, or if they have a disease which may affect their ability to recall data, such as dementia.

The value of the recorded data can vary; in a case-control study, exposure data which are recorded before the individual is diagnosed are particularly useful.[16] For example, if birth weight is the exposure of interest (recorded accurately for each individual *regardless of their disease status*), then it is more valuable than data recorded *after* assignment of the disease status. Hence this is an example of where data are collected before diagnosis. In other instances, the accuracy or completeness of the data may depend upon the individual's disease status, which is less desirable.

Prior knowledge about the exposure and disease of interest, including their mechanisms, will help

to determine which data should be collected[16] and help to form more specific questions regarding exposure, for example smoking habits *ever*, or smoking habits during the *last five years*.

2.2.5 Results of a Case-Control Study

The participants of a case-control study are selected because they either do or do not have the disease of interest. Once recruited, they are subdivided into groups which were exposed or not, as shown in Table 2.1.

	Case	Control	Total
Exposed	a	b	a+b
Not Exposed	c	d	c+d
Total	a+c	b+d	a+b+c+d

Table 2.1: Sample groups used to calculate odds ratios in a case-control study.

	Case	Control	Total
Exposed	a′	b′	a′+b′
Not Exposed	c′	d′	c′+d′
Total	a′+c′	b′+d′	a′+b′+c′+d′

Table 2.2: Unknown population groups represented by a case-control study.

The total numbers of exposed and unexposed cases in the population cannot be derived from Table 2.1, since a case-control study does not use a specific proportion of either category, and the individuals are not randomly sampled from the population, but instead are sampled dependent upon their disease status. Therefore, the incidence rate in the population and the incidence rate in those exposed cannot be derived. It follows from this that the relative risk (RR), usually calculated from the data, cannot be found. The relative risk is the risk of developing the disease relative to the exposure. It is calculated as a ratio of the risk in the exposed group over the risk in the non-exposed group. The true relative risk from Table 2.2 is therefore:

$$\frac{a' \times (c' + d')}{c' \times (a' + b')}.$$

Case-control studies instead use the odds ratio (OR) to approximate the relative risk. Usually, the proportion of the population who are cases is small; a rare disease. Therefore, a' will be small in relation to b', and c' will be small in relation to d' [32, p40]. It follows that d' approximates $c' + d'$ and b' approximates $a' + b'$, so the OR and RR are very similar when the disease is rare.

Odds Ratio, OR = $\frac{a'/c'}{b'/d'} = \frac{a'd'}{b'c'}$.

Relative Risk, RR = $\frac{a'/(a'+b')}{c'/(c'+d')} \approx \frac{a'd'}{b'c'}$, when the disease of interest is rare.

However, when the disease of interest is common, the meaning of the OR will depend upon the sampling scheme used for the controls.[36] The usual choice is to select controls from those who are still without the disease of interest at the end of the study. Therefore, any controls who develop the disease of interest during the study, are treated as cases. In this instance; OR = $\frac{a'd'}{b'c'}; 0 < |\text{RR}| < |\text{OR}|$.[22] This can be demonstrated using the hypothetical data in Tables 2.3–2.5.

	Case	Control	Total
Exposed	10	90	100
Not Exposed	5	95	100
Total	15	185	200

Table 2.3: Odds ratio and relative risk: Rare disease, as common in case-control studies.

The OR using Table 2.3 is $\frac{10\times95}{90\times5} = 2.1$ and the RR is $\frac{10(5+95)}{5(10+90)} = 2$, which is very close as expected when considering a rare disease. Table 2.4, with a less rare disease, gives an OR of $\frac{25\times88}{75\times12} = 2.4$ and a RR of $\frac{25(12+88)}{12(25+75)} = 2.1$, which is slightly different as expected when considering a less rare disease. Table 2.5 results in an OR of $\frac{40\times80}{60\times20} = 2.7$ and a RR of $\frac{40(20+80)}{20(40+60)} = 2$, showing how the difference between the RR and OR estimates increases as the disease becomes more common.

	Case	Control	Total
Exposed	25	75	100
Not Exposed	12	88	100
Total	37	163	200

Table 2.4: Odds ratio and relative risk: Less rare disease, not usually found in case-control studies.

Example 2.2.1 *This example uses a hypothetical dataset, where the exposure of interest is coffee consumption and the disease of interest is cancer. In a real study, the location and type of the*

	Case	Control	Total
Exposed	40	60	100
Not Exposed	20	80	100
Total	60	140	200

Table 2.5: Odds ratio and relative risk: Common disease, not usually found in case-control studies.

cancer would need to be specified along with the criteria to be categorised as a coffee drinker, including the frequency, number of cups and whether decaffeinated coffee is included.

	Case	Control	Total
Coffee Drinker	150	250	400
Not a Coffee Drinker	20	90	110
Total	170	340	510

Table 2.6: Case-control study example: Coffee intake and cancer.

Odds Ratio, OR = $\frac{150/20}{250/90} = \frac{150\times90}{250\times20} = 2.7$.

The odds ratio of 2.7 suggests that those who consume coffee are 2.7 times more likely to develop cancer than those who do not consume coffee.

In Example 2.2.1 the individuals were categorised as 'coffee drinker' or 'not a coffee drinker' and the approach will not be suitable for continuous exposures (unless dichotomised). There follows an example with a hypothetical dataset showing how to calculate the OR when the exposure is continuous, which is also applicable to exposures with multiple categories.

Example 2.2.2 *Let the exposure be blood pressure (BP) and the outcome be stroke. The blood pressure variable could be dichotomised (low/high), but in this example it will remain continuous to retain as much information as possible about the exposure. A table similar to Table 2.6 cannot be produced for the odds ratio to be calculated. Instead, a logistic regression model can be formed and the odds ratio calculated as the exponential of the regression coefficient for the exposure variable. There are 15 individuals, 4 of which are cases (stroke=1), shown in Table 2.7.*

The corresponding model is as follows, (where 'Stroke' is in fact the log odds of stroke)

$$\textit{Stroke} = \text{-}18.07 + (0.13 \times \textit{BloodPressure}).$$

BP	120	119	134	141	122	125	119	137	139	117	121	121	130	145	133
Stroke	0	0	1	1	0	0	0	0	0	0	0	1	0	1	0

Table 2.7: Hypothetical raw data: Blood pressure and stroke.

The odds ratio is then,

$$e^{0.13} = 1.14(2dp).$$

The 95% *confidence interval for the odds ratio can also be calculated;* $(0.98, 1.33)$.

The odds ratio of 1.14 suggests that for a one unit increase in blood pressure, the individual is 1.14 times more likely to have a stroke, hence a 14% *increase in the odds. This increase does not depend on the original blood pressure reading, only the unit increase. However, as the confidence interval for this odds ratio contains the value of one which indicates no association, it is possible there is no association between blood pressure and stroke.*

2.2.6 Confounding Variables in Case-Control Studies

A confounding variable is associated with both the exposure and disease of interest; causing the disease, but not on the causal path from the exposure.[37] DAGs are useful for identifying bias[20] and determining whether or not a variable is a confounder, since confounders are a common cause of the exposure and disease.[18] If a variable is predictive of both the disease and the exposure, then it should be incorporated into the analysis in an appropriate manner.[38] A classic example is if there is a study which is investigating the association between alcohol consumption (exposure) and heart disease (disease), a confounding variable may be smoking. This is because there is likely to be a greater proportion of smokers amongst the cases than the controls, since smoking is often correlated with heart disease, and smoking can be associated with alcohol consumption. Therefore, the apparent increased risk of heart disease which is associated with alcohol consumption, may in fact be the result of smoking.[32] A variable can only confound an association if it differs between the case group and the control group.[16] For further details on confounders, see §2.1.

One main concern in case-control studies, as well as most other study designs, is that confounding may be occurring which is not identified. If known, the confounding variables can be accounted for in the study design. For example, if age is a confounder, equal numbers of cases and controls for each age group could be selected (stratum matching). Alternatively, a matched design could

be used whereby the cases and controls are matched by age.[22] If the confounding is not accounted for in the study design, it may be considered during the analysis stage. Two proposed methods for dealing with confounding variables during analysis are as follows:

1. Analyse subsets of the data, defined by the confounding variable. For example if age is a confounder, the age of the individual could be restricted, and the effects calculated for each age group, resulting in estimates unconfounded with regard to age. The results can then be displayed as a set, with one result plus corresponding confidence interval for each age stratum, or combined to form one overall unconfounded estimate. One way to pool results is using the Mantel-Haenszel method;[39]

$$\text{OR}_{MH} = \frac{\sum ad/n}{\sum bc/n}, \tag{2.1}$$

where a, b, c, d are as specified in Table 2.1, where each table contains one strata, and $n = a + b + c + d$. The suitability of this approach will depend upon the data.[39]

2. Use an analysis technique to adjust for the effects of the confounding variable(s), by constructing a carefully chosen mathematical model. A logistic regression model is often used, since the disease status is usually binary and as it allows for several confounding variables to be adjusted for at once.[32] The risk of disease is expressed as a function of independent predictor variables,[16] in fact the dependent variable is defined to be the natural logarithm (ln) of the odds of disease, the logit. Let π_d be the probability of the disease, then $\frac{\pi_d}{(1-\pi_d)}$ is the odds of developing the disease of interest and the log odds of disease, or the logit, is written as $\ln\left[\frac{\pi_d}{1-\pi_d}\right]$.[16] The log odds of the disease is then:

$$\ln\left[\frac{\pi_d}{1-\pi_d}\right] = B_0 + B_1X_1 + \cdots + B_nX_n,$$

where the X_i are the recorded values for each individual of the independent variables and the B_i are the corresponding coefficients for each of the independent variables. This can be rewritten as:

$$\pi_d = \frac{1}{1 + e^{-(B_0+B_1X_1+\ldots+B_nX_n)}},$$

to represent the probability of disease.[16] The coefficients of the logistic regression model denote the magnitude of the increase or decrease in the log odds produced by a one unit change in the value of the independent variable. The coefficients therefore show the effect of an individual variable on the log odds of the disease while keeping all other variables

constant, and this model can also be adapted to include interaction terms if required.[16] The magnitude of the overall confounding is estimated by comparing the crude estimates, with the adjusted estimates. In case-control studies, this would involve comparing the two estimated relative risks.[16]

2.2.7 Summary of Case-Control Studies

Case-control studies are useful when the disease of interest is rare, and in this instance the OR is similar to the RR. Since the number of cases is known before the study commences, a significant result can be obtained with relatively small participant numbers. Compare this with a cohort study for a rare disease where the number of cases is unknown at the beginning of the study and hence a larger sample size is required to ensure an adequate number have the disease of interest by the end. The results from a case-control study can be generated relatively quickly since there is no need to wait for the disease of interest to develop as in a cohort study, resulting in case-control studies being relatively cheap to perform.[32]

Case-control studies are retrospective, relying on data which can be susceptible to problems such as recall bias. It may be that individuals cannot remember specific details or that the data records used are inaccurate or outdated. Problems associated with causality can also occur from retrospective data, since it is usually not proven that the exposure preceded the disease and therefore may have caused it. Finally, it has been well documented that controls can be difficult to recruit for case-control studies and this can affect the validity and generalisability of the results.[32] Further recent details of case-control studies are available.[40]

2.3 Participation Bias

The two most common sources of bias in causal inference studies are confounding bias and selection bias[41] (of which participation bias is a subset). However, before considering participation bias, one idea which is often linked is that of participation rates, but how are they calculated?

2.3.1 How to Calculate Participation Rates

Participation rates for epidemiological studies are rarely 100% and have been declining during the last 30 years or more,[27,42] as documented by academic researchers, governmental agencies and non-profit companies.[17] For example, cooperation rates;[43]

$$\frac{\text{participants interviewed}}{\text{number of eligible individuals contacted}},$$

in published population-based case-control studies declined by 3.33% per year in cases and 5.15% per year in controls from 1991 to 2003.[44] It is not uncommon to have cooperation rates of around 70% and response rates;

$$\frac{\text{interviewed}}{\text{interviewed+eligible non-participants+individuals of presumed but unconfirmed eligibility}},$$

of around 50%.[45] Some studies have managed to maintain relatively high participation rates over time, however even these studies have suffered from increased refusal rates and hence have only maintained high participation rates by increasing their effort to recruit the hard-to-reach individuals.[46]

For any study it is useful to assess the participation rates to determine how they differ between study groups. A number of different formulae for calculating participation rates have been proposed. In fact, it has been stated that "there are so many ways of calculating response rates that comparisons across surveys are fraught with misinterpretations".[47] However, generally the participation rate is considered to be the number who participated divided by the number who were eligible to participate; where the number eligible can be difficult to determine before the initial interview. Examples for how to calculate participation rates are shown in Equations (2.2) and (2.3), however these equations can be altered by a research group to 'improve' the study rates and hence make the study results appear more reliable.

Case Participation Rate:

$$\frac{\text{interviewed cases}}{\text{interviewed cases} + \text{case refusals} + \text{surrogate refusals} + \text{consultant refusals}} \times 100. \qquad (2.2)$$

Control Participation Rate:

$$\frac{\text{interviewed controls}}{\text{interviewed controls} + \text{control refusals} + \text{GP refusals} + \text{non-interviewable}} \times 100. \qquad (2.3)$$

The different refusals are defined as follows,

- *Case or control refusals:* When the initial potential participants refuse to take part.
- *Surrogate refusals:* When the initial potential participants refuse to take part, surrogate potential participants can be selected. Surrogate refusals are when these individuals also refuse to participate. (Note this differs from where surrogate is sometimes used to describe an individual who provides data on behalf of someone who is unable to, such as for a child, someone who is critically ill, or someone who has died).
- *Consultant/GP refusals:* When the consultant for the case, or the GP for the control does not allow their patient to be approached by the study group. This may be because the consultant believes the case is too ill, stressed or busy, or the GP has additional information for why they believe the control should not be asked to participate.

There are different ways in which the participation rates can be compared, depending upon the research aim. Examples include; overall, or by sex, age group, study region, social class estimators or deprivation indices.

It is useful for a study to provide a table or diagram showing the numbers of individuals who participated or refused, so statistics such as the participation rates can be calculated if not included. Comparisons can then be made between the participating and non-participating groups (cases/cases, controls/controls and cases/controls), to determine how the groups differ. These comparisons can help to decide whether the cases and controls represent the same target population, or whether they differ due to their sources or the motivation to participate between the groups. It can be difficult to compare participation rates between studies, since there are often different recruitment methods used and the individuals need to satisfy different criteria. Also, unfortunately some studies do not record enough information to allow statistics such as participation rates to be calculated.[44]

Alternatives to participation rate calculations have been suggested.[17] The American Association for Public Opinion Research (AAPOR) suggests that the *response rate* is the number of completed interviews, divided by the number of possible interviews, the *co-operation rate* is the number of cases interviewed, divided by the proportion of all eligible cases and lastly the *refusal rate* is the proportion of cases who either do not attend the interview or refused to be interviewed. However, there are still different ways in which to define the parts of these rates, and one document gives 6 response rate equations, 4 co-operation rate equations and 3 refusal rate equations.[43] Some

recommend reporting a variety of rates, along with the methods used to calculate them,[17] but space restrictions mean this is rarely practiced in epidemiological papers.

The confusion around which rate to use to assess participation and the choices of how to calculate that particular rate, may be reasons why participation rates from studies are often not reported.[45] An additional reason may be that authors do not wish for their findings to be dismissed if they reveal low participation rates. Lastly, some studies may not have recorded enough information to confidently report the participation rate.

It can be difficult to assess how the participation rate has affected the study results. It is not simply a matter of calculating the participation rate and using a cut-off value to determine whether the study will be 'good' or 'bad'.[17,27] The main concerns arise when the lack of participation leads to participation bias; the systematic error introduced into the study when the reasons for participation are associated with the epidemiologic area of interest.[17] It is therefore the differences between the participants and non-participants which determine the amount of bias.[48–50] Some studies with lower participation rates can result in less bias than those with higher participation rates, so a low participation rate does not necessarily indicate a poor study with a high level of bias,[50–52] although low rates can allow more opportunity for bias to occur.[27] However, sometimes this idea of low participation rates not leading to large amounts of bias is used in the discussion of a study without the actual amount of bias being assessed.[53] To encourage a participation bias analysis to be conducted, journals could insist all case-control studies detail a participation bias calculation, for judgment by the reader. Alternatively journals could adopt standardised formulae to calculate rates such as those proposed by The American Association for Public Opinion Research (AAPOR),[43] which would give guidance to researchers and allow easier comparison between studies. Similar advice is published.[27] Where participation bias may be a concern, methods such as those given in Chapter 4 should be considered.[27]

Differences between participants and non-participants can only usually be assessed if there is information available on the characteristics of those who did not participate and this information is often limited.[54] In some cases, such as when the study uses random-digit dialling, there is no means of identifying those who did not participate.[55] There have been efforts made to acquire this comparative information,[56–58] but there can still be constraints on being able to use this information within the study.[54] If at all possible, widely available information should be used,

such as census data or national databases,[59,60] which are virtually free from participation bias, since all households are required to partake. Some suggest that the profiling of non-participants may be just as important as that of the participants.[54]

There are currently different opinions on the effects of non-participation. Some consider the effects to be minimal, while others claim it can have dramatic effects on the results.[61] It would appear the effect of non-participation varies from study to study and depends upon the combination of a number of factors such as;

- How the participation rates differ between the case and control groups.
- What the participation rates are; high or low.
- How participation is related to the exposure and disease of interest.

2.3.2 Individual Characteristics Associated with Participation

2.3.2.1 Motivation to Participate

When conducting a study the success of recruiting, particularly a control, will depend partly on the willingness of the individual to be involved in medical research, which will rely upon their general attitudes towards research, any previous participant experiences, demographics, and the disease of interest. The required effort to recruit controls into studies is thought to have increased substantially from 1991 to 2003,[46] including factors such as the number of times an individual needs to be contacted.

There are a number of ways in which an individual may not participate, including;

1. Actively refusing to participate.
2. Refusing to respond to the data collection method, such as a survey or telephone call.
3. Being uncontactable, for instance if the GP records are not up to date.

There are authors[17] who believe there are four main reasons why an individual would not participate in a study, these are;

1. The increasing number of studies being conducted over recent years. It is likely that each individual is facing an increased number of requests for participation into studies, including those conducted by academics, the government or medical companies. This increase may

cause an individual to refuse all the requests, since they view it as an invasion of their privacy or a burden on their personal life. They may also believe their input is less valuable, since so many requests are made and hence refuse. Telemarketing is thought to have affected the participation rates for epidemiological studies, since it can be hard to distinguish between sales and research when the telephone calls at home seem similar.[17]

2. The general decline in volunteering in the United States and other Western countries. Studies have shown that participation into research is related to participation in other events such as community organisations or activities.[62,63]
3. The personal circumstances of the individual regarding the disease or exposure of interest, or their ability to relate to it. For example, studies which have looked into the effects of mobile telephone use and cancer[64] or the effects of fried potato consumption and cancer,[65] have both stated participation rates of 90% or more. This uses the same reasoning as to why cases are more likely to participate in case-control studies than controls.
4. The perceived amount of time, data and commitment the study will require. This may also include whether the study has the potential to be painful, personal or to require confidential information.

2.3.2.2 Incentives

Incentives, such as cash, can affect participation. The National Survey on Drug Use and Health recorded participation rates as follows:[66]

- 69% for those who received no cash incentive,
- 79% for those who received $20 incentive,
- 83% for those who received $40 incentive.

However, the effect incentives have on participation is unclear. One investigation[67] into the use of incentives in surveys found larger cash incentives were more effective at recruiting those with a lower level of education or those with a lower income. It also found in other instances the use of cash incentives was more effective in recruiting those with a higher income and higher education level, who expected compensation for their time.

2.3.2.3 Demographics

In case-control studies, participation is also dependent upon the satisfactory fulfillment of the case or control criteria. After selection, the probability of participation is often higher for cases than controls, which may be for a number of reasons. Cases may be willing to help with any research into their disease in the hope that a cure can be found, or a cause of the disease identified to help to prevent others from suffering. Controls do not have this same motivation and may only be interested in participating if they have personal connections to the disease, such as a friend or relative who is a case.

Studies such as the UK Childhood Cancer Study (UKCCS) have shown there are certain groups of individuals who are more likely to participate as controls such as women, the employed, the educated and those who are married,[68] with lower response rates in more deprived areas.[69] Other studies have agreed that these factors are associated with participation;[17]

- **Sex:** It has been well documented that women are more likely to participate than men.[70–74]
- **Age:** Older individuals have been shown to be more likely to participate in some studies,[70,72,73,75] yet younger individuals have been more likely to participate in others.[63] This may depend upon the research topic.
- **Ethnicity:** White individuals are more likely to participate in some studies,[76] while black or minority groups are more likely to participate in others.[63,77] Again this may be due to the area of research, or possibly the study location.
- **Socio-Economic Status:** Those with high SES have been shown repeatedly to be more likely to participate.[70,74–76,78,79] SES can be measured as, deprivation,[68] housing,[58,80–82] income,[81–83] education,[56–58,81,82,84–87] or occupation.[57,58,80–82,88]
- **Education:** Those more highly educated are more likely to participate.[63,76]
- **Employment:** Those who are employed are more likely to participate.[72,76,78]
- **Marital Status:** Those who are married are more likely to participate.[76,78]

It is also believed that cases are more likely to participate if they are at an earlier stage of their disease and have a higher chance of survival, which may be due to different attitudes at earlier stages or that later stages prevent individuals from being healthy enough to participate. These characteristics could be used to form ideas about the types of individuals who are more likely to participate and their other commitments such as employment or family, and possibly assumptions

about when would be most suitable to contact them and by which means. These ideas could then be used to try to increase recruitment in future studies. Most of the characteristics here are taken from a review article in 2007, but a more recent review will be conducted in §2.3.5, formed using publications from 2007–2015.

2.3.2.4 Study Area and Requirements

There can be increased difficulty in recruiting cases or controls for exposures which are deemed to be negative in some way. For example, those who smoke or consume alcohol may be less willing to participate due to the stigma attached to their lifestyle. This may also apply to cases with a less socially acceptable disease of interest, such as an eating disorder or a sexually transmitted disease.[17,89,90] The current circumstances of the potential participant could also affect their likelihood of participating. An individual in an abusive relationship may be less willing to participate in a study regarding domestic violence if they are still living with their abusive partner.[91] However, they may be more willing if they have started a new life away from that partner.

The requirements of the study can play an important role in determining the participation rates. As expected, studies which require more time, or which involve more invasive procedures such as giving blood, or personal questions, are likely to have lower participation rates.[49,92] Factors which affect participation can also have different effects on cases and controls. For example, a variable which may affect participation for cases, may not for controls. Those variables which do affect participation for both cases and controls are unlikely to affect both groups to the same extent. It may be that the variable may increase participation in one group, while decreasing it in the other.

2.3.3 Study Factors Associated with Non-Participation

While some areas of epidemiology are flourishing due to the wider availability of registers and databases, others are suffering due to the decline of participation into studies which require interviews, questionnaires, or biological samples.[45] In 2006, an assessment of the state of response rates was conducted, by analysing epidemiology studies which were published in the period leading up to 2003 in ten high-impact journals in epidemiology.[45] It was found that only

41% of cross-sectional studies, 56% of case-control studies and 68% of cohort studies provided information regarding response or participation rates. The publications analysed included studies from 1970 to 2002, and in this time it was seen that participation rates decreased for all study types and most dramatically in control groups. It stated that participation rates of 70% and response rates of 50% were no longer uncommon and that it must be considered how much the response rates affect the study results, if at all. The following examples were given;

- A 70% response rate in cases and 60% response rate in controls could result in very little bias for one exposure, while creating large bias in another exposure within the same study.
- Equal response rates in the cases and the controls may not be of any benefit if the non-participation differs between the two groups for the different variables, as is often found.
- Also, rather high non-participation rates may not be of any concern if the reason for non-participation is unrelated to the exposure. Even when the reason for non-participation is related to the exposure, this will not result in bias unless the participation regarding exposure differs between the two groups.

Non-participation can be associated with different stages in a study and some examples applicable to case-control studies follow. Many of these examples are taken from a review which was published in 2007, but a more recent review will be conducted in §2.3.5, formed using findings from 2007–2015. This new review will also report on up-to-date study design factors thought to affect participation in §2.3.5.4.

2.3.3.1 Study Design

Particular features of a study design may affect participation, for example when the study is retrospective. In this instance, the exposure and disease have already occurred, which may affect the motivation of the (non-)diseased and (non-)exposed. The sensitivity of the data and need for each participant to give active consent are also features which may affect participation.

It may be possible to retrieve information about those who have chosen not to participate, but often there will not be sufficient time or funds for this and even when there are, it may be difficult data to record accurately. This can subsequently prevent a comparison between participants and non-participants.

2.3.3.2 Selection Process

Ideally, the potential participants for a study should be chosen randomly from a well-defined population over a specific time interval to avoid selection bias, and then each of the randomly selected individuals would participate to avoid participation bias. In practice there are restrictions such as ethical constraints, which can affect the selection process and even when a carefully planned selection process is used, there is no guarantee that all individuals will participate. In addition to the decreasing participation rates noted by several authors in the 2007 review,[17] studies are generally looking for smaller effect sizes,[93] meaning the problems associated with participation are increasingly important.

In case-control studies, cases may be selected using hospital lists or disease registers, whereas controls may be selected using birth registers. However, the two groups are expected to represent the same population. In case-control studies, the cases and controls must be sampled from the same population.[94] It may therefore be more suitable to instead recruit the controls from the same hospital as the cases, but only from patients who have a disease unrelated to the study disease. However, this may result in the participants having similarities related to their hospital, such as a similar socio-economic background caused by a similar residential location and therefore they may not represent the general public due to their hospital attendance. For practicality, the controls are sometimes recruited from GP lists, which are thought to cover around 98% of the population,[95] which is high, but still excludes some individuals.

Non-participation may be due to not being involved and active. For example, if the study requires volunteers, non-participation occurs when an individual does not step-forward. Studies have shown that, as how participants differ from non-participants with the characteristics in §2.3.2, volunteers differ from non-volunteers. Another selection method is a postal or online questionnaire or survey, where some individuals may simply not respond. Again, those who choose to respond are likely to be different in some respects to those who do not.

The study criteria for participants can also result in non-participation. For example, if those at the later stages of a disease are excluded as they are deemed too ill to participate, this could result in willing cases being excluded. However caution should be taken with these criteria, since the study results will then not be generalisable to those who are most ill.

2.3.3.3 Method of Data Collection

Some methods used for recruiting individuals have been found to be more successful than others. Generally, studies which use face-to-face recruitment have higher participation rates than those using telephones or letters,[96] which may be due to the individual feeling more obliged to participate when asked face-to-face. In case-control studies, since the disease status is known prior to the study, it is common for the cases and controls to be recruited using different methods, which can therefore affect participation.

There have been studies into the methods used to collect data from participants and non-/partial-participants; with different methods often used for the two groups. Studies frequently use face-to-face interviews for participants but telephone interviews for obtaining basic background information from non-participants.[57,58,81,86,97] This may determine whether or not an individual participates at either level. If the study centre is far from the individual's home or only open when the individual is working, they may be more inclined to 'participate' as a non-participant rather than a full participant.

The different methods used for the two groups may introduce information bias, such as misclassification bias, interviewer bias, recall bias or reporting bias. One solution to reduce this bias may be to use information which is freely available to compare the participants with the non-participants, such as electoral registers, basic medical records, cancer registries or census data. This may be preferable since the information for both groups of individuals is obtained from the same source and hence reduces problems associated with information bias. Limitations of this approach include the availability of data and the possibility of inaccurate or out of date information.

There can be problems even when a suitable recruitment method is used. If telephones are used to recruit cases and controls, the recent increase in mobile telephones as well as some households having multiple landlines or unlisted telephone numbers, all contribute to the difficulty in contacting suitable participants.[17] Lifestyle changes, such as increased working hours and more women in work, have also made it more difficult to contact individuals at home, resulting in either alternative individuals being contacted or more attempts being required.[17]

Two main factors should be considered when planning the data collection method. Firstly, the

way in which the data will be collected; for example a questionnaire, interview or examination. Secondly, the individual collecting the data; whether they are a trained interviewer or not, and their awareness of the exposure or disease status of the participant. It is preferable for the individual collecting the data to be unaware of the disease status of the participant (blinded), so they do not ask questions regarding the exposure in a different way for cases and controls. The participant should be unaware of the exposure of interest if possible, and the questions regarding exposure should be masked by questions regarding other exposures. Whatever decisions are made, it is crucial the data for both cases and controls are collected in the same way.

Some studies,[98–101] have adopted more than one option for data collection in an attempt to increase participation rates. For example, some studies give the participant the option of completing their questionnaire by post or using the Internet.[98] Other studies use different data collection methods for those who did not respond to the first method.[99–101] However, these studies could raise concerns regarding differences in the participants who respond using varying methods.[17] Care should be taken when trying to avoid non-participation during data collection, as extreme efforts to recruit unwilling participants in order to improve the participation rates may be viewed as unethical.[17]

2.3.4 When Does Non-Participation Result in Bias?

Non-participation *can* lead to participation bias, but bias does not *always* occur.[102] A study with a very low participation rate may contain little or no bias, while another study with a high participation rate may have large problems associated with participation bias. Some work has already been conducted into the effect on regression estimates from non-participation or attrition,[103–109] with some showing little effect of participation bias resulting from non-participation.

2.3.4.1 Participation, Exposure and Outcome Relation

Participation bias can be explored using causal graphs.[20,94,110–112] Generally, if participation is associated with the outcome in the analysis (conditional on all the variables in the analysis model), then there is potential for bias (regardless of whether participation is associated with the exposure).

The exception to this is when the outcome is binary and logistic regression is used, as in case-control studies. In this instance, non-participation only causes bias if it is associated with both the exposure and outcome, and an interaction between them (on the probabilistic scale). The definitions which follow focus on this scenario which applies to case-control studies.

In case-control studies, bias is said to be present when the participation variable is a collider in the causal graph and is conditioned on.[94,113–115] This definition exists not only in the epidemiology but also in sociology.[116] For participation to be a collider, there is an arrow from the exposure variable to participation, and another from the disease variable to participation. Therefore, both variables must be causally associated with participation such that participation is a common effect of the exposure and disease, and participation must also be conditioned on.[111] This also applies to a cause of the exposure or cause of the outcome, in place of the exposure and outcome respectively.[111,117] Different definitions apply for other study designs.

A detailed explanation for why this causes bias is available from Pearl[113] or Spirtes.[118] Briefly, this can be explained as follows. If exposure causes participation (say positively), and disease causes participation (again say positively), if it is known that a given participant is unexposed, then the individual is likely to have the disease. This suggests an (inverse) association between the exposure and disease, even if there is not one in the population. In this instance, the odds ratio would be likely less than one, suggesting the exposure is protective with respect to the disease of interest. In epidemiological studies, the participation variable is conditioned on when only data from participants are used in the analysis, usually to estimate the exposure-disease association. This conditioning on common effects is the definition of selection bias,[18,94,111] of which participation bias is a subset. Bias is then said to be present when the association between the exposure and outcome variables is not solely due to the result of the causal effect of the exposure on the outcome.[111]

Participation bias can also be explained without using terminology from causal graphs. Bias can arise in estimating the effect between the exposure and outcome if participation is influenced either by the exposure or by factors that influence the exposure, and is also influenced by the outcome or factors which influence the outcome.[18,117] This would again assume that only data from participants are used in the analysis.

If participation is a collider of exposure and outcome, then conditioning on participation (by

using in the analysis only those who have participated), means that the exposure and outcome are associated, conditional on participation.[18] Therefore it is expected that, conditional on participation, the exposure and outcome are associated, even if the exposure does not affect the outcome.[18] This is also true when a descendant of a collider is conditioned on.[38] Recall from §2.2.6 that confounding is where there is a common cause, and bias results when this common cause is not conditioned on. In contrast, selection (or participation) bias is where there is a common effect, and bias results when this common effect is conditioned on.[111] It is therefore important to know the direction of the arrows in a causal diagram, rather than just the association. A reversal of arrows could result in the definition of confounding as stated in §2.2.6 instead, depending on whether or not the variable has been conditioned on.

In the literature, participation bias may be more widely referred to as selection bias, which is also termed collider-stratification bias or bias resulting from conditioning on a collider,[115] all of which support the definition given in this section. A spurious association can be induced by conditioning on a collider by design or during analysis.[114,115,119] In case-control studies, conditioning on a collider is often through the study design, as will be discussed in §2.3.4.2. This can result in bias in two ways, (i) since participation is a collider which is conditioned on and (ii) since participation is a descendant of a "virtual collider" (the disease), whose parents are the exposure and the unmeasured error term of the disease (also sometimes referred to as a "hidden variable"), which is always present but not often included in graphical models.[119]

The quantification of bias resulting from a collider which is conditioned on has been studied,[110] with findings indicating the magnitude to be largest when both the exposure and disease cause participation, smaller when only the exposure causes participation and smaller still when neither the exposure nor the disease influence participation.[110] Therefore, the definition of participation bias, which fits the first of these three scenarios, can result in the largest of these three biases, causing it to be an area of interest. This first scenario is also more likely in case-control studies than other study designs, since both the exposure and disease have occurred before participant recruitment, hence the focus on this study design here.

The size of the bias is affected not only by the association between the participation, exposure and disease variables, but also by the distribution of participation. If non-participation is very rare, then the exposure-disease OR would likely approximate the true OR, with negligible bias.[110]

There are minor differences between the definitions of participation bias used by different authors.[111] A key epidemiology textbook[18] acknowledges the difference between their definition and that given by a leading causal graph article,[111] and the causal graph article acknowledges differences within and between fields.[111] While each author has their reasoning for selecting a particular definition, these differences could lead to confusion amongst those using these definitions. For this thesis, the definition given in the causal graph article[111] will be favoured, since it is widely used in epidemiology literature [the article had been cited 848 times by 04/02/2016].

An Example of Participation Bias Using Causal Graphs

A hypothetical example of participation bias demonstrated using causal graphs follows, which is very loosely based on an example by Hernan *et al.*[111] It uses an example of an inappropriately selected control group which can be common in case-control studies, either through poor selection or through non-participation.

Let there be a case-control study interested in the association between an exposure, lactose intolerance, and a disease, hypertension. Case-control studies select a higher proportion of cases from the population than controls, and cases are more likely to participate than controls, hence there is an arrow from the disease to participation. Next, let the control group be recruited from patients with osteoporosis, hence there is an arrow from osteoporosis to participation. Since individuals who are lactose intolerant may consequently have osteoporosis, there is an arrow from the exposure to osteoporosis. A causal graph for this is shown in Figure 2.1, where the exposure is lactose intolerance, the disease is hypertension and the auxiliary variable is osteoporosis. Participation is therefore a collider between the exposure and disease, and once conditioned on (by only analysing data from participants) leads to bias resulting from an inappropriate control group. In the population, there is no known association between lactose intolerance and hypertension. However in the study sample, the controls were taken from a group who had osteoporosis, and hence controls were more likely to have osteoporosis than cases. Lactose intolerance increases the occurrence of osteoporosis, so the controls are more likely to be lactose intolerant than the cases. Therefore, the odds ratio for the association between lactose intolerance and hypertension is likely to be less than one, suggesting lactose intolerance is protective against hypertension.

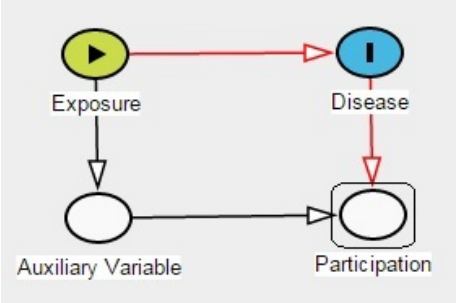

Figure 2.1: Example of a causal graph with participation bias.

2.3.4.2 Differences Between Common Study Designs

Some study designs may be more prone to participation bias than others. Here, three common designs; cohort studies, case-control studies and randomised controlled trials are considered, and examples are given in §2.3.4.4 – §2.3.4.10.

Participation bias can occur in cohort studies,[111] although this bias has generally been shown to be small.[61, 120, 121] Participation bias in this study design can occur from a biased sampling frame (and hence be closer to selection bias[120]), volunteering,[121] non-participation[61] or loss to follow up.[111] The effect of participation bias in cohort studies is often deemed negligible as the outcome has not yet occurred, and comparison between studies,[121] or simulations[61] confirm this theory, even when extreme values are used.[120]

Due to the retrospective nature of case-control studies, the exposure and disease of interest have already occurred, thus increasing the possibility of participation being dependent upon these variables. In addition, participation will be conditioned on, since only data from participants are used in the calculation of the exposure-disease association. By definition, case-control studies involve conditioning on a child of the outcome, specifically the participation variable.[18] Therefore, participation bias is more likely in this study design than other designs.

Randomised controlled trials (RCTs) can suffer from participation bias, through loss to follow up and non-participation post-selection.[111] The random element of RCTs may lead researchers to believe they cannot suffer from participation bias, but unfortunately dropout after the study has started allows participation bias to be possible, although less common than in other study designs.

Factors such as the study design make some studies more susceptible to participation bias

than others. The retrospective nature of case-control studies mean potential participants have already encountered both their exposure and disease status, which could affect their probability of participation. Some studies may blind their participants to the exposure of interest, but the disease status will always be known since it is a requirement for participation, and the mere fact the exposure has occurred, may cause differences between the participants and non-participants. Randomised controlled trials (RCTs) and cohort studies have the prospective advantage that the outcome of interest is not known at study enrollment and hence participation bias is less likely to occur. For this reason, case-control studies are the focus of this thesis.

2.3.4.3 Does the Analysis Method or the Structure of the Outcome Matter?

There are different ways to report the association between two variables such as an exposure and outcome, and these include the odds ratio (OR), relative risk (or risk ratio) (RR) and absolute risk reduction (ARR). The relative risk is the number of times more likely the outcome is in the experimental group than the control group and hence is often the desired outcome. In some instances, the relative risk will not be possible due to the required data not being available and instead an odds ratio must be used as an approximation to the relative risk, such as in case-control studies as discussed in §2.2. Case-control studies utilise odds ratios rather than relative risks so they can sample from the case and control population separately to reduce the risk of recruiting very small numbers of cases when the disease in rare. The odds ratio also has the advantage of being invariant to the labelling of the event/non-event.[122] Although individuals participate in a case-control study with probability dependent upon their binary disease status, the odds ratio for the effect of the exposure on the outcome in the study group is unbiased, as a consequence of its reversibility.[18,38] In other words, the bias produced by conditioning on a child of the outcome will cancel out of the odds ratio from the study, provided participation is only associated with the exposure through the outcome.[18] Therefore for the OR to be biased, the exposure variable must also affect participation.[38] The exposure and outcome can affect participation independently without causing bias. The examples and scenarios used in this thesis will focus on the case-control study design, with logistic regression as the analysis method.

A binary outcome variable such as case/control often uses logistic regression. Logistic regression can underestimate the probability of rare events[123] and methods have been proposed for this.[123]

It is also biased for finite samples and investigations into the bias have been conducted.[124–126] It has been known for some time that some models can be biased when analysing a non-random sample[127] and guidance has been provided for this.[127] The main advantage of the binary outcome in case-control studies, is that it can be summarised using an odds ratio (OR),[110] the advantages of which are given above.

2.3.4.4 Example 1: Participation Bias in a Case-Control Study of Alcohol

There are instances where non-participation leads to bias, as the next few examples will demonstrate. Let there be a hypothetical case-control study to determine whether excessive alcohol consumption is associated with bowel cancer. Let the cases be more likely to participate as they have greater motivation and interest than the controls, and let those who drink excessively be motivated to participate as they are interested in the potential risks of their lifestyle.

A causal graph can be formed with this information. Let D denote the disease status of the potential participants, E represent the excessive drinking and P show whether or not an individual participates in the case-control study. This leads to;

- $E \longrightarrow P$, since those who excessively consume alcohol are interested in their long-term health risks;
- $D \longrightarrow P$, since a higher proportion of cases are sampled and cases are more interested in the study than controls, as they would like an explanation or cure;
- $E \longrightarrow D$, which is the association of interest.

Since only the participants can be studied, as the non-participants do not have their details recorded, the participation variable, P, is conditioned on. There are now the requirements for participation bias; a collider between the exposure and outcome, which is conditioned on. Therefore, this scenario would result in participation bias, which would then need to be accounted for during the analysis.

2.3.4.5 Example 2: Participation Bias in a Case-Control Study With Multiple Variables

If a case-control study has multiple variables, these can disguise the conditioning on a collider, but participation bias can still occur. For example, let there be a hypothetical exposure E and

disease D, with participation P affected by the disease, since the selection criteria is based upon the disease status of an individual, and cases are often more likely to participate than controls. Now let the controls be selected from a group with a particular attribute A, such as asthma. If this attribute is affected by the exposure, and since participation is conditioned on, there is again the scenario whereby a collider is conditioned on, which is a common effect of the outcome and the exposure. This can be considered to be participation bias through poor recruitment of the control group.

- $E \longrightarrow A$; $A \longrightarrow P$; $D \longrightarrow P$; $E \longrightarrow D$.

2.3.4.6 Example 3: Berkson's Bias

Berkson's bias[128] is another example of bias, which occurs when two unassociated diseases are compared within a hospital environment. If all participants are hospitalised due to one of the two diseases, hospital admission is conditional on and a spurious association between the diseases is formed. This can be extended to case-control studies if the exposure of interest is a risk factor for the first disease, since it will also be shown to be a risk factor for the second disease (now the disease of interest) through the same reasoning.

2.3.4.7 Example 4: Participation Bias in a Longitudinal Study

In a hypothetical longitudinal study there may be loss to follow up, where individuals may 'drop out' of the study for reasons such as moving away from the study area, or death from an unrelated cause. In this instance, if the loss to follow up is caused by the exposure and caused by a variable which also causes the outcome of interest, such as a poor immune system, then bias will be caused. The bias is again due to conditioning on a variable which is a collider of the exposure and outcome, or in this case, a cause of the outcome.

2.3.4.8 Example 5: Participation Bias in Randomised Controlled Trials

Randomised controlled trials (RCTs) are often thought to contain little or no bias, since the treatment or similar is assigned randomly. However, individuals may still be lost from the study

after random allocation in a study, for reasons such as an adverse effect from the treatment drug or a disinterest resulting from being allocated to the placebo group. In these instances, the exposure has caused an effect, which has lead to the loss of a participant. If the effect is also caused by a factor which causes the outcome, which could again be a poor immune system, then bias is caused.

These examples need not be caused by loss to follow up, they could also be due to non-response in a survey or missing data in an interview. Failure to answer particularly sensitive questionnaire sections or to attend clinic meetings are also possible.

2.3.4.9 Example 6: Participation Bias in a Real Case-Control Study

One real example is a case-control study which was conducted to investigate the efficacy of Papanicolaou (Pap) smears in reducing mortality from cervical cancer.[129] The controls were recruited using a household survey, but for each of the 1060 controls used, an average of 12 households had to be contacted before a control was found. The most common reason for non-participation was that nobody was at the household at the time. The time of day, time of year and whether the day was a weekend or week day are likely to result in different types of individuals being present at the home. If the presence or not at the time of the survey is related to the likelihood of having a Pap smear within the last five years, which is the exposure of interest, then the estimated relative risk could be biased. For example, women who were not home may be those who are employed and it may be that the likelihood of having a smear test differs between those who are employed and those who are not. In this instance, replacing the unavailable employed women with available unemployed women may bias the results, since by design the outcome of a case-control study also affects participation. In situations such as these, it is advisable for the investigators to try to gain information regarding any differences between the available and unavailable individuals. That is, whether there is a difference in the likelihood of having a smear within the last five years between employed and unemployed women.

2.3.4.10 Example 7: No Participation Bias in a Randomised Controlled Trial

There are also instances where non-participation does not lead to bias. Let there be a hypothetical randomised controlled trial (RCT) for a new hayfever tablet. Potential participants are recruited to

try to determine whether the new tablet is effective. Let the participants be randomly assigned to either the drug group or a placebo group. Now let any one of the following three scenarios occur;

- The new tablet produces some unexpected side effects and some participants in the drug arm suffer from fainting or severe vomiting. Half of the participants in the drug arm withdraw from the study, as they decide that their hayfever symptoms are preferable to the side effects;
- The trial is conducted during a particular summer which is known to be one of the worst on record for hayfever suffers due to the high pollen count. Participants in the placebo arm are suffering with hayfever symptoms and feel their 'drug' is not effective, so withdraw from the study to allow them to take other hayfever remedies;
- The trial is advertised but only attracts those who have the most severe hayfever symptoms, possibly those who have not found any of the current hayfever remedies to be effective, hence they are seeking an improved drug.

Again a causal graph can be formed from these three scenarios. Let Ta be the new tablet being tested, let H denote the severity of the hayfever symptoms, and let P be whether or not the individual participates in the RCT. Each of the three scenarios has $Ta \longrightarrow H$ as the association of interest. The first two scenarios will result in a directed edge from Ta to P ($Ta \longrightarrow P$), while the third scenario will result in an edge from H to P ($H \longrightarrow P$). However, none of the three scenarios result in both edges to P which is conditioned on, and hence participation bias is not present in these examples.

2.3.5 Participation: Recent Developments in the Field

In 2007, a detailed review of participation rates in epidemiology studies was conducted, including what was known about who participates in epidemiologic studies[17] as discussed in §2.3.2. However, there seems to be no such review since then. Advances in technology, increased use of the Internet, more open data and increased data sharing have all occurred in recent years. These changes may have affected the way in which data are sought and recorded, and in turn may have affected participation rates. In addition, societal shifts may have led to differences in participant characteristics. Therefore, there follows a new review conducted using articles since the 2007 review, for a more up-to-date summary of participation in published literature.

2.3.5.1 Inclusion Criteria

Web of Science[130] was used to search titles of English articles from 2007–2015 for a range of synonymous words concerning participation rates. The title search used on 8th September 2015 was `TI=("selection rate*" OR "participat* rate*" OR "nonresponse rate*" OR "response rate*" OR "nonparticipat* rate" OR "cooperat* rate*" OR "noncooperat* rate*")`. This returned 626 articles for further consideration.

The abstract of each of the 626 articles was read to determine whether the article met the next phase of the inclusion criteria, which ensured the term referring to participation rates was in relation to a study or survey. Specifically, participation here refers only to the willing enrollment, or involvement, of an individual to a survey or study, where adequate data are provided to assist with the research question. Synonyms of participation include '(self-)selection', where an individual volunteers, 'cooperation', where an individual agrees to be involved, or 'response' relating to, say, the return of a completed questionnaire. Therefore these synonyms are provided in the context of participation in research rather than the general definition of the term. Linking these terms is the willingness of the individual to contribute data. Similarly, non-response, non-cooperation and non-participation were of interest, to understand those individuals who declined a survey or study. If the abstract was not sufficiently detailed to determine inclusion or not, the full text was sought and read. All study designs were included, such as cohort studies, case-control studies, trials and surveys, to obtain as much information as possible about factors affecting participation, with the overarching requirement that the individual had to consent to involvement in the data collection, that is, willingly participate.

From the 626 article abstracts read, 162 articles satisfied the inclusion criteria and so the full text was read thoroughly. Notes were recorded for each article, including a brief summary of the article, the year it was published and any participation findings. The results were later split into two sections; those concerning the individual participating, and those relating to the study design.

2.3.5.2 Exclusion Criteria

Unintended interpretations of the search terms such as 'response' to an intervention, 'participation' in a physical activity, or 'cooperation' with an event were not of interest, and hence these articles were excluded from the review.

During the final phase of the inclusion criteria, 464 articles were excluded, with the main reason being that the term 'response' related to a patient response to a drug or treatment (282 articles). Other reasons were repeated articles (6), articles regarding best practice (67), where 'participation' described the uptake or acceptance of an intervention (26), articles investigating the labour force participation rate (22), articles where 'participation' described involvement in a sport or activity (27) or where 'response' described a reaction using a stimulus or similar (34).

2.3.5.3 Participant Characteristics

The characteristics of the individuals found to be most or least likely to participate are listed, starting with the most reported theme, and their correspondence with previous findings noted.

Age was found to differ between participants and non-participants, as in the 2007 review,[17] with studies reporting findings such as those who were 30+,[131] 40+,[132] 51+,[133] 75+[134] or older[135–139] being more likely to participate than younger individuals. Although these studies used different age categories, they each concluded that older individuals were more likely to participate than younger individuals. One study simply stated that age was important,[140] while another found those who were younger[141] were more likely to participate in a text messaging study, although this may be a finding unique to text messaging.

Higher education levels were associated with higher participation rates in studies,[142–144] or the education level of participants was found to differ by sampling technique.[136] Higher education was a known characteristic associated with increased participation in 2007.[17] Being a homeowner was also found to be associated with an increased participation probability.[144] However there may be an association between education levels and homeownership, or between homeownership and age. Employment type was associated with participation;[140] full-time employment was associated with lower participation rates,[136,139] while unemployment was associated with increased participation rates for studies offering incentives.[145] This may be related to the amount of free time potential

participants have to complete a survey or be involved in a study, but does contradict the findings from 2007.[17]

Race and ethnicity differed between those who chose to participate and those who did not.[140] Those more likely to participate were found to be non-Asian,[146] white,[147–149] or Western,[150] generally agreeing with the previous review.[17] Participation was found to differ by country,[151] which may incorporate factors such as ethnicity and race. Location generally was also found to differ between participants and non-participants,[140, 152] with those in rural locations more likely to participate[148] than those in urban. Location may be associated with other factors discussed earlier, such as employment status, education level and homeownership.

Gender was found to be associated with participation,[138, 140] with females more likely to participate than males,[132, 135, 148, 150, 153, 154] as commonly found in studies through time.[17]

Smoking status was found to be associated with participation,[155] with non-smokers (or those who are not lifelong smokers) usually more likely to participate,[53, 139, 144, 156] as also found in the earlier review.[17] Smoking may be a factor specifically related to the study of interest, since it is unlikely to be recorded routinely for all studies.

Marital status was found to differ between participants and non-participants; with those classed as married[144] or not single[131] being more likely to participate, again agreeing with previous findings.[17]

Socio-economic class was associated with participation, with those categorised as not lower class[150] or not manual social class[144] being more likely to participate. Similarly, previous work has concluded that upper class individuals or those with a higher socio-economic status are more likely to participate.[17]

Physicians with less than 15 years experience were found to be more likely to participate than those with more experience,[148] which may be specific to physicians or even this particular study. Mental health problems were associated with lower participation,[150] although this is a variable which may only be recorded in studies where mental health is of interest. Obesity was found to be associated with lower response rates,[144] but again obesity is a factor which is often only recorded in studies associated with weight. Multiparous women, or women with pre-term deliveries were less likely to participate in a pregnancy study;[131] variables which are likely to only be recorded

in pregnancy or pregnancy-related studies. Lower pain intensity was found to be associated with increased participation probability[139] when the study considered surgery; which may or may not be generalisable to other surgery studies. These factors are less commonly recorded and hence cannot easily be compared with the 2007 review findings.

Heavy drinkers were assumed to be less likely to participate in alcohol consumption studies.[157] Although specific to this study, or studies of alcohol consumption, it may be that individuals who indulge in habits with negative connotations are less likely to participate in a study regarding that aspect of their lifestyle. Alternatively, ones function may be impaired by overindulgence in particular areas such as alcohol consumption or drug use and hence this may affect their participation in a study or their completion of a survey. Finally, cases were found to be more likely to participate than controls,[141] as found in the previous review and frequently in case-control studies.[17]

2.3.5.4 Study Design

Study factors associated with participation are summarised here, with the most frequently reported themes listed first within each topic. Some factors are specific to particular studies, whereas others are more generalisable.

Study Design: Prior to the Study

Participation was found to increase with incentives or free gifts in some studies,[136,142,158–174] but not in others.[137,175–189] Some studies found that small incentives were not quite sufficient to encourage potential participants,[190] while larger incentives were.[191] Some studies compared the size of incentives with participation rates, which could help to determine a threshold amongst certain populations of interest, but this may not generalise to all populations. Often studies which found incentives to not help study enrollment, were those offering less valuable incentives. Incentives were also usually more successful in studies which sought to enroll those who are less wealthy, or those who are busy and expect compensation for their time. A small incentive such as a free pen may be sufficient for a short survey for non-personal data, but a larger incentive may be required for a survey requiring a blood sample, sensitive data or a significant time commitment. The immediacy of the incentive was also important,[192,193] i.e. whether the incentive was given

at the time of enrollment, or promised at a later date. This mixed influence of incentives on participation was also found in the previous review.[17]

Pre-notification of the study was found to be helpful for recruitment in some studies,[138,171,191,194–196] but not in others,[197–200] even when personalised.[201] In 2007 it was thought to be a positive measure.[17] The type of pre-notification used was generally found to be unimportant.[202] However, advanced mailing of the questionnaire before a telephone survey was found to be associated with reduced participation rates.[203]

Study Design: Mode of Contact

Paper surveys have been found; to be effective,[167] to be required in addition to electronic surveys,[204,205] to be better than web surveys (completed online),[165,166,190,206,207] or electronic surveys (completed electronically but not necessarily using the Internet),[188,208–210] or advantageous over telephone surveys.[211] Although conversely, an investigation into organisation surveys found participation rates in electronic studies to be as good as or higher than mail.[177] Web surveys were found to be better than mail surveys in a study of PhD holders,[212] although offering a web option was associated with decreased participation in another study.[184] Item non-response was similar in web and mail surveys,[213] but online surveys were better for open-ended and text answers in a study of item non-response.[214] For web surveys, a welcome screen describing a survey with a short length and including less information regarding privacy, was found to be most effective.[215] Recruitment using a direct email was more successful than through a newsletter,[164] and tablet device surveys[216] or facebook[217] were found to help recruit reluctant or hard-to-reach potential participants. Exclusively online surveys were found; not to be suitable for a doctors survey,[218] generally not effective in a medical practitioner survey,[219] or less effective than other modes.[220]

Telephone calls can be useful[158,196,221] and there exists a simple positive association between the number of calls made and the response rates.[222] Utilising multiple sources to obtain a telephone number, followed by multiple phone call attempts and postal approaches, was successful at increasing participation rates in one study,[223] although this could be viewed as unethical and as a form of harassment or coercion.

Short message service (SMS) was successful for recruitment in an arthritis study[224] and an SMS reminder was found to increase response rates.[153] Text messaging an invitation received faster

responses than email invitations[225] and particular combinations of modes were found to be highly effective, such as an SMS pre-notification followed by an email invitation.[226]

The previous review also found differences in participation between survey modes,[17] but with less emphasis on modes utilising modern technology such as web surveys and SMS. Recent advances in technology may alter the effectiveness of each mode of recruitment now and in future research.

Study Design: Survey Delivery Mode & Design

Mailing was found to be an effective delivery mode for recruitment,[154,158,171,227–229] and better than emailing,[178] although being handed a survey by an acquaintance was found to be more effective than mailing in studies involving older communities.[230] Priority[189] or registered mail[171,231] were found to be associated with higher response rates,[160] but tracked mailing was associated with lower rates.[232]

Repeated mailing[189] and reminders[233–236] successfully increased participation rates, as did rewording the reminder.[170] Follow-up generally was viewed as useful,[167,237] with follow-up more effective for mail than web surveys,[238] but not helpful in all cases.[228] One study even found reminders to be associated with decreased participation rates.[177] Sending a newsletter initially was found to be more beneficial than sending a reminder later[239] and electronic reminders were not found to improve response rates in postal studies.[240] This generally supports the previous review finding of increased participation with follow-up.[17]

Response rate does not differ with envelope type,[241] envelope colour,[242] whether the material was aesthetically pleasing,[243] enhanced,[133] or contained an envelope teaser regarding an incentive.[185] However, the invitation design was found to be important[244] as were the size and colour of the paper.[245] The location of the respondent code (on the survey itself or on the return envelope) was not found to significantly affect participation rates,[246] neither was numbering the questionnaires.[247] Inclusion of a return stamp aided participation rates,[158] and stamped envelopes were found to be more effective than business reply envelopes.[239] Investigations into these factors were not so common in 2007[17] and so show a recent shift in focus of how to improve participation rates.

Study Design: Choice and Personalised Surveys

The illusion of a choice between surveys (but in fact just a different ordering of questions)

was found to increase participation,[248] as was locating the demographic data at the start of the survey.[249] Presenting the survey in multiple languages also increased participation rates,[250,251] whereas single (opposed to double) sided questionnaires and the Internet, were not found to produce significantly improved response rates.[252] Survey length was found to be significant in some studies,[158,235,236,253] but not in others.[178,179,202,254] Participation differed with the time of day[153] and with the day of the week in some studies,[153,171] but not in others.[232] These are again areas not covered by the 2007 review,[17] so show recent developments for investigations into participation rates.

A choice of survey mode (i.e. electronic, paper, etc) possibly increases participation,[158,163,255–258] but does not necessarily reduce the bias associated with non-participation.[256] These views were also found in 2007.[17] However in another study, the addition of a fax option was found to increase response rates, but other electronic options were not.[259] Multiple contact methods can increase participation rates[260] and it was found that the preferred survey mode differed between participants of different professions.[238] Similar findings were reported in 2007.[17]

Personalisation of the survey, such as through tailored letters or interaction with the potential participants, was associated with increased response rates in some studies,[158,170,171,235,237,261–263] but not in others.[264–266] Personalisation is another more recent consideration in studies of participation.[17] A persuasive message can be helpful[267] and surveys at an institutional level are found to be more successful at recruiting respondents than those conducted nationally.[210]

Study Design: Specific Studies

Participation rates were associated with features exclusive to particular studies, such as the number of days prior to surgery in an arthroplasty study,[268] or the type of cancer amongst cancer patients.[179] A child-focused protocol was also found to be more effective in children's health research, than a parent/teacher or teacher-only protocol.[269] A survey into male escorts[270] found increased response rates when the researcher posed as a client rather than a researcher, but this approach using deception may be seen as unethical. Sending a female responder to recruit male participants increased participation rates,[194] as did having a dedicated centre for data collection rather than using a generic centre.[152] Generally, the survey content was found to affect participation rates,[271] including whether samples were required such as saliva or blood.[272] These findings specific to particular studies are not easily comparable with the 2007 review.

Expert help was useful in one study,[273] as was endorsement,[158] but the additional of a logo or senior faculty's signature was not found to be helpful.[274] One view is that the potential participants need to be intrinsically motivated for participation to occur,[180] although offering the results from the study was not found to increase participation rates.[275]

Study Design: Opt-Out

Allowing the potential participants to actively decline a postal questionnaire, rather than actively agree, may be one way in which to increase participation rates,[276] since active consent was found to reduce participation.[277] Alternatively, using default settings in a web survey could be useful,[278] but this approach has the potential to lead to biased results with an excess of default responses.

2.3.5.5 Consistency and Changes Through Time

Changes over time have not generally affected the demographic of participants. Only employment status contradicted previous findings,[17] with three studies concluding a negative association of employment with participation.[136,139,145] One of these studies could be explained through the inclusion of incentives[145] raising participation rates in unemployed individuals, but the other two studies concluded full-time employment to be associated with decreased participation, possibly showing a shift in participant demographics. However the small sample size of these studies is not sufficient to draw any definitive conclusions.

In recent years, greater attention has been paid to techniques which increase participation. Studies researching envelope size, colour, style and composition are examples, with the results seen to differ by target population. This valuable information can be used to inform future studies, to ensure resources are not wasted. However, increased participation does not necessarily lead to reduced participation bias, since those participating may still differ from those who do not.[172,256]

The greatest change over time relates to participant recruitment and interaction. Although paper surveys remain the predominant survey mode, increasingly web-based approaches are being employed for recruitment, and electronic tools are being utilised during data extraction. Technology has advanced greatly in recent years and is expected to continue to do so, suggesting an even greater involvement of electronic devices in future research. The availability of tablets and smartphones has allowed users to participate 'on-the-go' and complete surveys at

a time convenient to them. Facilities such as facebook enable studies to be advertised easily and encourage the involvement of previously hard-to-reach participants. The Internet grants researchers the ability to quickly contact and enroll participants from all over the world, rather than be restricted to those locally. Advances in technology and the wider availability of devices in conjunction with social media, could result in significantly higher participation rates, particularly for studies where physical contact is not required. Even for studies requiring contact for blood or urine samples, advertisements can be circulated more widely. There will of course be studies for which this information will not be helpful. Examples includes recruitment in locations where modern technology is not common, or in populations which are not able or not willing to use technology. In some instances, this 'digital divide' could lead to increased participation bias.

2.3.5.6 Review Limitations and Assumptions

In the review, 282 of the results related to treatment response and therefore did not satisfying the inclusion criteria. Although common words such as 'virologic' or 'pathologic' were used, there was no obvious list of terms would have excluded all treatment articles. These results increased data collection time, but ensured no relevant studies were missed. The search was conducted using keywords from titles, assuming relevant research would use this or a similar title word. The abstract and keywords were trialled for inclusion, but the frequency of words such as 'cooperation' and 'participation' in the English language meant many unrelated results were returned. One article met the inclusion criteria, but could not be included as the article was unavailable using the means available.[279] It compares email and postal survey methods, but the conclusion is unknown.

Some articles reported the same dataset, either because the data appeared in multiple studies or since meta-analyses, which were included to contribute studies otherwise not captured, contained the same data. This may have altered the findings, but the effect should be reduced by the large sample of articles reviewed.

Study-specific findings were included, perhaps questionably, to demonstrate successful tactics for participation. Since the future direction of studies requiring participation is unknown, it may be that topics rarely studied now will increase in frequency, rendering these specific findings generalisable, hence they were not excluded.

Some articles assumed a causal link between study design and response rates, but it is recognised that these may only be associations. Some, such as reminders resulting in reduced participation rates, seem unlikely to be causal.

2.3.5.7 Associations Between Participation Factors

Many of the variables found to be associated with participation may be linked, for example it may be that higher proportions of older individuals live in rural locations or that more employed individuals live in urban areas. These are merely speculations, but these apparent reasons for participation or non-participation may be due to another recorded or unrecorded factor for which the identified factor acts as a proxy. Also some variables may differ between participants and non-participants, but may not have been recorded. For example, sex and age are often recorded, but factors such as obesity or pain intensity may only be recorded if relevant to the study. There is always the possibility of unidentified or unrecordable factors being associated with participation.

2.3.6 Links to Survey Non-Response

Non-participation in case-control studies could be considered to be similar to non-response in surveys. Surveys can suffer from both unit (e.g. individual 3 failed to return the survey) and item (e.g. question 5 answer missing) non-response.[280] Case-control studies may be comparable when basic information regarding a potential participation is known, for example their disease status and postcode, but other information is not recorded if they do not participate. This could be considered to be item non-response. Alternatively, the potential participation may be excluded completely from the study and treated as a non-responding unit.

Survey non-response could differ from non-participation due to the *reason* for the incomplete data. For example, item non-response may occur if the participant does not understand the question, if they do not think a particular question applies to them, or if they accidentally miss a question. Unit non-response may occur if the survey is damaged during delivery (e.g. by rain in a postal delivery or corruption in an electronic delivery) or if the potential participant chooses not to respond. Survey non-response could potentially be caused by more reasons than non-participation, although the two areas are likely to overlap. In some instances, surveys may form part of the data

collection process in case-control studies.

Epidemiology authors[281–283] have derived formulae to show how bias arising from non-participation can be adjusted for when the participation probabilities are known or can be estimated.[94] By consequence, these formulae are the same as those proposed in the survey literature.[284] Post-stratification, a method which will be discussed in §4.2, is also used in a survey context.[94,285]

Participation bias, although sometimes under different names, can be found in other fields such as econometrics[286] and machine learning.[287,288] In econometrics, it led to the development of a two-stage correction process,[286] which will be discussed in §4.4.4.

2.3.7 Links to Missing Data

Missing data are often grouped into one of three categories; missing completely at random (MCAR), missing at random (MAR) or missing not at random (MNAR).[289] MCAR is where the missingness does not depend upon the observed or missing values, MAR is when the missingness depends only upon the recorded or observed data, and MNAR is the remaining scenario when the missingness depends upon the missing or unknown values.[289] In case-control studies, all three forms of missingness mechanism are possible. In a case-control study where smoking habits are the exposure of interest for a given disease, data could be MCAR if the questionnaires are randomly lost by the postal system, data could be MAR if males choose to participate but not to reveal their smoking habits, and data could be MNAR if heavy smokers do not wish to complete the questionnaires. The remainder of this section discusses missing data in general, rather than specifically for case-control studies.

Rubin and Little are key authors in the missing data literature who have published several articles and textbooks surrounding missing data,[289–291] and methods such as multiple imputation (see §4.4.2) have been proposed to account for missing data in studies and surveys.[292] Unfortunately it is usually unknown which of the missingness mechanisms is acting within a given dataset, and this can lead to additional, untestable assumptions being required when applying methods to account for missing data.[38]

When data are MCAR, MAR or MNAR from only the exposure, complete case analysis can be

valid.[38] However, when missingness is driven by both the exposure and the outcome, the data are MNAR,[38] provided the missingness is in the exposure or outcome, and complete case analysis is no longer valid. If the missingness occurred only in a covariate then the data would be MAR (given fully observed exposure and outcome variables), but this would still result in bias from complete case analysis of the association between the exposure and outcome, adjusted for the partially observed covariate. When missingness is driven only by the outcome variable,[38] this still results in a distorted distribution of the disease, leading to a biased exposure-disease association estimate when using complete case analysis.[38] However, as already mentioned, this does not apply when ORs are used,[38] as discussed in §2.3.4.3. Therefore, while causal reasoning can be used to identify scenarios which can lead to bias, this is generally a cautious approach which cannot report specific instances where estimates may be unbiased (such as ORs).[38] This is due to the non-parametric nature of causal diagrams.[38]

Non-participation can be considered to be a form of missing data, since those who have chosen not to participate are missing from the study. However, non-participation is the (passive or active) refusal to engage in a study, whereas missing data could result from other means such as data which have been lost or destroyed, or values which were never recorded. Therefore the *reason* for incomplete data resulting from non-participation is likely to differ from that of other missing data scenarios.

Missing data is a large research area and the entire field is beyond the scope of this thesis. Non-participation can be thought of as a subset of missing data, where often the entire individual is missing, rather than just a few items (often referred to as unit non-response rather than item non-response). As already stated, case-control studies are of particular interest here since they are prone to non-participation in the control group and this can sometimes lead to participation bias as shown in §2.3.4.2. Some of the methods available to account for missing data may also be applicable to non-participation and this is the main focus of the missing data literature in the thesis. Suitable methods will be introduced in Chapter 4.

Chapter 3

Assessment of the Treatment of Non-Participation in a Sample of Published Epidemiology Literature

3.1 Aim

The aim is to summarise the treatment of non-participation by epidemiology authors in publications from October, November and December 2011. This includes recording the study designs involved, the actions taken by the authors, and which of the methods for non-participation (described in detail in Chapter 4) were used. The sources and strategy used will be explained, before the results of the assessment are presented and discussed. This summary is not intended as an attempt to disregard the findings from thoroughly planned and well-conducted studies, nor is there any attempt to present trends, but just to give a view on practices at one particular point in time.

3.2 Data Collection

Three journals in epidemiology were used for the assessment, with the most recent issue of each at the time selected. The specific journals and issues were;

- *International Journal of Epidemiology*, October 2011, *40*(**5**):1135–1428;
- *Epidemiology*, November 2011, *22*(**6**):753–881;
- *American Journal of Epidemiology*, December 1, 2011, *174*(**11**):1211–1325.

These journals were selected based upon their impact at the time of the assessment; they were the top three for impact factor and five year impact factor in epidemiology,[9] with the impact factors from 2010 (the most recently reported at the time) as shown in Table 3.1. Since the journals used have different publication frequencies, issues from different months were used.

Journal	Impact Factor	Five Year Impact Factor
Epidemiology	5.866	6.249
International Journal of Epidemiology	5.759	6.404
American Journal of Epidemiology	5.745	6.105

Table 3.1: Impact factors (2010) of the journals assessed.[9]

Each article in each issue was read thoroughly to assess non-participation and the potential for participation bias. The entire article was read since participation could be discussed anywhere in the article; in the abstract, methods, results or discussion section. Also, not all authors use the specific term 'participation'; some used 'selection', some discussed 'recruitment rates', while others did not name or highlight the potential bias, but listed the limitations of their study, which included suggestions towards participation bias.

Any data used within the article were considered for participation bias. This included data which were collected specifically for the article or data taken from a previous study. The selection process used during recruitment was considered in addition to any non-participation. To gain a greater overview of non-participation and to increase the size of the sample of articles, all study designs were considered, although it was recognised that different designs may suffer from non-participation in different ways as discussed in §2.3.4.2. The same researcher (see 'Contributions') categorised all of the articles in each of the three journals to minimise observer bias, although at the risk of increased subjectivity. In addition to the assignment of categories, the researcher also recorded the data source (for example case-control study or database), whether the article include the term "participation bias", and any methods implemented in the article which could be used to reduce the effects of participation bias.

3.2.1 Category Allocation

Within a study which suffers from non-participation, there are different options available to the the authors such as,

- ignoring that participation bias from non-participation is a possibility;
- acknowledging that bias is possible and including a statement to this effect in the article; or
- realising that participation bias is a possibility but considering it to be negligible (through, possibly incorrect, reasoning such as low rates of non-participation or the choice of study design).

Some authors may choose to apply a method designed to reduce the effects of non-participation; examples of such methods can be found in Chapter 4. As with most methods, these approaches have assumptions which must be satisfied, and the authors must have verified these assumptions and applied the approach correctly. In other instances, the article will either not contain any data, or it will be such that the study cannot suffer from participation bias. The categories used in this assessment were selected to try to encompass all of these possible outcomes regarding non-participation, without being so specific that any category would be poorly populated. Table 3.2 shows the categories used and Figure 3.1 contains a flowchart showing how the categories were assigned. Further explanations of the categories, and accompanying examples, follow.

Category Code	Explanation
N/A	The article could not be connected in any way to participation bias.
Ignored	The authors ignored the potential for participation bias.
Reasonable	The authors used a reasonable method to try to reduce the effects of non-participation.
Acknowledged	The authors acknowledged the possibility of participation bias, but they did not take any action to try to reduce it.
Dismissed	The authors considered the bias to be negligible.
Method	The authors suggested a new method for reducing participation bias.

Table 3.2: Assessment categories.

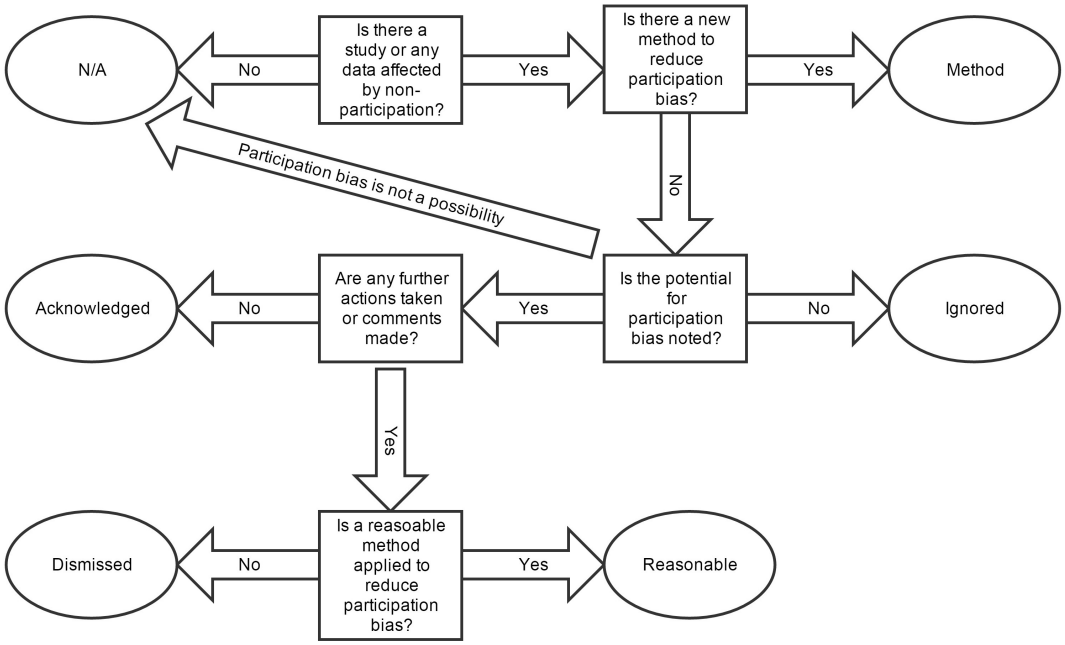

Figure 3.1: A flowchart showing the steps taken to categorise the journal articles, with the start indicated by a bold outline.

3.2.1.1 Category: N/A

The article cannot be connected in any way to participation bias, for example there is no study or there are no data. This may be in the form of a letter to the editor or a simulation study. The N/A category may also be reached if the study design is such that participation bias is not possible (refer to §2.3.4.2 and Examples 2.3.4.4–2.3.4.10 for further details).

3.2.1.2 Category: Ignored

The non-participation described in the article has the potential to result in bias as defined in §2.3.4.2, but this possibility is not mentioned. Therefore the article is considered to have ignored this possibility.

For example, a cohort study may have suffered from loss to follow up, where ten individuals have left the study due to an illness caused by the exposure. If these participants suffer from a poor immune system which has led them to leave the study and their poor immune system also makes

them more susceptible to the disease of interest, then participation bias is a possibility, since the participation bias is conditioned on, and it is a collider of the exposure and a cause of the disease. In this example, the non-participation caused by loss to follow up may lead to participation bias, but the authors may believe that cohort studies are not susceptible to participation bias and hence ignore this possibility.

3.2.1.3 Category: Reasonable

The authors have acknowledged that participation bias may have occurred and have used a method, such as those detailed in Chapter 4, to reduce the effects of the non-participation.

An example of a reasonable method could be where data are missing and a sensitivity analysis (see §4.1) is conducted using the largest and smallest plausible values in place of the missing values, to explore the possible effect on the conclusions.

3.2.1.4 Category: Acknowledged

The authors have commented on the non-participation within their study and concluded that the results may have been affected by participation bias. No further action is taken.

For example, a case-control study interested in the association between illegal drug use and mental health disorders has missing data which are caused by the users of illegal substances not wishing to participate for fear of repercussions such as legal action. In addition, the participants are selected based on their mental health status; with or without a disorder, and participation is more common amongst those with a disorder. Participation is conditioned on, since only data from participants are analysed, and so the study is susceptible to participation bias. The authors may include a qualitative summary of the possible bias, rather than a quantitative assessment or adjustment.

3.2.1.5 Category: Dismissed

The authors have commented on the non-participation within their study and the possibility of participation bias, but concluded, often without quantification, that their study conclusions remain valid. No further action is taken as the bias is classed as negligible.

Referring to the drug and mental health case-control study above, the authors acknowledge the possibility of participation bias, but qualitatively conclude its effects to be sufficiently small so as to not affect the study conclusions.

3.2.1.6 Category: Method

The article suggested a new method to reduce the effects of participation bias, or introduced an existing method from another field adapted for use with epidemiology data.

In these articles the focus is usually not on a dataset or study, but instead the methodological article educates readers of a new approach which may be used, with demonstration of its validity through simulation or an example dataset. This may also include a method from another field, such as non-response in surveys, adapted for use with epidemiology studies.

3.2.1.7 Study Design

Some study designs may be less prone to bias resulting from non-participation than others (see §2.3.4.2), therefore authors of articles using such study designs may be less likely to consider this form of bias, often with good reason. This could subsequently result in fewer articles which actively aim to reduce participation bias and this was accounted for in the assessment by also recording the study design used in each article.

3.3 Findings

The results are presented for each journal separately and then combined in §3.3.4 for an overview of non-participation across the three epidemiology journals.

3.3.1 Epidemiology

Epidemiology had 27 articles which covered a wide variety of topics, with both medically orientated articles and others with a theoretical focus. Some of the theoretical articles did not use any data. The results for the non-participation assessment are shown in Tables 3.3 – 3.5. Not

all articles included data or specified the data source, and not all articles used a method while others used more than one, hence not all the values in Table 3.5 sum to 81. However, percentages have been included where the denominator is the total number of articles in the given journal in the first three columns of numbers, and 81 for the final column. In Table 3.5, *'Possibly'* refers to methods which may be suitable for participation bias adjustment, but which are not directly stated to have been used for participation bias.

There were 18 articles classed as 'N/A' as some were purely mathematical or used simulations, along with a high number of letters (7) as well as some corrections (2). Of the nine articles considered to be relevant, four (44%) used a reasonable method to account for participation bias. Of those remaining, four articles (44%) ignored the potential problem and one (11%) dismissed it as irrelevant. This finding showed that although there were some attempts made to account for potential participation bias, there were still studies published which either did not consider the bias at all, or which concluded it to be insignificant, without any quantification of the bias. However, it is possible that the results from these studies could have been altered if the participation bias was taken into account.

In the journal Epidemiology there was one case-control study recorded, six cohort studies and three studies which used national databases. As discussed in §2.3.4.2, case-control studies are more prone to participation bias than cohort studies, which may contribute to the high proportion of articles which ignored the possibility of participation bias. National databases are intended to include the entire target population, and individual consent was not required, reducing the probability of non-participation in these studies. No article used the phrase "participation bias" and the most commonly used method to account for non-participation in Epidemiology was sensitivity analysis, see §4.1.

3.3.2 American Journal of Epidemiology

The American Journal of Epidemiology contained fewer articles than Epidemiology (16) and fewer were categorised as 'N/A' (6). As with Epidemiology there were theoretical articles, but some of these considered datasets to demonstrate their ideas, and these datasets were included in the assessment. In addition there were original research articles using data which could be assessed for non-participation. The results for the number of articles in each category are shown in Table

Category	Epidemiology	American Journal of Epidemiology	International Journal of Epidemiology	Combined
Articles Considered	9	10	23	42
Ignored	4	5	6	15
Reasonable	4	3	4	11
Acknowledged	0	1	6	7
Dismissed	1	1	5	7
Method	0	0	2	2
N/A	18	6	15	39

Table 3.3: Number of articles in each category from each journal and combined results from all three journals.

3.3. Of the ten articles considered to be relevant, five (50%) were in the category considered to have ignored the possibility of participation bias, three (30%) used a reasonable method to account for this bias, one (10%) acknowledged the possibility of participation bias and one (10%) dismissed that it was an issue; seemingly without a thorough investigation of the potential bias. Corresponding percentages and confidence intervals can be seen in Table 3.4. For some authors the possibility of participation bias was not a consideration, or possibly it was deemed to not be a problem and hence not mentioned. However without further information, results from these studies should be treated with caution.

In the American Journal of Epidemiology there were two case-control studies and seven cohort studies, see Table 3.5. As mentioned in §2.3.4.2, cohort studies are less likely to be affected by participation bias than case-control studies, which may explain why half of the articles included in the assessment were in the category considered to have ignored the possibility of participation bias, why one article dismissed its effects and why another article only acknowledged the bias without applying a method such as those in Chapter 4. As with the Epidemiology journal, no article used the term “participation bias” and sensitivity analyses were the most popular of the methods to account for non-participation.

	Epidemiology	American Journal of Epidemiology	International Journal of Epidemiology	Combined
Category				
Ignored	44%(18.9, 73.3)	50%(23.7, 76.3)	26%(12.6, 46.5)	36%(23.0, 50.8)
Reasonable	44%(18.9, 73.3)	30%(10.8, 60.3)	17%(7.0, 37.1)	26%(15.3, 41.1)
Acknowledged	0%(0.0, 0.0)	10%(1.8, 40.4)	26%(12.6, 46.5)	17%(8.3, 30.6)
Dismissed	11%(2.0, 43.5)	10%(1.8, 40.4)	22%(9.7, 41.9)	17%(8.3, 30.6)
Method	0%(0.0, 0.0)	0%(0.0, 0.0)	9%(2.4, 26.8)	5%(1.3, 15.8)

Table 3.4: Percentage (and 95% Wilson[10] confidence interval) of articles in each category from each journal and combined results from all three journals. Point estimates and confidence intervals are only calculated over those articles for which participation bias is relevant. Rounding to the nearest percentage leads to not all totals equaling 100%.

3.3.3 International Journal of Epidemiology

The International Journal of Epidemiology was the journal which contained the most articles (38) of the three considered. Fifteen articles were classed as 'N/A'; some being letters (7), others being editorial (2) and others for reasons such as being entirely theoretical, with no data to consider. There was more variability in the categories for this journal compared with the previous two as to how to deal with the issue of non-participation; see Tables 3.3 and 3.4. It can be seen that of the 23 articles considered to be relevant, six (26%) were in the category considered to have ignored the possibility of participation bias, four (17%) used a reasonable method for non-participation, six (26%) acknowledged there may be participation bias in the study but made no attempt to reduce it, five (22%) dismissed participation bias as a problem, and the remaining two (9%) proposed methods relating to participation bias. This higher proportion of positive categories (reasonable, acknowledged, method - a total of 52% for the International Journal of Epidemiology compared with 44% and 40% for the American Journal of Epidemiology and Epidemiology respectively) may suggest a greater awareness of the effects of non-participation in the articles in the given issue of the International Journal of Epidemiology compared with the previous two journals. This was despite there being one case-control study, 13 cohort studies and four database studies; therefore still a high proportion of studies which are less likely to be affected by participation bias, see Table 3.5. This journal also contained the only article which used the phrase "participation bias".[293]

Data Sources & Methods		Epi.	AJE	IJE	Total
Data Source	Case-Control Study	1 (4%)	2 (13%)	1 (3%)	4 (5%)
	Cohort Study	6 (22%)	7 (44%)	13 (34%)	26 (32%)
	National Database	3 (11%)	0 (0%)	4 (11%)	7 (9%)
Methods Used for Non-Participation	Sensitivity Analysis	2 (7%)	2 (13%)	5 (13%)	9 (11%)
	Adjust for the Variable	1 (4%)	0 (0%)	0 (0%)	1 (1%)
	Stratification	1 (4%)	1 (6%)	0 (0%)	2 (2%)
Methods Used *Possibly* for Non-Participation	Sensitivity Analysis	1 (4%)	0 (0%)	2 (5%)	3 (4%)
	Adjust for the Variable	5 (19%)	5 (31%)	10 (26%)	20 (25%)
Participation Bias Term Used?	Yes	0 (0%)	0 (0%)	1 (3%)	1 (1%)
	No	27 (100%)	16 (100%)	37 (97%)	80 (99%)

Table 3.5: The different types of articles in each of the three journals. (Epi. = Epidemiology, AJE = American Journal of Epidemiology, IJE= International Journal of Epidemiology).

However, this may not be representative of all issues of these three journals and is not intended to compare or rank the journals for their treatment of participation bias.

In the International Journal of Epidemiology, there were more studies which used national databases as a source of data (4 articles, although this was 11% of the articles in the issue and the same percentage as for the American Journal of Epidemiology, whereas Epidemiology has none). Databases should contain information regarding every member of the population and this can help to reduce participation bias, since the data are anonymised and hence individual consent is not usually required. Provided the database captures the entire nation as intended, which many do by using identification numbers for all residents assigned at birth or during immigration, and provided the database is accurate, then this is a way to reduce participation bias. For example, Denmark use a CPR-number; a unique personal identification number[60] and Sweden use a PIN; personal identification number.[59] These unique identifiers could be useful in reducing participation bias for studies in Denmark or Sweden, as they are allocated to all those born in the country at birth, and any immigrants during immigration. These identifiers can also be linked to other databases for more extensive research.[60] Sensitivity analyses were again the most popular approach when considering non-participation.

3.3.4 Combined Results

The results from all three journals were combined to give an overview of the treatment of non-participation in articles published in the field of epidemiology. The results are shown in Tables 3.3 – 3.5. It can be seen that 39 articles had to be excluded on the basis that they were not connected to participation bias. Of the 42 which could be related to participation bias, there were 15 (36%) which ignored participation bias may be a problem; note that this is more than a third of the articles considered. There were 11 (26%) articles which applied a reasonable method to correct for the bias and 7 (17%) which acknowledged participation bias may have affected the results, which is useful and allows the reader to treat the results with caution. Of those left, 7 (17%) discussed participation bias and concluded that it could be dismissed since it was negligible, which if true, demonstrates to the reader that the author has considered the effects of the bias. However, if this potential bias has not been quantified, it is difficult to class as negligible and may be providing the reader with more confidence in the published results than is warranted. Finally, 2 (5%) of the articles were specifically related to participation bias and proposed methods to help reduce it; an encouraging finding for future studies.

Overall, there appeared to be some awareness of participation bias and attempts were being made by some authors to reduce its effects. However, there are still some authors who did not appear to consider participation bias in their studies. This may partly be due to fear from authors that journals may not accept their article or that readers may not trust their results, if they suggest possible bias relating to participation. However it may also be due to the study designs used and the evidence that some study designs do not need to be as concerned about the possibility of participation bias compared with others, see §2.3.4.2.

Overall there were four case-control studies, 26 cohort studies and seven database studies as shown in Table 3.5, with cohort studies and databases less prone to participation bias than case-control studies. However, participation bias could always be mentioned, and those in the International Journal of Epidemiology seemed to consider non-participation more, despite having only one case-control study and several cohort or database studies. Overall 36% (23.0%, 50.8%) of the articles ignored potential participation bias, and 17% (8.3%, 30.6%) dismissed it, usually without published quantification. Only 26% (15.3%, 41.1%) of the articles used reasonable methods, although 5% (1.3%, 15.8%) of the articles proposed new methods to reduce participation bias.

3.4 Impact on Research

The assessment results showed that in 2011 the possibility of participation bias was still not reported by some authors, although others were attempting to reduce its effects or include it as a limitation in their study. By realising participation bias may have an effect on their study, even if less likely in some designs, the author is allowing the readers to judge whether the results are valid.

One reason for some authors to acknowledge rather than try to reduce the effects of participation bias may be uncertainty in how to select an appropriate method since so many are available. This issue has been addressed in §4.5 with a newly developed flowchart tool in Figure 4.4.

There may be concerns amongst authors that their results may be disregarded if the article shows high rates of non-participation. There may even be fear that it would not be published as a consequence. However, it would be better practice if this information was displayed for the reader so they could make their own, informed decision regarding the validity of the results. It may be useful for authors to provide a set of results; initial results and those reanalysed using possibly more than one method to try to reduce the effects of the bias. This approach has been adopted by some of the authors in these three journals and could be a useful idea for future articles. This general approach has also been used with the diabetes dataset (in Appendix A) throughout the thesis to compare different methods. However, the limited space within journals may not always allow for this. Studies less prone to participation bias could state that participation bias is unlikely to be a problem to strengthen their findings.

Sensitivity analyses, stratification and adjusting for variables were the most common approaches found in this assessment and for this reason these methods are the focus in Chapter 4. Sensitivity analyses (§4.1) often compare the different unadjusted and adjusted results and this gives valuable information to the reader, but may not always be possible with restricted article space in many journals. It could, however, form useful analyses which could be displayed as supplementary material or on a linked webpage. Stratification and variable adjustment were other methods adopted and §4.5 will show that different methods for reducing participation bias are more convenient for some studies than others depending upon the information available.

The term “participation bias” was used by one article[293] in this assessment; showing how

infrequently the phrase is included. This may be caused by authors avoiding a term with negative connotations or it may be that other terms were used, such as "selection bias" or "low participation rate".

The articles included in the assessment were carefully planned and well conducted. There was also attention paid to minor details in order to produce accurate results and to minimise many different forms of bias. However, the results of this review showed that participation bias is a form of bias which may need more consideration.

Although there was a total of 81 articles in the three journals issues included, 39 could not be considered for non-participation due to the lack of a dataset or the nature of the article. This reduced the sample size of the study, but the sample still allowed an overview of the treatment of non-participation in epidemiology journals. However this smaller sample size did prevent any further analysis, such as by each study design. Factors such as the study design are important here, since some designs are more susceptible to participation bias than others, as discussed in §2.3.4.2. There were only four articles across the three journals assessed which reported on case-control studies, which can be considered to be more frequently prone to participation bias. Therefore, other authors may not have considered or accounted for participation bias when using study designs which are less likely to be affected by this bias. They may instead have chosen to focus on other forms of bias which are more likely to occur in their particular study design. The finding here, that participation bias is not always accounted for, may be partly explained by the study designs included in these journals. Had there been more case-control studies, they may have been more consideration of potential participation bias. This may also suggest that case-control studies have declined in popularity, possibly due to traits such as a susceptibility to participation bias.

It is appreciated that the summaries and categories used are subjective, however one researcher (see 'Contributions') was used for all data collection to minimise observer bias and maintain consistency. Each article was also read thoroughly to ensure all references to non-participation were considered, regardless of the section in the article or the terminology used to describe the bias. The journals used for analysis may not be representative of all journals which could be affected by participation bias, but it would be impractical to consider all journals, so those with the highest impact factors were selected. It is also possible that there may have been changes over

time, since these journals (from October–December 2011) are no longer current, despite being the most recent at the time. However, the more recent (2015) literature review regarding participation rates which was given in §2.3.5 suggested that there had been few changes in recent years for participation rates, and it is possible that a similar trend may be found in the treatment of non-participation in studies. A 2016 review may report different findings, but this assessment from a given point in time is still an informative summary. Finally, it is accepted that the selected articles from each journal, which were the most recent at the time of data collection, may not reflect the overall articles which the journal publishes.

As mentioned previously this is not an attempt to compare journals, nor is it a criticism of the articles which have been published. It is instead an assessment of how non-participation is considered in a range of typical articles. The results showed that participation bias is not always considered and hence remains a possibility in epidemiology. More awareness and clearer guidelines on how to reduce the effects of non-participation are required and attempts will be made in Figure 4.4 for this in the form of a guidance tool to select an appropriate method. Additionally, new solutions related to non-participation have been proposed in Chapters 5 – 7, but first Chapter 4 describes the current methods in further detail.

Chapter 4

Methods to Reduce Participation Bias

This chapter uses the results from the assessment in Chapter 3 to identify then describe in detail the three most frequently used methods to account for non-participation. To ensure a more thorough report, this is followed by a brief description of alternative methods encountered during the literature search executed for Chapter 2. Each of the three methods used most frequently, sensitivity analysis (9 uses in the Chapter 3 assessment), stratification (2 uses) and adjustment for the participation variable(s) (1 use), will be assigned one section of this chapter and will include a method description, numerical example, application to the diabetes data used throughout the thesis (see Appendix A) and a critical evaluation. Other available methods are discussed in §4.4, which includes an overview of each method and a critical evaluation. In §4.6.1 similarities between the methods in the chapter will be stated, and a discussion for how the methods can be used to complement one another.

Throughout the chapter, terms such as selection bias and non-response bias will be used in addition to participation bias. This is to distinguish between methods specifically designed for use with non-participation and those applicable to more scenarios such as selection bias, of which participation bias is a subset (see Chapter 2).

4.1 Sensitivity Analysis

Sensitivity analyses are the repeat of an analysis which substitutes alternative decisions or ranges of values for decisions that were arbitrary or unclear.[294] This includes values which were missing

or estimated through the result of non-participation. In this section, approaches which are suitable for use with case-control studies and which go by the name of a sensitivity analysis are included. However, other approaches could be seen as sensitivity analyses, such as analyses before and after multiple imputation (§4.4.2) or before and after weighting (§4.4.1), but these approaches have been allocated their own section and are not repeated here.

4.1.1 Explanation

Greenland[295] describes a sensitivity analysis rather concisely as the quantitative extension of the qualitative elements which are found in good discussions of results. Hence sensitivity analyses can be seen as an attempt to link traditional statistics, where the assumptions may not hold, with more informed but less formal inferences that acknowledge biases.[295]

The first sensitivity analysis of bias in an observational study was in 1959[296] and considered the claim made by tobacco companies that the high rates of lung cancer amongst smokers may not be due to smoking.[297] However, the sensitivity analysis[296] found that any unmeasured factor responsible for causing lung cancer would need to be a near perfect predictor of the cancer and approximately nine times more common amongst the smokers than non-smokers.[297] While this unmeasured factor is possible, it is unlikely.[297]

Sensitivity analyses have been used in epidemiology[298,299] in an attempt to assess the effect of non-participation on a study, which can be performed in different ways depending upon the available information. One approach may be to assume the exposure levels for those who have not participated and re-analyse the data.[14] Various assumptions can be made about the unknown exposure levels and each tested in turn, with their effect on the odds ratio (or similar) reported.[14] This includes testing the extreme instances whereby all non-participating cases/controls are exposed/unexposed. Conclusions can be drawn about the maximum possible differences between the study-calculated odds ratio and the possible odds ratio if all invited individuals had participated. If the difference between the calculated and hypothesised odds ratios is small, the study odds ratio can be reported with confidence. However, if this difference is large or causes the exposure to change from a risk to protective factor or vice versa, then the effect of non-participation should be discussed in the reporting of the results. Such discussion should also include the likelihood of the missing exposure levels including the extreme possible values,

despite being a subjective likelihood.[14]

One author[300] stated that sensitivity analyses were rarely used between 1959 and the time of their publication in 2003, and thought it due to the method requiring values from background information which may be unreliable or controversial and which the study results may be sensitive to.[300] This limitation is similar to Bayesian analysis where there is the need to select a prior, upon which the posterior may be sensitive.[110]

Bias resulting from non-participation can be difficult to assess, due to the lack of sufficient information available to perform a quantitative analysis,[301] but sensitivity analyses attempt to estimate the effect of this bias. Various forms of sensitivity analysis have been suggested; some examples designed for use with participation bias follow.

4.1.1.1 Approach 1: Decomposition of the Odds Ratio

This sensitivity analysis for participation bias involves the decomposition of the odds ratio.[295,301] Let p_{Ca_j} and p_{Co_j} be the participation probabilities of cases and controls respectively at a given exposure level, j. In case-control studies where the exposure is binary, j will take the value 1 for exposed individuals and 0 for unexposed individuals. Population case counts can then be estimated using $\frac{Ca_j}{p_{Ca_j}}$ and population control counts can be estimated using $\frac{Co_j}{p_{Co_j}}$. The adjusted odds ratio which compares exposure level j to level 0 is then,

$$\frac{\left(\frac{Ca_j}{p_{Ca_j}}\right)\left(\frac{Co_0}{p_{Co_0}}\right)}{\left(\frac{Ca_0}{p_{Ca_0}}\right)\left(\frac{Co_j}{p_{Co_j}}\right)} = \frac{Ca_j Co_0}{Ca_0 Co_j}\left(\frac{p_{Ca_j}p_{Co_0}}{p_{Ca_0}p_{Co_j}}\right)^{-1}, \tag{4.1}$$

which is calculated by dividing the sample odds ratio by a value termed the 'selection bias factor', $\left(\frac{p_{Ca_j}p_{Co_0}}{p_{Ca_0}p_{Co_j}}\right)$,[295,301] but could equally be named the 'participation bias factor'. Equation 4.1 can be applied to the whole sample, or within strata of confounding variables, and the selection bias factor need not be constant across the different strata.[301] Note that there will be no bias if the selection bias factor is equal to one,[295,301] since the sample odds ratio will be returned. Although this is less common for case-control studies since the exposure and disease are known before the study commences.[301]

4.1.1.2 Approach 2: Cornfield's Inequality

In 1959, Cornfield *et al.*[296] achieved the first formal sensitivity analysis on an observational study by deriving an inequality for a risk ratio which was defined as the probability of death from lung cancer for smokers divided by the probability of death from lung cancer for non-smokers.[297] The statement given by Cornfield *et al.* was,

> "If an agent, A, with no causal effect upon the risk of a disease, nevertheless, because of a positive correlation with some other causal agent, B, shows an apparent risk, r, for those exposed to A, relative to those not so exposed, then the prevalence of B, among those exposed to A, relative to the prevalence among those not so exposed, must be greater than r."

The key action here is the conversion of a usually qualitative statement applicable to all observational studies to a quantitative statement specific to a particular study.[297] In other words, rather than state that associations do not imply causation and that hidden biases may explain the observed associations, the authors instead state that to explain the observed association in a given study, the hidden bias would need to be of a certain magnitude.[297] Therefore a strong association would need to be explained by a large bias.[297]

4.1.1.3 Approach 3: Rosenbaum's Extension of Cornfield's Inequality

Rosenbaum[297] aimed to extend the inequality suggested by Cornfield[296] such that it was suitable for any type of response (not just binary) and could take into account sampling variability.[297]

The assumption of this approach is that, before any matching or stratifying, individuals are assigned to exposed or unexposed groups independently, but with unknown probabilities.[297] In other words, two individuals who have the same recorded characteristics may differ by unrecorded characteristics, such that their odds of exposure differ. The odds of exposure are denoted using Γ, with $\Gamma \geq 1$. Therefore, in a simple randomised controlled trial, $\Gamma = 1$ since each individual has the same odds of being exposure, or receiving the treatment. A value of two for Γ would suggest that an individual is twice as likely to be exposed than another individual, and this difference is due to unrecorded characteristics.

This approach by Rosenbaum quantifies the amount of bias required from unrecorded factors before the qualitative study conclusions change.[297] A very sensitive conclusion would be changed when Γ is not much larger than one, while a more robust or insensitive study conclusion would only be changed by large values of Γ.

4.1.2 Hypothetical Example

Let there be hypothetical data from a case-control study as shown in Table 4.1, with 50 cases and 100 controls. The odds ratio for these data is $\frac{70\times35}{15\times30} = 5\frac{4}{9} = 5.44$(2dp). The `rbounds` package[302] in *R*[303] can be used to perform the sensitivity analysis as in §4.1.1.3. The results are presented in Table 4.2 and show that for an increase of 2.5 in Γ, then p-value increases to 0.09, which is above the usual 0.05 threshold.[304] So, if the case or control individuals are 3.5 times more likely to be exposed due to an unobserved variable, then the inference changes. This indicates a fairly reliable case-control result, as the presence of an unobserved variable of this nature seems unlikely.

	Control	Case
Not exposed	70	15
Exposed	30	35

Table 4.1: Hypothetical data for the sensitivity analysis example.

Γ	p-value lower bound	p-value upper bound
1.0	5.56e-07	5.56e-07
1.5	8.04e-10	1.03e-04
2.0	3.54e-12	1.89e-03
2.5	3.42e-14	1.16e-02
3.0	5.91e-16	3.90e-02
3.5	1.59e-17	9.05e-02
4.0	6.11e-19	1.65e-01

Table 4.2: Sensitivity analysis hypothetical data results.

4.1.3 Sensitivity Analysis of Participation Bias in the Diabetes Data

Sensitivity analyses similar to the hypothetical example in §4.1.2 can be applied to the diabetes data used throughout the thesis. The results from using caesarean or amniocentesis as the exposure of interest, with type I diabetes as the outcome of interest, are given in Table 4.3. For caesarean, an increase in Γ of just 0.1 (upper bound) is sufficient to change the conclusion from significant to insignificant, whereas for amniocentesis Γ would need to increase by 0.6. This suggests the amniocentesis conclusion may be more robust than the caesarean conclusion, since an unobserved variable would need to cause the case or control individuals to be 1.6 or 1.1 times more likely to be exposed respectively for the inference to change.

	Γ	p-value lower bound	p-value upper bound
Caesarean	1.0	2.31e-02	0.023
	1.05	1.46e-02	0.035
	1.10	9.22e-03	0.052
	1.15	5.79e-03	0.072
	Γ	p-value lower bound	p-value upper bound
Amniocentesis	1.0	3.04e-03	0.003
	1.2	7.40e-04	0.011
	1.4	1.98e-04	0.026
	1.6	5.78e-05	0.053
	1.8	1.83e-05	0.091
	2.0	6.21e-06	0.140

Table 4.3: Sensitivity analysis results for the diabetes data.

4.1.4 Critical Evaluation

As stated at the start of this section, sensitivity analyses could also include approaches such as multiple imputation (§4.4.2) or weighting (§4.4.1), but here methods designed for use with case-control data, which go by the name of a sensitivity analysis, are included and evaluated.

Sensitivity analyses were the most popular method found in the assessment in Chapter 3 with

nine uses, but there are limitations with using this approach. For example, the decomposition of the odds ratio in §4.1.1.1 will only be possible if the participation probabilities for Ca_j and Co_j are known or can be estimated, which may require the inclusion of a survey in the study to collect additional data regarding participation.[301] Although, it is possible to use Equation 4.1 with more than one value for the participation probabilities and report a range of adjusted odds ratios dependent upon given assumptions. The approach by Cornfield *et al.*[296] in §4.1.1.2 is restricted to binary outcomes and ignores sampling variability.[297]

Geneletti *et al.*[305] state that the "choice of adjustment method depends on the assumptions that are considered plausible regarding the nature of the non-participation and the type of additional sources of data that are available. However, any chosen model will generally be based on untestable assumptions, because by definition we do not observe the characteristics of primary interest of the non-participants. Thus any method that attempts to correct for non-participation bias is essentially a sensitivity analysis". While this is a relatively strong statement, it is true that methods to adjust for participation bias do rely upon assumptions which are often untestable and which may not be true. For example, some methods (see §4.4.2 and §4.4.1) often require the data to be MAR, and testing between MAR and MNAR data is not possible.[305] The safest option with respect to the missingness mechanism, would be to choose a method which can be used for MNAR data to err on the side of caution. Even when causal graphs are adopted to aid the adjustment or not of variables relating to participation bias (see §4.3), there is still the assumption that the causal graph has been draw correctly and that all relevant variables are included.

Therefore, a sensitivity analysis may be a wise choice, and it can be included in conjunction with another method. The sensitivity analysis need not be complicated, but could be as simple as replacing all the missing data with the highest or lowest plausible values to generate a range of results and determine whether the conclusion may be altered. In many instances, this simple sensitivity analysis approach may strengthen the study findings if it shows the conclusions to be robust to these extremes. Where assumptions are made during the analysis, the impact of these assumptions being incorrect should be considered, and the likelihood of the conclusions changing should be reported. The suggestion[305] that studies should report a range of models or a base model plus a series of sensitivity analysis is a sensible one, since no model will be 'correct', despite epidemiologists often reporting findings based on a single most suitable model.

For this to be common practice, support would be needed from journals to allow authors to publish extended results, as current space restrictions may prohibit this. However, many journals allow for supplementary materials, which could be used to present sensitivity analyses.

A quote from Allen and Holland[306] twenty-seven years ago states that "You must be prepared to think as hard about your non-respondents as you do about your substantive research and to incorporate this into a sensitivity analysis. Otherwise, you have not handled selection bias but have only ignored it". This rightly encourages researchers to consider non-participants more frequently and more deeply in their study, and to conduct and report a sensitivity analysis, even if it is only discussed briefly in the main article.

Sensitivity analysis may be a particularly useful approach for when the area of research is new or not well understood as it is suitable in most instances and different forms of sensitivity analyses are available depending on the available data and the required outcome. The main disadvantage of this method is that it typically only assesses bias, rather than directly reducing it. However if the author reports different results from different analyses, it allows the reader to assess how plausible the results are and which conclusions to draw.

4.2 Stratification

4.2.1 Explanation

Stratification usually refers to stratified matching at the beginning of a study, where cases and controls with similar characteristics are paired.[14] However post-stratification[14,285,307] allows strata to be used during study analysis.

The advantages of allocating cases and controls to strata after data collection are (1) the matching variables need not be chosen before the sample characteristics are known and (2) participant recruitment is not restricted to accommodate matched characteristics. In addition, the ratio of cases and controls does not need to be constant across strata, although a disadvantage is that comparisons will be prevented where there are not both cases and controls in a given stratum.[14]

The reasoning for stratification as a method to reduce bias is as follows. Within each stratum the cases and controls are similar with respect to the stratification variables, so it is assumed that any

differences in exposure between the cases and controls cannot be due to these variables.[14] While these 'similar' individuals can be compared across the disease categories, information regarding the stratification variables is lost, hence strata should be formed using factors which are already well-understood.

The number of strata is a matter of compromise. The more stratification variables and the greater the number of variable categories, the more similar the individuals are for comparison, but the fewer the individuals within each strata, and the greater the possibility of no cases or no controls for comparison. One advantage of post-stratification over stratified sampling is that strata specification can be amended during the analysis to accommodate the data, although the choice of strata should be supported with reasoning. Both the number of stratification variables and the number of variable categories are limited by the number of participants.

In the literature, stratification is commonly used for confounding variables,[111] but this can have the consequence of introducing selection bias[111] and in these instances it should be considered whether the size of the biases can be estimated and whether the smaller of the two biases should be selected. If individuals are stratified by age (e.g. four groups), sex (two groups) and race (e.g. four groups), this would result in 32 subgroups ($4 \times 2 \times 4$), and within these groups individuals should be relatively homogeneous on each of these three variables. Within the 32 subgroups any difference in exposure between case and control groups must be due to a variable other than age, sex and race. The smaller the number of stratification groups, the more dissimilarity there may be within subgroups, and this may account for some variation in exposure between the cases and controls.

An example of a stratification-based method is the Mantel-Haenszel analysis which is often used for confounders[39] (see Equation 2.1 in §2.2.6). Odds ratios can be calculated for the exposure-disease association in each strata as described in §2.2.5, and the results may then be reported as a set of odds ratios, one for each stratum, or be combined into one odds ratio using the Mantel-Haenszel adjustment. One example is an article[308] which used the Mantel-Haenszel approach to analyse data from a case-control study which had three sets of participants (full, partial and non), to assess possible bias resulting from non-participation. Although the term 'stratification' was not used in their publication, the data were stratified by participation group (full, partial or non), by disease group (case, control) and by sex (male, female, all).

4.2.2 Hypothetical Example

In a case-control study, let the hypothetical exposure, outcome and factor (race, which affects both exposure and participation) be binary. The data are given in Tables 4.4–4.5. Standard odds ratio can be calculated using Table 4.4 as $\frac{150\times20}{120\times15} = \frac{3000}{1800} = 1\frac{2}{3} = 1.67$(2dp). The Mantel-Haenszel odds ratio can be calculated after stratification by race using $\frac{\sum ad/n}{\sum bc/n}$ in Equation 2.1 where a, b, c, d, n are as given in Table 2.1. Table 4.5 enables the calculation to be $\frac{\sum ad/n}{\sum bc/n} = \frac{(1050/200)+(400/105)}{(550/200)+(100/105)} = \frac{9\frac{5}{84}}{3\frac{59}{84}} = 2.45$(2dp), which is higher than the unadjusted odds ratio of $1\frac{2}{3} = 1.67$(2dp). These calculations can be verified using the `epiR` package[309] in R[303] with the `epi.2by2` command. The output in Table 4.6 confirms the calculation by hand and provides confidence intervals. The crude odds ratio of 1.67 has a confidence interval which includes one, suggesting the exposure is not a significant risk nor protective factor. However after stratification by race, the Mantel-Haenszel adjusted odds ratio is 2.45 and shows the exposure to be a significant risk factor since the confidence interval excludes one.

	Outcome	
Exposure	1	0
1	20	120
0	15	150

Table 4.4: Hypothetical data for the stratification example.

Race=0	Outcome	
Exposure	1	0
1	15	110
0	5	70

Race=1	Outcome	
Exposure	1	0
1	5	10
0	10	80

Table 4.5: Hypothetical data for the stratification example, split by race.

	Outcome +	Outcome -	Total
Exposed +	20	120	140
Exposed -	15	150	165
Total	35	270	305

Estimate type	Point estimate (CI)
Odds ratio (Wald CI) (crude)	1.67 (0.82, 3.39)
Odds ratio (Mantel-Haenszel)	2.45 (1.06, 5.64)
Odds ratio (crude:Mantel-Haenszel)	0.68

Table 4.6: Stratification data and analysis output.

4.2.3 Stratification During Analysis of the Diabetes Data

Stratification can be applied to the diabetes data. Let the association of interest be between caesarean deliveries and childhood type I diabetes. Say that control mothers who underwent an amniocentesis were more likely to participate and more likely to have a caesarean delivery.

The results are shown in Table 4.7 for standard analysis, then analysis after stratification by amniocentesis. In this instance the odds ratio and confidence interval do not alter by much, but it could be said that the odds ratio has changed from significant in the crude estimate to insignificant using the adjusted estimate. Stratification of these data by amniocentesis may not be necessary.

	Outcome +	Outcome -	Total
Exposed +	34	35	69
Exposed -	162	290	452
Total	196	325	521

Estimate type	Point estimate (CI)
Odds ratio (Wald CI) (crude)	1.74 (1.04, 2.89)
Odds ratio (Mantel-Haenszel)	1.65 (0.99, 2.77)
Odds ratio (crude:Mantel-Haenszel)	1.05

Table 4.7: Stratification data and analysis output: Diabetes data.

4.2.4 Critical Evaluation

The number of stratification variables and the number of variable categories dictate the number of models required, with a large number of models potentially leading to extensive computational time, because as the number of variables to control for increases, the number of strata grow exponentially.[310] The participants also need to be distributed across the categories of the stratification variable in such a way that no categories are left poorly populated or unpopulated and such that both cases and controls are present for comparison. While 'poorly populated' is not defined as such, the fewer individuals present, the wider the confidence intervals. Therefore this approach usually requires a substantial number of participants, possibly more than other methods, which may be difficult at a time when participation in studies is declining (see Chapter 2). However, if the participants are heavily biased by a particular variable, or if certain combinations of variables are impossible, there may be some empty categories regardless of the sample size and then this approach may not be preferable. Multiple stratification variables can be used at once, but will again be limited by the number and distribution of participants.

Stratification is a relatively simple idea to implement since the usual analysis is applied (albeit to subgroups) and hence it is suited to a range of study designs. This may explain why it was the second most popular approach in the assessment in Chapter 3 (although the small number of methods returned meant it only had two uses). A limitation of stratification for if the subset results are not pooled, is that the confidence intervals will tend to be wider due to the smaller numbers in each stratum. It may then be unclear whether an exposure is a risk (or preventative) factor.

To apply this method the data need to be presented with sufficient detail to stratify by variables and still be able to perform the intended analysis. If further stratification is required, for example by a second variable, this additional level of detail would be needed. If only summaries of some variables are given, for example the number of individuals in each age group, which is not linked to the rest of the data, or if any variables have been categorised, for example age has been split into old/young, then stratification may be restricted to these categories or in some instances may not practical or possible.

Stratification includes the variable associated with participation in the analysis using a similar approach to adjusting for that variable (see §4.3). In fact the term stratification is sometimes used to

describe when a variable is included in the analysis model. Stratification is a more time-consuming approach than regression adjusting, since several models are created and then usually combined, whereas when adjusting for a variable, just one model is formed. For this reason, stratification could be viewed as less desirable than adjusting, although in Chapter 3 stratification was used more obviously for bias whereas adjusting appeared to be used more often for confounding. Of course is it possible that the inconsistency in the literature of the use of the terms bias and confounding (discussed in more detail in §4.3 and §4.6.2) may have affected the ordering of the popularity of these methods.

When a variable is conditioned on (or controlled for[115]), or when a subset selected,[111] this is the same as stratification, but in some instances just one strata is available. For example, when only participants are included in a study and non-participants are excluded, this is the same as using strata of participation (yes/no), but rather than analyse participants and non-participants separately, the non-participant data are unavailable.

4.3 Adjusting for the Variable Associated with Participation

4.3.1 Explanation

In case-control studies, participation is conditioned on since information is only available on those who have participated in the study. In other words, the participation variable has been stratified, but only one strata is available for analysis. In some instances, the binary participation variable will also be a collider between the exposure and disease variables (as defined in §2.1.1 and shown in §2.3.4.1), and this is the definition of participation bias.[111] However, it may be possible to eliminate this bias by conditioning on another variable. Conditioning, controlling or adjusting can refer to restriction, stratification, or regression adjustment[18,115] due to the overlap in the reasoning behind these approaches. It is regression adjustment[111] which is focused on here, which can also be referred to as variable adjustment,[311] controlling for a variable, or in matched case-control studies, conditional logistic regression.[312] These terms are also often used interchangeably with stratification techniques.[111,312] For confounders, Cochrane[313] use "controlling for" as an overarching theme, and use matching, stratification and modelling as subthemes. Therefore, while it has been assumed here that adjustment refers to regression adjustment, it may refer to other

similar methods since there are common themes between these approaches. Adjustment was the third most popular approach found in the assessment in Chapter 3 (but due to the small number of methods returned, had just one use).

Adjustment for variables is common when the variables in question are confounders. As already mentioned, there is an overlap (or possibly confusion) between the description of biases such as selection bias, participation bias and confounding bias, or terms such as selection confounding. Therefore, some references to adjustment for selection bias using regression models may in fact be describing adjustment for confounding.

Which variable(s) to adjust for, or simply the identification of a collider which has been conditioned on, is best achieved by using causal diagrams. Recall from §2.1.1 that a causal diagram generally consists of vertices which represent the variables in the analysis, with directed arrows between them representing direct causal effects.[115] The absence of an arrow indicates a strong claim of no causal effect between two variables, and conditioning is represented by placing a box around a variable.[115] Directed acyclic graphs (DAGs) are often used for such analyses and have the advantage of no parametric (often untested) assumptions such as linearity.[12] A graphical approach using causal graphs may be appropriate to know when adding variables to regression models is beneficial, or when it could lead to further bias. There are examples in the literature where adjusting for confounding bias by conditioning on a variable, results in selection bias.[111]

Let there be participation bias as defined in §2.3.4.1, and as a consequence a misrepresentation of a particular variable in the sample compared with the target population, and let this particular variable also be associated with the exposure of interest. Even if this variable is not a true confounder between the exposure and disease of interest, a confounding effect can be seen in the data, which can be controlled for during analysis in the same way as controlling for a confounder.[314,315] However, it is advisable to only adjust for variables which are thought to lead to bias, since each adjustment can contribute to an increase in the variance associated with a parameter estimate.[314] This is also referred to as over-adjustment, which is used to describe controlling for a variable that increases net bias or that does not affect bias but which decreases precision.[316]

4.3.1.1 Causal Graphs

Some terminology for causal graphs was introduced in §2.1.1 and used in §2.3.4.1. When the exposure and disease share a common effect, they will be conditionally associated when the association is calculated within strata of the common effect,[111,113] regardless of whether or not there is a true association. This also applies to a cause of the exposure or outcome in place of the exposure or outcome.[111] The common effect in case-control studies can be the binary participation variable. Figure 4.1 shows the simplest form of participation bias using causal diagrams. The box around the participation variable indicates that participation has been conditioned on, i.e. only those who have participated are included in the analysis. Therefore, a collider (or common effect) of exposure and disease has been conditioned on, leading to an association between exposure and disease regardless of whether or not there is a true association between them.

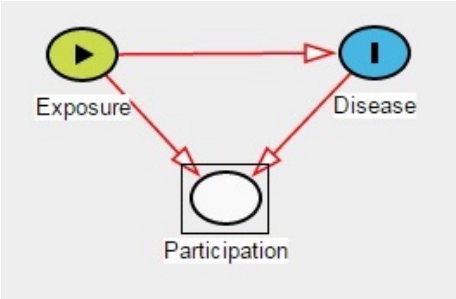

Figure 4.1: An example of a causal diagram showing exposure and outcome affecting participation.

Next, causal graphs can be used to determine when it is appropriate to adjust for participation bias. In the simplest scenario when there are just three variables of exposure, disease and participation, it is thought that when participation is conditioned on and caused only by the exposure, then the bias resulting from non-participation can be removed.[119] However, when participation is conditioned on and caused only by the disease, the bias resulting from non-participation cannot be eliminated, unless an odds ratio is used to describe the association,[119] as was discussed in §2.3.4.3. When participation is conditioned on and affected by both the exposure and disease, the bias also cannot be removed,[119] even by using an odds ratio. Of course, non-participation may result from variables other than the exposure and disease, but this scenario is the simplest.

Overall, confounding is where there is a common cause, and bias results when this common cause

is not conditioned on. Causal graphs can be particularly useful for defining a minimally sufficient set of confounders to adjust for. In contrast, selection (or participation) bias is where there is a common effect, and bias results when this common effect is conditioned on.[111] While causal assumptions can be displayed in the causal diagram, the underlying parametric assumptions should be checked using the data.[38]

4.3.2 Hypothetical Example

As seen in §4.3.1.1, causal graphs or directed acyclic graphs (DAGs) can be useful when considering confounding and participation bias in studies. Let there be a hypothetical matched case-control study, where the association of interest is between hypertension and stroke, with cases and controls matched on age. Participants are selected by their disease status and participation is more common amongst cases than controls, hence stroke status influences participation. Previous studies (see §2.3.2) have found that age affects participation, with older individuals generally more likely to participate. Age is also known to affect hypertension, with older individuals more likely to have hypertension.[317] By design, case-control studies condition on participation, hence the collider between stroke and age is conditioned on. The causal graph corresponding to this is given in Figure 4.2, which is simplified to assume no causal link between age and stroke.

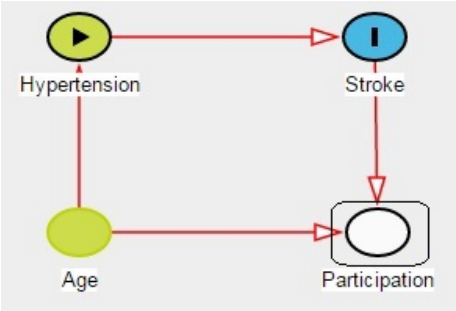

Figure 4.2: Causal graph for the hypothetical stroke example.

In Figure 4.2, age is not a confounder between hypertension and stroke. However, since participation has been conditioned on, by only including participants in the analysis, the path from hypertension to stroke via age and participation is now unblocked and hence acts as a biasing path.[18] To estimate the association between hypertension and stroke, this biasing path must be blocked, which can be achieved by conditioning on age. This would also be true had age and

hypertension been associated rather than causal (for example if there was a double-headed arrow between age and hypertension, or a third variable between them which caused both hypertension and age).[18]

Therefore in this example the case-control study matched on a non-confounding variable which was associated with the exposure. This resulted in the need for the matched variable to be controlled for, which would not have been required, had the matching not have taken place. It has been suggested that in these situations, either conditional logistic regression or the Mantel-Haenszel adjusted odds ratio be used.[312]

4.3.3 Adjusting for Participation Bias in the Diabetes Data

This method can be applied to the diabetes dataset used throughout the thesis. Let the association of interest be between caesarean deliveries and childhood type I diabetes. Say that control mothers who underwent an amniocentesis were more likely to participate. It is known that the age of the mother can influence whether or not there is an amniocentesis, and also whether a caesarean is required. The causal graph for this situation is given in Figure 4.3.

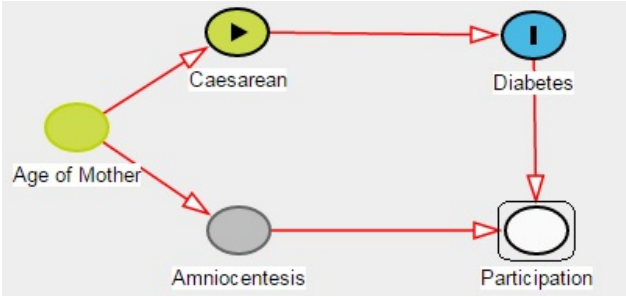

Figure 4.3: Causal graph for the diabetes data example.

Since participation is conditioned on, there is a biasing path opened between caesarean and diabetes, via (age of mother), amniocentesis and participation. This biasing path can be blocked by conditioning on amniocentesis, so the association between caesarean and diabetes can be estimated. Table 4.8 shows the models before and after controlling for amniocentesis. In this instance, the estimate for caesarean does not change by much (from 0.553 to 0.507) but if significance is of interest, the estimate is significant at the 5% level in the unadjusted model, but not in the adjusted model. Both models suggest that diabetes is more likely amongst children

who were delivered by caesarean.

Model	Variable	Coefficient (CI)	p-value	Odds Ratio (CI)
No adjustment	Intercept	-0.582 (-0.776, -0.392)	2.91×10^{-9}	0.559 (0.460, 0.676)
	Caesarean	0.553 (0.042, 1.064)	0.033	1.739 (1.043, 2.899)
Adjusting for	Intercept	-0.633 (-0.832, -0.438)	2.84×10^{-10}	0.531 (0.435, 0.645)
amniocentesis	Caesarean	0.507 (-0.011, 1.024)	0.054	1.661 (0.989, 2.786)
	Amniocentesis	1.208 (0.361, 2.132)	0.007	3.348 (1.434, 8.435)

Table 4.8: Logistic regression model results for diabetes status and caesarean, before and after controlling for amniocentesis. CI = confidence interval.

4.3.4 Critical Evaluation

The method of adjusting for participation bias is an apparently simple method to apply, since it only requires the variable(s) associated with participation to be added to the analysis model, but it does require the selection of appropriate variables. This approach is well established for reducing confounding bias, and therefore the general principles are widely understood. This method was the third most commonly used in the assessment in Chapter 3 and therefore seemingly accepted amongst authors, although the ease and speed of this method are likely reasons why it is a popular choice.

This method is suitable for continuous, binary or categorical variables, and multiple study designs where regression models are used for the analysis. Multiple variables can also be adjusted for simultaneously. However, this method will not be possible if the required variable has not been recorded, such as when the variable is discovered after the study has taken place or if the variable is unknown.

There are instances where using adjustment to reduce confounding bias may result in the introduction of selection bias, which could be larger than the initial confounding bias.[111] Care should be taken when adjusting for variables, and causal diagrams are recommended to check which variables to adjust for and to see the consequences of any adjustments. Online software such as DAGitty[112] are available to assist with this process, as well as a recently developed (2016) *R* package.[318]

There may be multiple plausible causal diagrams, and it is a worthwhile activity drawing each (or at least some if there are many) to see how the analysis and assumptions vary. If the same adjustments would be required for multiple diagrams, this strengthens the reasoning for the adjustment. Causal diagrams also have the ability to incorporate unmeasured variables, even if they cannot be included in the analysis. Their role in the association of interest may be important and may dictate other variables which need to be recorded to conduct the required adjustment. Since a measure of the strength of an association is not included in a causal diagram, an association of zero is permitted and hence allows the diagram to be constructed even if the association between variables is not fully understood. The absence of an association is a strong assumption in a causal diagram[38] and should be used with care. Thus as a general approach, uncertainty of associations should be depicted as an association, which could later be concluded to have a value of zero.

Causal diagrams can be used in other scenarios, such as to determine when complete case analysis provides unbiased estimates. Daniel *et al.* explain how the joint distributions of the exposure and outcome can be estimated without bias when the outcome is MCAR,[38] but not when the outcome is MAR given the exposure. However, in both scenarios the causal effect of the exposure on the outcome can be estimated without bias in complete case analysis.[38] There are many extensions to causal diagrams such as this, but the focus in this thesis is specially for their use with case-control studies.

Causal diagrams could be described as cautious, since they may suggest there is bias when there may not be.[38] This allows them to be generalisable but also requires them to interpreted for the given scenario. One example of this is where a causal diagram suggests bias from conditioning on participation which is affected by the outcome, when the odds ratio used for case-control studies would remove his bias.[38]

It is possible that adjusting for variables in regression models has become common practice and that not all authors consider whether adjustment is required or whether adjustment can lead to other biases. Drawing a causal diagram and considering the effects of any adjustments before analysis would be wise.

4.4 Other Methods

Other methods to account for bias resulting from non-participation were encountered while conducting the literature search for Chapter 2, which are used in the literature less frequently according to the assessment in Chapter 3, hence these methods will be described only briefly, with references included for further reading.

4.4.1 Inverse Probability Weighting

Inverse probability weighting (IPW) assigns a weight to each of the participants so that in the analysis they represent themselves plus non-participants which possess similar characteristics.[111] The weight is equal to the inverse of their probability of participating[111] and can be calculated using external target population data to ensure those participating represent the intended population.[117] This weighting results in a pseudo-population which contains individuals with similar characteristics to the initial population.

4.4.1.1 Critical Evaluation

Since IPW is an approach which essentially removes any missing data resulting from non-participation by replacing missing individuals with similar participating individuals, it has the potential to ensure the confidence intervals remain narrow with respect to the sample size. However, there could be additional uncertainty in the estimate of the odds ratio which should be taken into account during the interpretation of the results.

To conduct the analysis it must be possible to estimate the weights and all weights must be non-zero. If either of these points fail, reliable analyses will not be possible. Members of the population can only be represented if there is a participant with similar characteristics, since weights cannot be applied without a recorded value and so some individuals may not be accounted for. This is particularly likely when individuals refuse to participate for a reason related to a given characteristic, such as refusal on religious grounds. Typically IPW requires some knowledge of the drop-out mechanism and is therefore not as easily applicable to case-control studies where the participation probabilities are usually unknown.[94] When known, the

probabilities are usually much higher for cases than controls,[319–323] resulting in large variability in the probabilities, causing the weighting method to be inefficient.[324,325] Additional data may be required to determine the characteristics of missing individuals[117] and hence calculate the weights. While census data or published literature are useful in some instances, in others these data may be unavailable, particularly in new areas of research or where data relate to sensitive or personal issues. Collecting data from non-participants is another option, but for many studies will not be possible due to uncooperative non-participants, or possibly due to restricted resources.

Weighting methods can be complex and computationally intensive,[94] although they can usually be implemented using statistical software packages.[117] The assumption that the data are MAR may not hold in case-control studies, particularly in studies of sensitive exposures or outcomes such as sexually transmitted diseases or drug use, and since this is an untestable assumption, it may not be valid even when thought to be.

IPW has an advantage over stratification in that it produces unbiased estimates of the exposure-disease association in more scenarios.[111] Generally, stratification is limited since it calculates the exposure-disease association conditional on the stratification variable which causes non-participation, and is hence not suitable when the exposure causes the stratification variable, or when the exposure and stratification variables share a common cause,[111] whereas IPW does not have this assumption.

IPW has been used in the literature specifically to adjust for selection bias in case-control studies[117] and is a flexible approach which can be used in a range of scenarios when participation probabilities can be estimated. This includes in trials with drop-out,[94] since the data consist of a pseudo-population to which standard analyses can be applied. Weighting has the advantage of being a reasonably well-established method with a relatively simple idea, which was developed in the survey literature in 1952.[284] However, it may be more suited to survey data where the weights often vary little between comparison groups, than to case-control data where weights can vary drastically between groups.

4.4.2 Imputation

One idea proposed is to list the basic characteristics of those who have not participated in the study and rather than recruit further individuals, take the data from those who did participate and impute the missing data. Imputation is an approach for entering a value for a specific data item where the response is either missing or incomplete.[326] This idea assumes that at least some information is known about the non-participants to be able to predict the missing values, such as their location, age or sex, and usually that the data are MAR.[38,305] The approach then uses these available data to estimate the missing variables, to complete the dataset. Imputation results in a sample of individuals 'similar' to the original sample, but with increased uncertainty in the results due to the imputation and this is usually reflected in the confidence intervals of any estimates.

Imputation is frequently used in other fields, such as to fill in missing items in survey non-response,[327–330] and justification of this approach for use with survey data has been published.[331] There are also a variety of different types of imputation methods available,[49] including single and multiple imputation.

Single imputation involves replacing the missing value with a value based upon a given rule.[292] Examples include using the mean value for the missing variable, or the last measured value for that variable.[292] This approach has little statistical grounding and can result in bias, plus it does not take into account the added uncertainty from imputing missing values.[292]

Multiple imputation was first introduced in 1978 in the survey literature[290] and has since become a well-recognised approach for missing data, being used widely and included in reviews and texts.[332] Areas of use include observational studies in medical research and implementation is now possible via a range of statistical packages.[332] Multiple imputation imputes the missing values multiple times to create several datasets with the same recorded values but differing imputed values.[292,332] The chosen data analysis is then applied to each of the datasets and the set of estimates combined using Rubin's rules[331] which average over the estimates while allowing for the additional uncertainty in the final estimate.[292,332] Multiple imputation techniques[292] can be used as a method to quantify possible bias resulting from non-participation in case-control studies.[53] Examples[333] for and discussions[334] on using multiple imputation for case-control studies can be found in the literature, as can further details about multiple imputation in general.[292,332]

4.4.2.1 Critical Evaluation

Multiple imputation is an efficient approach[332] which is used frequently for missing data, hence many resources are available, including packages for implementation in statistical software.[335] The general concept seems to be well understood and it has been successfully implemented for survey non-response. Multiple imputation can be applied to most study designs, even if the dataset is large and the missingness complex.[332] It can also be applied easily in conjunction with a sensitivity analysis.[332]

Multiple imputation has the advantage that the imputation model and overall analysis model are separate, therefore a different variable subset may be used to impute missing values and to analyse the data. In some other methods such as regression adjustment, the adjustment and analysis are performed as one step, but multiple imputation allows this additional flexibility. However this extra step may be seen to be time-consuming and deter some researchers.

Choices made during the imputation procedure can vastly affect the conclusions drawn, and hence care needs to be taken when defining steps such as the imputation model.[292] For example, it is not always known that the outcome often needs to be included in the imputation model when imputing predictors, and omission of the outcome can affect the association of interest.[292] Many multiple imputation procedures also assume that data are normally distributed so problems can occur for non-normal data, or when data are binary or categorical.[292] Finally, since the MAR assumption is untestable, it is unknown whether it holds for the variables, yet the suitability of this method depends upon the level of and the patterns in, the non-participation.

Multiple imputation can be computationally intensive when the dataset is large, if there are a large number of variables or if the missingness percentage is high,[292] and then imputation is also known to perform less well.[331] Multiple imputation would be best suited to studies which can reasonably assume the missingness to be MAR and where the missingness is not too high; although the percentage missingness to be classed as ‘not too high’ is arguable and study-specific. Since approximations are used,[292] some algorithms may need to be run repeatedly, which can also add to the computation time.

The entire multiple imputation process can be complicated and restricted space in journals may not allow for such detail. Guidelines have been suggested[292] for details regarding multiple imputation

which should be included in supplementary material. This approach is sensible as it provides the reader with all the required information and allows the research (provided the data are available) to be reproducible. It also enables the reader to judge the integrity of the final conclusions and encourages the researchers to test assumptions and highlight potential pitfalls.

4.4.3 Propensity Score

The propensity score is the probability of the exposure given recorded baseline variables[336,337] and it can be used to match samples using the univariate propensity score, or for multivariate adjustment using the propensity score.[338] The exposure groups are comparable with respect to the recorded variables, conditional on the true propensity score.[336] The true score is often estimated as the predicted probability of the exposure given the measured variables.[336]

Propensity score matching uses the standard idea of matching two groups for comparison, but matches using a single indicator known as the propensity score instead of multiple variables.[310,339] Matching is achieved by pairing a control to a case with a similar propensity score.[310,339] The aim is a dataset comprising of cases and controls with similar characteristics for the variables used to define the propensity score.[339] Logistic regression models are often used to calculate the propensity scores since they do not make assumptions about the distributions of the variables in the model on the dichotomous outcome.[310] A score is calculated for each individual, whether case or control, and seen to adjust for the differences between the groups with respect to the recorded variables[339] and for this reason it is often referred to as a balancing score.[338,339] Matching through techniques such as nearest neighbour matching can be achieved using statistical software packages.[340]

Propensity scores can be used for matching, regression adjustment or stratification,[310,339] with regression adjustment the most commonly found approach in the medical literature.[337] The process of matching pairs cases and controls with similar scores, stratification divides the individuals intro strata dependent upon their scores, and regression adjustment techniques use the propensity score as a variable in the model or as a weight.[339] Where regression adjustment is used, the estimate of the exposure remains the same as if all the variables included in the formation of the propensity score were in a regression model with the exposure. However the propensity score approach enables the propensity score model to be (possibly) over-parametrised and include

interaction and higher order terms, and allows the model with the outcome and exposure of interest to be simple and hence model-fit tests can be performed more easily.[310]

Propensity score matching has been used in the analysis of case-control studies,[336,341–345] but not frequently.[2] Propensity scores are more often used to match cases and controls before study commencement than during analysis.[341] When used in observational studies, the role of propensity scores tends to be to reduce bias and increase precision.[310]

4.4.3.1 Critical Evaluation

The true propensity score is usually unknown for observational studies[336] and hence must be estimated. However, propensity scores cannot be estimated for individuals who have missing data for the variables needed in the propensity model, so some individuals may be excluded on these grounds. In addition, if those with missing data differ from those with recorded data, additional biases may be incurred. While propensity scores can be used to balance comparison groups with respect to measured variables, they are unable to balance unmeasured variables, which is an advantage of randomisation which propensity scores aim to replicate.[340,346] However, by estimating the score there is the advantage that it accounts for some variability occurring by chance, unlike the true score which would only account for systematic bias.[336,346]

Propensity scores are particularly useful for matching multiple controls to a case and a simulation study has also shown propensity score matching to be superior to other forms of matching.[346] Propensity score matching permits a large number of variables to be adjusted for at once using techniques such as nearest neighbour matching, whereas methods such as stratification and variable adjustment can be limited by the number of variables due to concerns of the sample size.[310,339] Implementation can also be through a statistical software package, offering an alternative for matching with relative ease.

Propensity scoring assumes that factors which predict participation are similar for case and control groups, which is unlikely, and it requires the correct model-specification when estimating the propensity score.[341] The inclusion of more variables (possibly more than needed) to the propensity model ensures the comparison groups are similar[347] therefore if there is doubt as to which variables to include, adding more can be beneficial. This approach may therefore be suitable for large

studies which have recorded information on a range of variables for calculation of the propensity score. Tests of the groups should be performed before and after matching to determine whether the propensity score approach has made the groups more similar as intended, or sensitivity analyses could be used to determine their robustness to variation in the propensity score model. However these options involve additional analysis steps to a method which is already more difficult to apply than others and hence may be undesirable, especially when other suitable methods are usually available.

The use of propensity scores with case-control data has not been well-studied[336] and hence there is little guidance on their suitability or application. Propensity scores, while useful for matching or adjustment in cohort studies, are less useful in case-control studies, where they are more complicated to apply and less accurate than other available methods.[336] One problem is that the probability of selection or participation in a case-control study for the control group is often unknown and so must be estimated.[336] Propensity scoring should therefore not be the preferred choice to adjust for participation bias in case-control studies and may be best suited to cohort studies and the reduction of confounding rather than participation bias, or to form matched case-control studies from cohort studies.[339] Its lack of presence in the epidemiology literature (see Chapter 3) may reflect these factors.

4.4.4 Related Methods

The Heckman correction was introduced in econometrics in 1979 and is a two-stage method to adjust for bias arising from non-randomly selected individuals.[286] Heckman was subsequently awarded the Econometrics Nobel Prize in 2000 for this work. Briefly, this approach works by developing two models, (i) a regression model considering mechanisms determining the outcome and (ii) a selection model.[348] In this way, it is similar to propensity scores as in §4.4.3, and hence for case-control studies is likely to suffer from similar limitations as discussed in §4.4.3.1. A Web of Science[130] search suggests this approach has not routinely been used in conjunction with case-control studies. The topic search terms used were “Heckman” plus “case-control” or “case-control” and only three results were returned on 19/02/2016. The first was a presedential address,[349] the second was the reanalysis of a prospective case-control study[350] and the third referred to a different author named Heckman.[351]

The bias breaking method was introduced in 2009 specifically to adjust for selection bias in case-control studies.[352] Briefly, a model is constructed which incorporates additional variables to adjust for selection bias and which can be combined with study data to improve inference.[352] This is achieved by defining a "bias breaking" variable which separates the risk factor from the selection criteria, provided such a variable exists. In addition, this approach requires data to be available for the bias-corrected estimate of the distribution for this variable. These requirements will not be applicable to all studies and the method is more complicated than other approaches to implement, hence this approach is unlikely to be suitable for common use with case-control study data. While the authors demonstrate both the presence of bias breaking variables in epidemiology studies and the application of the method, its use in the medical literature is limited. A Web of Science[130] topic search on 19/02/2016 using the terms "bias-breaking" or "bias breaking" returned just six results; two of which were published before this method was proposed[352] and referred to the breaking of atomic chains[353] and diffusion.[354] For the remaining three (since one was the method itself) one discussed oxidation[355] and the other two were conference abstracts from the same research team.[356,357] Its apparent lack of use in the literature has led to just a brief summary here. The same author group also published an article in 2011 stating that sensitivity analyses were the only solution to selection effects,[305] hence possibly dismissing the bias breaking method.

4.5 Guidance Tool for Researchers

With several methods available to reduce participation bias, researchers may be deterred from implementing a method for fear of choosing an unsuitable approach. In addition, implementation of a method to reduce bias may be viewed as an undesirable feature which could lead to criticism of their study or potentially reduced chances of publication. In instances such as these and often with time-constraints to adhere to, investigations into participation bias may not be prioritised.

Each method in §4.1–§4.4.3 has its own requirements and assumptions; some require external data,[284,352] some require non-participant data,[292] and some assume the variable associated with participation is known and measured.[111] Here, a straightforward flowchart to aid the selection of an appropriate method is provided and three examples are presented. Table 4.9 includes the data requirements for each method which are grouped into three categories; only one category is needed but some methods have a choice of category.

(Alphabetically ordered) Methods	Participation variable	Population data	Non-participant data
Imputation			✓
Propensity Score	✓		
Regression Adjustment	✓		
Sensitivity Analysis		✓	✓
Stratification	✓		
Weighting		✓	✓

Table 4.9: The required data to implement the methods.

- **Participation variable:** It is assumed that the variable associated with participation is known and can be recorded during the study. Proxy data are permitted.
- **Population data:** External information (such as census data or hospital registries) is available, which is assumed to be unbiased with respect to participation and to represent the population of interest.
- **Non-participant data:** Basic characteristics of those unwilling to participate are available, either taken from the individual directly or from external sources.

Although the three data categories in Table 4.9 are sourced differently (from participants, the population and non-participants) there are relationships between them. For example, if the target participants are representative of the population of interest, and relevant information is known for all non-participants, then the non-participant data in conjunction with the participant data, could be used to approximate the population data. Therefore under certain circumstances it may be possible to use a different column from Table 4.9 for the data source, other than the one(s) ticked. Table 4.9 and the consequent flowchart tool can be interpreted as a generalisation or guide, which can be adapted by the researcher if these conditions are met.

Figure 4.4 gives an example of a flowchart (which begins with the square towards the top-left corner, shown using a bold outline), based on the data from Table 4.9 which could be used by researchers to shortlist methods for further investigation. Researchers could extend this flowchart to meet their specific needs for the variables or data they encounter, or alternatively disciplines could form a subject-specific chart to which new methods could be added over time.

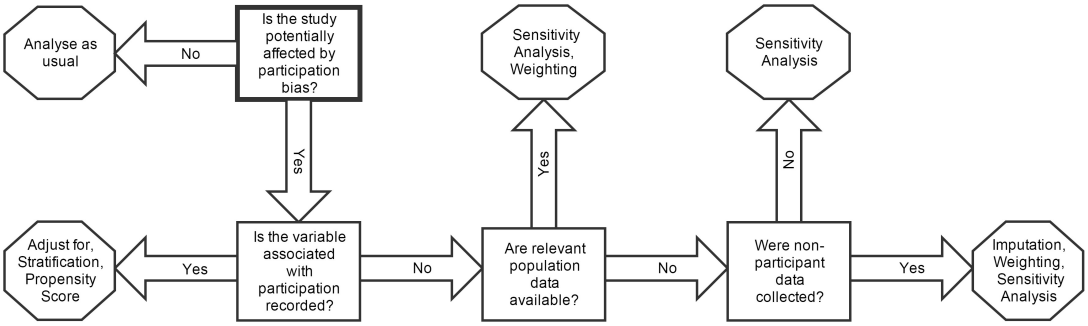

Figure 4.4: A flowchart tool to aid the selection of a suitable method to reduce bias.

4.5.1 Examples

Three examples of hypothetical studies utilising the flowchart in Figure 4.4 follow. To answer the first flowchart question, the requirements for bias stated in §2.3.4.1 must be known. After completion of the flowchart, it is the responsibility of the researcher to consider each method in turn to see which is most suitable for their study; all assumptions must hold.

4.5.1.1 Example 1

A randomised controlled trial (RCT) is conducted for a new eczema cream; sufferers are recruited and randomly allocated to the drug or placebo group. The new cream produces some unexpected side-effects and some participants in the drug arm suffer from rashes. Half of the participants in the drug arm withdraw from the study as they decide that their eczema symptoms are preferable to the side effects.

- Is the study potentially affected by participation bias? The association of interest is between the new cream and the severity of the eczema symptoms. For bias to occur both the treatment and eczema symptoms need to influence participation. The side-effects from the cream causing withdrawal mean that the treatment does affect inclusion in the analysis. However, eczema suffers were randomly allocated to either the treatment or placebo group, so the severity of eczema symptoms was balanced between the two groups and therefore the severity of the symptoms did not affect participation. Since only the treatment group and *not* the severity of the symptoms affects participation, bias is not a problem here and the

results can be analysed as usual. This example re-emphasises that RCTs are less likely to be affected by participation bias than other study designs, as was discussed in §2.3.4.2. As an aside, RCTs can conduct an intention to treat (ITT) analysis, which is the assessment of individuals in a trial regardless of whether they left the study or did not use the treatment as instructed,[358] which could be applicable here.

4.5.1.2 Example 2

A study investigates the association between coffee and migraines, with postal surveys sent randomly to households in the United Kingdom (UK) with a return envelope. In addition to migraine and coffee questions, the survey requires basic demographic data such as sex, age, general location and employment status. Migraine sufferers may be more interested in the study and hence more likely to respond. Previous studies have shown that older individuals are also generally more likely to participate in surveys.[135–139] Finally, let age and coffee consumption be positively correlated.

- Is the study potentially affected by participation bias? Coffee is the exposure and migraines are the outcome. Since older individuals and migraine sufferers are more likely to return the survey, and older individuals are more likely to be coffee drinkers, participation is affected by both the outcome and a cause of the exposure and so bias is possible after conditioning on participation.
- Is the variable associated with participation recorded? Migraine occurrences and age are recorded and affect survey returns. However, only variables other than the exposure and outcome can be used in the analysis to control for bias; age in this instance. The following methods can be considered:
  - Adjust for the variable associated with participation; Age can be added to the analysis, for example as a variable in a regression model between coffee and migraines.
  - Stratification; The analysis can be conducted within age strata; for example by analysing in age groups of ten years, to reduce the effect of age on participation.
  - Propensity score; The propensity score can be calculated using all the variables and included in the analysis.
- The assumptions for these three methods should be checked to assess suitability.

4.5.1.3 Example 3

A UK case-control study investigates the association between excessive alcohol consumption and brain tumours. Researchers attempt to recruit cases who have brain tumours and controls who do not and collect data regarding alcohol consumption retrospectively. The participants and interviewers are blinded to the exposure to reduce the effects of other biases such as interviewer bias. Blinding is in the form of an extended questionnaire with questions relating to several possible exposures such as alcohol, smoking, mobile phone use, exercise routines and family history. Some participants intentionally avoid questions such as those to which they have undesirable answers. For example, some heavy smokers may ignore the question regarding cigarettes, those who do not exercise frequently might miss the question regarding exercise, and frequent drinkers may be more likely to avoid the question about alcohol consumption.

Let data from the questionnaire be available, along with a national UK brain tumour database. The Office for National Statistics (ONS) also records data regarding adult drinking habits.[359]

- Is the study potentially affected by participation bias? It is well-documented that cases are more likely to participate than controls[17,141] and individuals were selected based upon their disease status, so the outcome is affecting participation. Alcohol answers are being recorded only for those willing to declare their consumption levels; with those who consume amounts not deemed to be excessive being more likely to participate, so inclusion depends upon exposure. Since only those who are willing to participate in the study *and* who answer the question regarding alcohol consumption are used in the analysis, participation is conditioned on and so bias is possible. This example has shown how participation bias can occur in case-control studies as discussed in §2.3.4.2.
- Is the variable associated with participation recorded? Only the exposure and outcome are recorded and affect participation, hence methods using the variable associated with participation to reduce bias are not suitable here.
- Are relevant population data available? The national database for brain tumours and ONS data for drinking habits are available so the following methods can be considered:
  - Sensitivity analysis; Population data could be used to conduct a sensitivity analysis.
  - Weighting; This is possible, but if there is the extreme scenario where there are no heavy-drinking participants who answer the question about alcohol consumption,

then a weight cannot be applied to this category and the weighing method would be unsuitable. Alternatively, if there are a few heavy-drinkers who answer the question regarding alcohol consumption, there may be large variability in the weights causing the weighting approach to be inefficient as discussed in §4.4.1.1.

- Therefore, two methods may be suitable, but the assumptions of both must be checked before an approach is selected.

4.6 Summary

4.6.1 Links Between Methods

Many of the methods in this chapter originated in the survey literature and have been either directly transferred to the medical literature or adapted for use with medical data. Examples include multiple imputation,[290] inverse probability weighting[284] and post-stratification.[94]

The different approaches need not be used independently, since they can complement one another. For example, authors may use causal diagrams to identify bias, but IPW to adjust for the bias.[117] Plus there are sensitivity analyses which use multiple imputation, and it is becoming increasingly popular to suggest the use of a sensitivity analysis alongside imputation.[292,332] One author correctly highlights that propensity scoring is not a replacement for other approaches, yet is available to complement them.[346] Propensity scores can be used in regression adjustment, as a weighting, or alternatively quintiles of propensity scores can be used to stratify analyses.[336,339,341] Comparisons may be made before and after adjustment and this could be viewed as a form of sensitivity analysis. In addition, the sensitivity of odds ratios under different weightings in IPW could be investigated. If the OR is very sensitive, then the study conclusions may not be robust, whereas if the weightings can vary greatly while maintaining the same study conclusions, the findings may be deemed to be more reliable.

Combinations of these methods can also occur, which can increase efficiency. For example, stratification on the propensity score can be used to balance the distribution of covariates among groups, without the need for an exponential increase in the number of strata[360] as in standard stratification. However, while using a combination of methods may be useful, there is also the need for the data to satisfy both sets of assumptions from the two (or more) methods being used.

Hernan *et al.* mention equivalent conditioning, matching or regression adjustment techniques, showing a similarity between these methods.[111] Some of the methods have also been derived from one another, for example, the bias breaking method[352] is a form of post-stratification, which is a type of stratification. There are also overlapping themes, such as splitting the data into smaller, more similar groups, which can be compared more easily, or weighting the data with the aim of making the sample data more similar to the population data.

4.6.2 Overview

Both qualitative and quantitative approaches to non-participation occur in the literature, with quantitative including the methods listed here, and qualitative often offering a discussion about the possible effect of the bias. Examples of both approaches can be found in the articles included in the assessment in Chapter 3. To determine the suitability of a method, several aspects should be taken into account such as the definition of participation bias, the study design, the available data and the required summary value.

Jiany *et al.* have stated that "all methods for correcting for non-response either require some sort of information about at least some of the missing units or essentially assume the problem away",[24] which may be true. For example, IPW assumes that the missing individuals possess similar characteristics to the participants, and multiple imputation assumes that the missing values for non-participants can be estimated from the recorded values for similar participants. Sensitivity analyses are slightly different, since although they require some assumptions to be made, such as the highest and lowest possible values for missing data, it is often an exploratory analysis rather than requiring specific values. In this respect though, it may not be classed as 'correcting' as quoted.

Although each of the methods described and evaluated in this chapter are designed to account for participation bias, they do so by using different techniques and assumptions. Therefore, a method which may be optimal for one study may not be suitable for another and hence each study should be considered on an individual basis. Also, throughout this chapter *the* variable associated with participation has been referred to, whereas in practice there may be multiple variables. The number of variables which need to be adjusted for will also help to determine a suitable method.

The study design will also affect the suitability of a method. Propensity scores may be most useful for matching in cohort studies, but least useful for case-control studies, where sensitivity analyses may be most appropriate. Provided assumptions hold and there are sufficient data, weighting, regression adjustment, imputation and stratification may also be suitable. To aid selection, a user-friendly flowchart tool has been provided, which can be adapted for particular research areas or depending upon the data resources available. The demonstration of this versatile tool through examples aimed to increase the consideration of participation bias and consequently the implementation of appropriate methods.

Chapter 5

Chain Event Graphs for Missingness in Case-Control Studies

In the previous chapter, methods to investigate and reduce participation bias were described, demonstrated and critically evaluated. The requirements and assumptions of these methods were included and it was found that several of the methods required the data to be missing at random (MAR). A graphical approach suitable for incorporating missingness resulting from non-participation in case-control studies, or exploring the missingness mechanism requirement for these methods was sought, and research of the literature led to chain event graphs.

The chapter continues with a recap of graphical models, which were introduced in Chapter 2, and highlights the limitations of the models seen thus far, before introducing chain event graphs in §5.1. In §5.1.2 a literature review is conducted which concludes that case-control studies have not before been used with chain event graphs. In §5.2 a description of how chain event graphs are formed is given and illustrations are provided. The remainder of the chapter describes how chain event graphs can be used to investigate missingness resulting from non-participation in case-control studies, forming the main focus of this chapter.

Illustrative examples of how chain event graphs can be used to investigate missingness and incorporate missing data are given in §5.3. The real diabetes data are then used with chain event graphs in §5.4, where extra variables are introduced, one of which has missing data. The missingness mechanism is investigated and suggestions are made for the missing values in the

variable with missing data. The chapter concludes with §5.5 by describing how this analysis can aid the selection of a suitable method to reduce participation bias, and critically evaluates the method overall.

The main aim of this chapter is to use chain event graphs as a tool to investigate the missingness within a case-control study, and to use the findings to select an appropriate method to reduce any bias resulting from non-participation. The findings here will be linked to the flowchart in §4.5 which summarises methods to reduce participation bias. Since chain event graphs have not been used before with case-control study data (as will be shown in §5.1.2), chain event graphs have also not been used as a tool by which to investigate non-participation in case-control studies and to guide the next stage of analysis.

5.1 Introduction to Chain Event Graphs

Chain event graphs (CEGs) are a graphical modelling technique for discrete probability models which were developed in statistics and artificial intelligence. Introduced in 2008, they are a form of directed graph that can be used to order and equate combinations of variable categories with respect to their probability of an outcome of interest.[21,361–365]

CEGs are an extension of Bayesian Networks,[366,367] hence can incorporate prior information into the analysis. Therefore, population level data, data from previous studies, or expert opinion, can be incorporated. CEGs are also able to predict the missingness mechanism when data are missing from a study, and suggest the likely values for the missing data.[365]

Formal, mathematical definitions for CEGs are provided in the literature.[21] Here, the definitions are given using an alternative and less formal explanation with the intention of being more accessible, and in the hope that these graphs will be adopted by the medical community. The focus here is to use the published theoretical findings to apply and adapt (see Chapter 6) CEGs for use with case-control data, particularly with respect to missing data resulting from non-participation.

5.1.1 Recap of and Comparison With Other Graphical Models

Chain event graphs (CEGs) form part of a family of probabilistic graphical models (PGMs). This family includes Bayesian Networks (BNs),[366,367] acyclic probabilistic finite automata (APFAs)[368] and chain graphs,[369] some of which have been successfully used with medical data.[370,371] Recall from Chapter 2 that graphical models are statistical models represented concisely using a graph. Graphical models offer an intuitive data representation and a means of communicating complex statistical models to medical experts in a more easy-to-interpret form. Graphical models may be used for data representation, inference or learning. Commonly used graphical models include directed acyclic graphs (DAGs) (§2.1.1), directed graphs and chain graphs.

Probability and decision trees are another form of model which can be used to display data. Theses graphs have a natural ordering from start to finish and tend to be utilised when there are asymmetric dependencies between variables. Trees are frequently used as an interim step before the graphical models described above, when statisticians are conversing with an expert in a given field regarding the variables in a dataset requiring analysis.[21] Event trees describe a sequence of events which may occur and give each of the outcome options of an event as an edge. The edges of these trees may be labelled with conditional probabilities, and in this instance, are defined as probability trees. The label is the probability of the next event given the previous events.

Each of these graphs use vertices or nodes to represent variables in a dataset, and edges to connect the vertices, which display conditional dependencies between the variables.[8] Conditional independence is where two variables are independent given a third variable. Therefore, information about the third variable and one of the first two variables, provides no information as to whether the other of the first two variables has occurred. The general consensus is that if there is doubt as to whether or not an edge should exist, it should be included, since the probability along that edge can be close to zero.

The information generally required for a graphical model is the list of variables which act as vertices, the set of conditional independence statements, and the conditional probability vectors. Without the final component of conditional probabilities, qualitative analysis only may be performed. This information is usually gathered and compiled with an expert in the field, to ensure plausibility of the associations and variables used in the analysis. The conditional independence

statements often do not result in a unique graphical representation, so this expert guidance is usually invaluable.[8]

It is generally agreed that inferring causality from a graphical model requires some assumptions which cannot be derived from observational data. Pearl, a key author in causal modelling, has stated just this[20] and Holland agrees by stating that there is "no causation without manipulation".[372] Therefore, causal graph findings should generally be interpreted as an association rather than a cause, and this will be adhered to while using CEGs.

Medical data, including data from case-control studies, can be asymmetric, whereby one event can have an impact on the options for subsequent events. Many of the current graphical models such as Bayesian Networks, do not allow for such asymmetry and hence are not suitable for all data structures. For example, it may be that given one event a second event is impossible, such as male gender followed by having breastfed their newborn child, or that during the recruitment of case-control studies, it may be that reminders to participate are sent to all those who have not yet responded, and these invitations will not be sent to participants. It is preferable that these forms of asymmetry are incorporated into the study analysis and can be when using CEGs.

APFAs are useful when one wishes to consider the dependence structure between variables, and when this structure is expected to vary over time, such as in longitudinal data.[373] The use of APFAs generally has been in handwriting recognition or with speech data,[368,374] but has also been with DNA data.[374] Case-control studies do not analyse changes over time and hence AFPAs are likely to be of less use here than other graphical models. APFAs may be more suited to cohort studies, since they represent longitudinal data and changes are expected through time in the dependence structure between variables. While CEGs can be used for cohort studies,[375] they can also be applied to data which are not longitudinal.

Chain graphs permit both directed and undirected edges, and the DAGs discussed in Chapter 4 are a special case (or subgroup) of chain graphs which allow only directed edges. As shown in Chapter 4, DAGs offer a useful means by which to explore bias resulting from non-participation or confounding, and they can be used to direct the adjustment of variables in regression, but alone they do not offer a form of analysis for non-participation. Therefore, while useful for identifying bias, there may be other graphical models more suited to the analysis of case-control studies which may suffer from non-participation. For each of these reasons, CEGs are the focus in this chapter.

5.1.2 Literature Search

A literature review was conducted to ensure CEGs had not been used before for case-control studies, in particular to investigate missingness from non-participation. Four searches were conducted on Monday 7th September 2015, using four main databases: Web of Science,[130] PubMed,[376] Scopus[377] and Google Scholar.[378] Each returned article was checked to determine whether CEGs had been used with case-control studies, and no evidence of such use was found. Full details can be found in Appendix D.1.

The literature search showed that there have been relatively few CEG publications, since many of those listed are presentations, technical reports and PhD theses on the same topics. The main topics published have been introductory work, causality, model selection, plus some applications. There has also been the introduction of CEGs for informed missingness and binary outcomes which have not yet been extensively applied, but both prospects have been demonstrated using a cohort study.[8]

Case-control studies have a binary outcome, namely the case or control status of the participants. This study design is also known to suffer from missingness either through non-participation or through the refusal to answer particular questions or to be involved with certain activities such as a face-to-face interview. Therefore, work published for binary outcomes and for data with informed missingness should be relevant to case-control studies, yet according to the literature search here, these approaches have not yet been applied to case-control data. It is of course possible that CEGs and case-control data have been used together and not been returned through the literature search. However, it appears that almost all the research has stemmed from one research group who upload their work, including technical reports and conference presentations, to an online repository which is accessible via the Internet. Therefore it is likely this application of CEGs and case-control data has not been missed. The remainder of this chapter aims to fill this gap in CEG usage, as CEGs can incorporate missing data, and non-participation can be viewed to be a form of missing data.

Reasons for why CEGs and case-control study data have not yet been used together were considered. However, since CEGs have already been shown to be suitable for medical data in the form of cohort studies, and for binary outcomes[8] in the form of ordinal CEGs (see §5.2.7 which follows), then the format of case-control data will be valid. In addition, case-control data

can suffer from non-participation and this limitation in the data can be addressed using the methods proposed to assess informed missingness.[365] Since CEGs are suitable for exploring missingness mechanisms (§5.3) and can make suggestions for the values of the missing data (§5.3.4.2), they will be applied in this chapter in the context of data missing due to non-participation. Knowledge regarding the missingness mechanism, while useful in itself, can also be used to determine the suitability of a method to reduce participation bias if needed, as will be shown in §5.4.4.2.

5.2 Formation of Chain Event Graphs

This section will describe, with illustrative examples, how CEGs are formed. No data will be used, but example trees and CEGs will be shown to demonstrate the steps needed to form a CEG, since CEGs are a relatively new approach and not widely used. The algorithm required when using real data is described in §5.2.2.1 and §5.2.2.2. Real data examples, using the diabetes data from Appendix A, will be given in §5.4.

5.2.1 The Tree

Chain event graphs are defined from finite probability trees, with the vertices in the tree arranged in a logical ordering, often chronological. An example of a tree is given in Figure 5.1, which shows hypothetical measures in chronological order through the lifetime of an individual; birth weight, height, blood pressure and whether or not they have kidney disease. Figure 5.1 also shows examples of some of the terminology of a tree. The order of height and blood pressure could be swapped, but since the height was likely reached before the blood pressure, this order has been selected. The variable of interest, which may be a disease outcome or a dependent variable, should be placed at the final stage of the probability tree.

If there are different plausible orderings of the variables, it is possible to trial the different orderings to test the effect on the CEG, but in many instances there will be a natural ordering which should be adopted. A chronological ordering is used, not to suggest causality, but since decisions or characteristics earlier in life, may restrict the options for later variables, which the tree can reflect. The variables in Figure 5.1 confirm that the tree (and subsequent CEG) cannot be interpreted causally, since although hypertension is thought to cause kidney disease, kidney disease can also

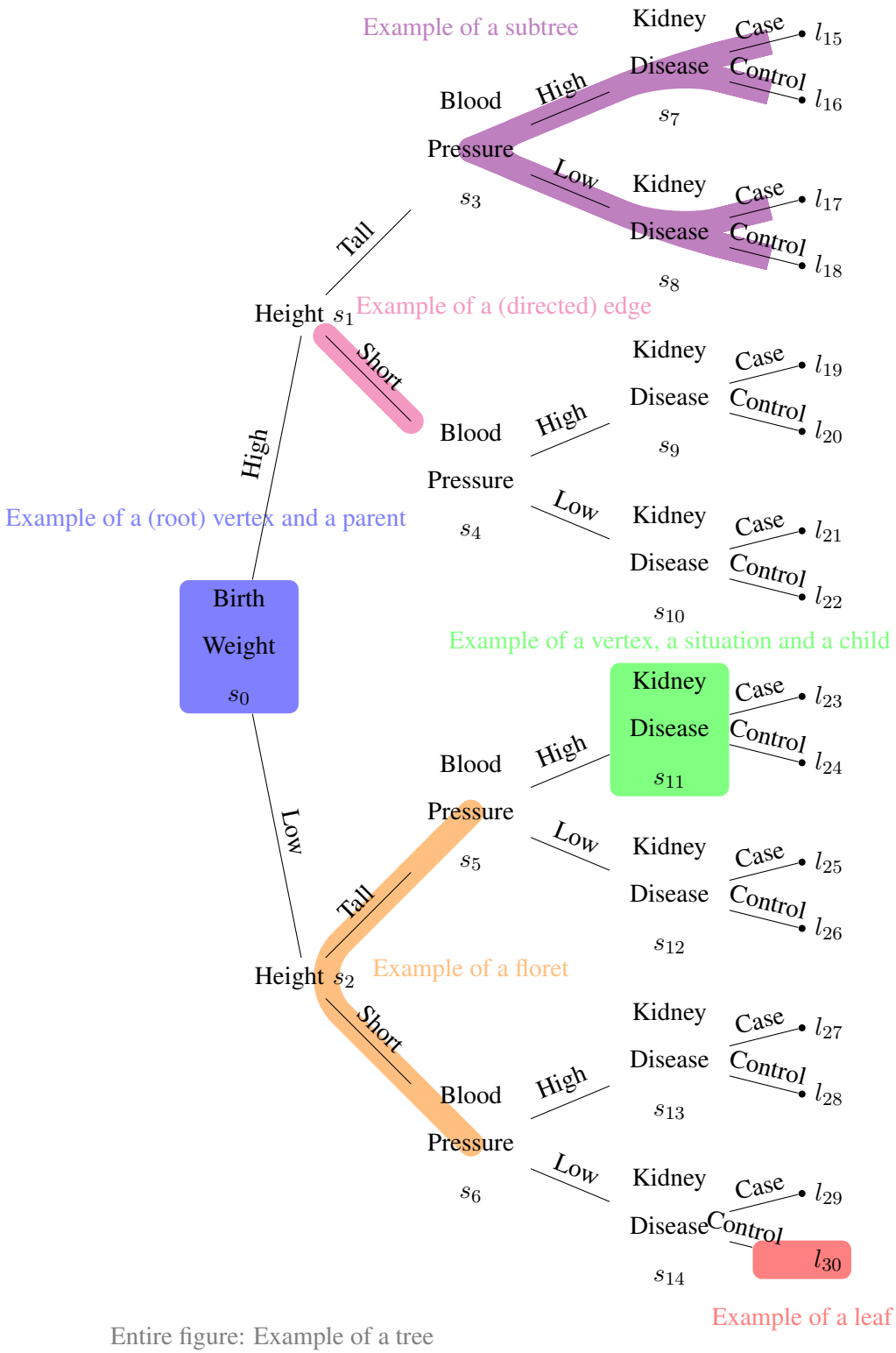

Figure 5.1: Tree for the hypothetical example with four variables.

result in hypertension. Therefore the tree is merely a display of the data, with the outcome of interest as the final variable in the tree. The tree is used as a intermediary step in the process of CEG construction, and conclusions are not draw directly from the tree here.

The vertices are labelled, starting with s for the situations and l for the leaves, as shown in Figure 5.1, and edges show the binary variable categories; low/high, tall/short or case/control. Each edge from a situation shows the possible subsequent steps which may be taken by an individual given they are at that situation, and a series of these steps forms a path. When the probability tree is converted to a CEG, these paths and orderings are maintained so no information is lost. The paths through the tree in Figure 5.1 allow every combination of the four variables, assuming all combinations are plausible. Consultation with an expert is recommended to ensure the variable ordering and possible paths are sensible, and trees may be easier to achieve this than other approaches such as models which can be more difficult to interpret.

Since the variables in the tree and corresponding CEG have categories which form the edges leaving a vertex, continuous variables must be categorised, although several categories may be used to approximate continuous data. The categorisation of continuous variables must be supported by clinical (or other) reasoning for the choices of the number of categories and the position of the cut-off points. Since the categorisation of continuous variables is an undesirable step in the formation of a CEG, this approach may be best suited to studies which have (i) categorical variables, (ii) variables with important clinical cut-off values, or (iii) studies which will be used in conjunction with a traditional analysis which can accommodate continuous variables. The use of CEGs here is not intended to replace analyses which produce OR estimates and which can accommodate continuous variables, but is instead intended as an exploratory tool which allows easier communication with clinical experts and which can be used to investigate missingness resulting from non-participation.

Each floret has an associated random variable which describes each of the children of the situation. The number of edges and hence the number of values in the random variable is equal to the number of children possessed by the situation. The initial tree has each vertically aligned floret representing the same options, but it is possible that these options have different probabilities associated with them, which correspond to the path already taken. When the event tree has conditional probabilities added, it is defined as a probability tree. The probabilities associated

with each floret are written as

$$\pi_{s_i} = (\pi_{s_i 1}, \pi_{s_i 2}, ..., \pi_{s_i m_{s_i}}), \tag{5.1}$$

where $\pi_{s_i k}$ is the probability of taking the kth edge from situation s_i. The sum of the probabilities along the m edges must equal one. Hence the conditional probabilities show the distribution of the random variable associated with the floret. For Figure 5.1 to be a probability tree, conditional probabilities must be assigned to each floret, denoted by π with subscripts to describe the corresponding situation and edge. For example, the conditional probability of situation three, edge two, which relates to low blood pressure conditional on having a high birth weight and a tall height in Figure 5.1, can be shown using $\pi_{s_3 2}$, i.e. the second edge from s_3.[365]

5.2.2 Staged Tree

Next there is the concept of a stage. Two situations are said to be in the same stage if and only if the topology of their associated florets is the same and the probability distributions associated with these florets are the same under a bijection. The stages are determined using an algorithm which will be described in §5.2.2.1. When two (or more) situations are in the same stage, they are assigned the same colours for their edges.[21] Trivial stages, where the stage consists of only one situation, are left uncoloured. Once the tree has been partitioned into stages, there is a set of stages formed, represented by $U(T)$. It follows that each situation in a stage has the same number of edges leaving it. A staged tree is simply the probability tree with the colours assigned according to the stages, as illustrated in Figure 5.2, which uses example colours assuming the algorithm in §5.2.2.1 has been run. From Figure 5.2 the stages, U, are,

$$u_0 = \{s_0\}, u_1 = \{s_1\}, u_2 = \{s_2\}, u_3 = \{s_3, s_4, s_5\}, u_4 = \{s_6\}, u_5 = \{s_7, s_9, s_{12}\},$$
$$u_6 = \{s_8, s_{10}, s_{14}\}, u_7 = \{s_{11}, s_{13}\},$$

since the edges leaving these vertices are shown with the same colour. Vertices whose edges are assigned the same colour are similar enough with respect to the distribution of the current variable such that they can be grouped. Therefore, in Figure 5.2, those who had a low birth weight and who are currently tall (s_5), are thought to have a similar blood pressure distribution to those who were born with a high birth weight (s_3, s_4), as shown using the green and red edges. Those with a low

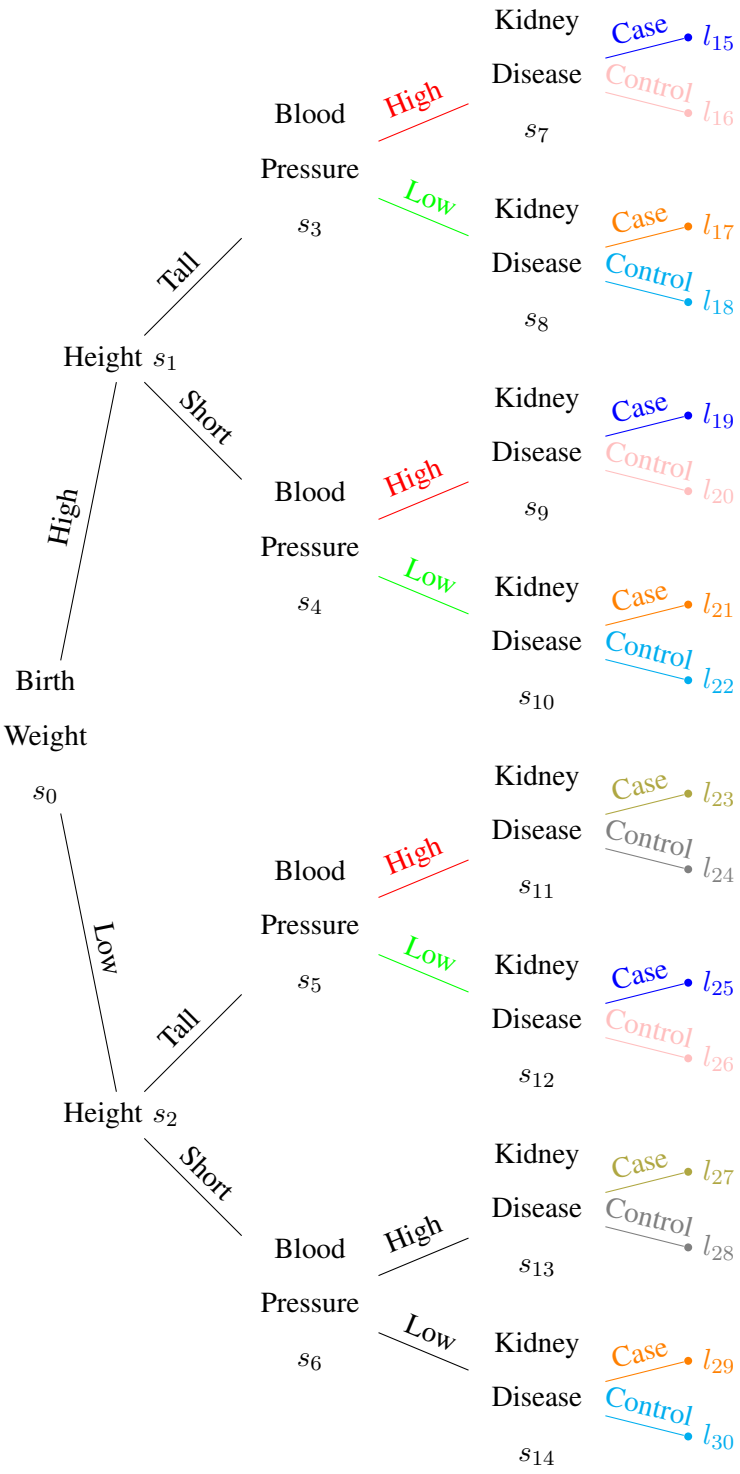

Figure 5.2: Staged tree for the hypothetical example with four variables.

birth weight who are currently short, are thought to have a different blood pressure distribution, hence the edges from s_6 are not coloured red and green.

Another concept, which can result in a finer partitioning of the vertices than stages, is that of positions. Two situations are said to be in the same position if and only if the topology of their subtrees is the same and the probability distributions of corresponding florets in the subtrees are the same under a bijection. Therefore the edges and colours between the subtrees from the situations must correspond. With stages, just the colours of the current edges must match to be in the same stage. However for positions, the entire subtree must have colours which match. Therefore in Figure 5.2, s_3 and s_4 are in the same position as their subtrees are the same colour, whereas s_5 which was in the same stage, is not in the same position as its subtree differs. When a situation leads to a leaf, the definition of a stage and a position are interchangeable. Elsewhere, the positions could lead to a finer partitioning of the vertices than the stages, since vertices may be in the same stage yet not in the same position. The set of all positions is represented by $W(T)$.

From Figure 5.2 the positions, W, are,

$$w_0 = \{s_0\}, w_1 = \{s_1\}, w_2 = \{s_2\}, w_3 = \{s_3, s_4\}, w_4 = \{s_5\}, w_5 = \{s_6\},$$
$$w_6 = \{s_7, s_9, s_{12}\}, w_7 = \{s_8, s_{10}, s_{14}\}, w_8 = \{s_{11}, s_{13}\},$$

resulting in a finer division of the vertices than stages.

5.2.2.1 The Bayesian Agglomerative Hierarchical Clustering (AHC) Algorithm

The following description is taken from Freeman *et al.*[379] and is concisely summarised here. The Bayesian agglomerative hierarchical clustering algorithm (AHC) is used throughout this chapter to identify vertices in the event tree which are in the same stage and the R[380] code used[8] to implement the algorithm is provided in Appendix D.2. Very simply, the Bayesian AHC algorithm is a *clustering* algorithm which starts with all vertices in the tree from a given variable separate and merges (or *agglomerates*) them, and is *hierarchical* since clusters have sub-clusters, which in turn have sub clusters and so on. The *Bayesian* element allows prior knowledge to be incorporated into the algorithm. More formally, the AHC algorithm is a local greedy search algorithm for finding the maximum a posteriori CEG. The algorithm starts with the finest partition of the vertices in the tree and seeks to combine vertices at each iteration which will result in the highest scoring CEG.

An example of the output from this algorithm when applied using R^{380} to the diabetes dataset is provided in §D.4 and the steps of the algorithm are as follows,

1. The event tree is used to form the initial CEG, C_0, which is identical to the tree except the leaves are collapsed to one terminal vertex. An initial score for the CEG is defined as the logarithm (log) of the posterior probability of the CEG given the data. The score is calculated using Bayes' theorem as the sum of the log of the prior probability, the log of the marginal likelihood of the model and a constant which does not depend upon the CEG. This score is used to search over the set of CEGs for the model which best describes the data.
2. For each pair of situations in C_0 from the same variable which have the same number of edges, calculate the log of the ratio of the scores with the score for C_0 as the denominator, and with the numerator as the CEG formed by grouping the pair of situations into the same stage and keeping all other situations in separate stages. Only do not calculate this value if the prior probability of the new CEG is zero.
3. Let C_1 be the CEG which maximises the ratio of the scores from step (1).
4. Next calculate the ratio of the scores,with the score for C_1 as the denominator and with the numerator as the CEG formed by grouping pairs of stages from C_1. Again do not calculate this value if the prior probability of the new CEG is zero. Record C_2 as the CEG which maximises the ratio of the scores.
5. Continue this process for C_3, C_4, etc until the coarsest partition, C_∞, has been achieved.
6. Select the CEG from C_0 to C_∞ which has the highest score. This is the maximum a posteriori (MAP) model.

With the log of the posterior probability of the CEG model C as the score for the CEG, the act of searching over the set of candidate models for the CEG with the highest score is equivalent to seeking the maximum a posteriori (MAP) model. Searching over *all* possible models can be very time consuming when there are multiple variables and variable categories, hence an algorithm for efficiently searching over the model space for the MAP CEG is desired. A disadvantage of this approach is that there is no way to ungroup any vertices which have been incorrectly merged and therefore the resulting stages should be checked to ensure they are sensible.

The coarsest CEG, C_∞, is that which has all situations in the same stage and the finest partition CEG, C_0, has each situation in a separate stage. The problem of searching over the CEGs for

the highest scoring therefore becomes a problem of searching over the clustering of stages for the highest scoring. This search can be computationally intensive and hence steps to simplify the search are required. One such step is to assume that the probability distributions of stages which are formed from the same vertices in the tree are equal in all CEGs and hence the differences between model scores, i.e. the logarithms of the relevant Bayes' factors, can be calculated rather than derive each score separately. This ensures the calculation of the logarithm of their posterior Bayes' factor (i.e. the ratio in step (4)) depends only upon the situations which have been merged, since all other stages remain unchanged.

The score therefore depends on two elements; the prior probability of the CEG being the true model and the marginal likelihood of the data. Hence these are the two elements required for the algorithm to run. Uniform priors can be allocated for the prior probability of the CEGs, or priors can be set using knowledge from previous studies or experts in the field, who can advise which paths or clusters are more likely. In some instances the prior Bayes factor for a CEG will be zero meaning that at least part of the CEG is impossible. For example, it may be impossible for a woman to be diagnosed with prostate cancer, or impossible for an individual to have undergone open heart surgery given that they have never had surgery. Information such as this helps to reduce the computation time when searching over the CEGs.

Setting the marginal likelihood for each CEG is equivalent to setting the priors over the CEG's parameters. With the assumption that the stage priors are independent for all CEGs and that equivalent stages in different CEGs have the same prior distributions on their probability vectors, this becomes a process of setting the parameter priors of the florets in C_0.

The usual prior for probability parameters of finite discrete Bayesian Networks is the product Dirichlet distribution. It has been shown[379] that Dirichlet priors are also required for CEGs and hence will be used throughout this chapter. In addition it has been shown how the priors can be allocated to the paths in the event tree rather than the florets and stated that when no prior information is available, it is assumed that each path through the CEG is equally likely.

The code used for this algorithm can be found in Appendix D.2 and an example of the output can be found in Appendix D.5 which corresponds to §D.4. The code was adapted here to include non-uniform priors where prior knowledge is available, as shown in Appendix D.3.

5.2.2.2 Equivalent Sample Size

The equivalent sample size is a measure of the strength of the prior beliefs. A small equivalent sample size corresponds to less confidence in the prior beliefs than a larger equivalent sample size. The equivalent sample size is often selected using one of two rules: (i) select the equivalent sample size such that each path in the event tree has an integer value, hence the equivalent sample size is at least equal to the number of paths in the tree, or (ii) select the equivalent sample size such that it is equal to the largest number of possible values a variable in the dataset can take, to ensure that the fractions are simple when a weak uniform prior is appropriate.[8] Both of these two options are designed to retain simplicity in the calculations (by including only integers or simple fractions) and are only guides.

In the real data examples which follow in this chapter, uniform priors are selected unless otherwise stated and in these instances rule (ii) is adhered to, where the equivalent sample size is equal to the largest number of possible values a variable in the dataset can take. This ensures the prior beliefs are weak and the data can hence play a more dominant role in the construction of the CEG from the event tree. Where non-uniform priors are selected, rule (i) is generally applied but the priors for each example are stated, along with the equivalent sample size chosen. It is recommended that the sensitivity of the CEG results to the equivalent sample size is investigated, since the findings can be sensitive to both the priors used and the strength of the belief in such priors. Where the diabetes data are used in conjunction with CEGs in §5.4 the sensitivity of the findings is reported.

As the equivalent sample size increases, the CEG can often become more complicated with fewer vertices in each position. Therefore a CEG formed with weaker prior beliefs and hence a smaller equivalent sample size, can often result in a simpler graph. This is demonstrated with the diabetes data in §5.4. Conclusions which are not sensitive to the equivalent sample size may be more reliable than those which require a given strength of prior beliefs.

5.2.3 The Chain Event Graph

The chain event graph, C, is a finite staged tree collapsed over its positions. The positions form the new vertices of the CEG and all leaves are collected into one final vertex represented by the position w_∞. Additionally, any two positions in the same stage are connected using a dashed line

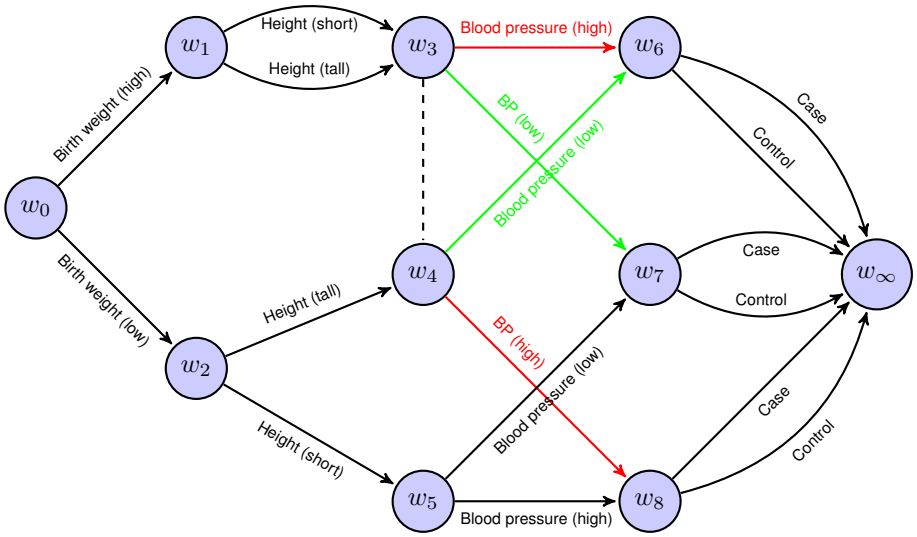

Figure 5.3: Chain event graph corresponding to the staged tree in Figure 5.2. BP = blood pressure.

and their edges shown with corresponding colours as in the staged tree, representing probability distributions which are indistinguishable. Edges are drawn between positions to represent each child from the position. A simple CEG is one where the positions and stages are equivalent and hence the CEG is uncoloured.[8]

The corresponding CEG for Figure 5.2 is shown in Figure 5.3. The CEG aligns vertices describing the same variable vertically, so position w_0 represents birth weight, positions w_1 and w_2 represent height, positions w_3, w_4 and w_5 represent blood pressure, positions w_6, w_7 and w_8 represent kidney disease and w_∞ represents all the leaf vertices. Edges are labelled with the possible categories and in Figure 5.3 both categories for a variable are available from each vertex associated with that variable. For example, for the vertices relating to height (w_1, w_2), both options of tall and short are available ($w_1 \to w_3$, $w_2 \to w_4$, $w_2 \to w_5$). In some instances, for example $w_1 \to w_3$, both categories within a variable follow the same edge and lead to the same vertex, w_3. These similarities and patterns within the CEG allow conclusions about the variables to be drawn.

5.2.4 Chain Event Graph Conclusions

The association of variable combinations with the outcome can be read directly from the CEG and the topology of the graph can be used to draw conclusions more specific than those relating to the conditional independence statements. It may be that a variable is associated with the outcome

for one subset of the population, but not another. For example, it may be that women have an increased association with a disease with age, but men have the same association with the disease, regardless of their age. This level of detail can be read directly from the CEG but not usually from other forms of graphical representation.[8] The CEG can also show which combinations of variables lead to the same positions. For example, a low birth weight and being male, may lead to the same position and consequently the same subsequent options for forthcoming variables as a normal birth weight and being female (this is particularly applicable in CEGs derived from asymmetric trees). Hence the CEG retains all the information provided by the corresponding tree, despite the collapse into positions, and the ability to retain multiple edges between vertices means no paths are lost.

Figure 5.3 shows that for low birth weight w_2, height determines the next vertex in the path, w_4 or w_5. However, when birth weight is high w_1, height is less important since both edges lead to vertex w_3. Regardless of height, if birth weight is high w_3, blood pressure will determine the vertex after w_3 (w_6 or w_7). Low birth weight can lead to any vertex relating to kidney disease (w_6, w_7, w_8), with low birth weight and tall height w_4 leading to vertices w_6 and w_8, and low birth weight and short height w_5 leading to w_7 and w_8. All vertices relating to kidney disease status (w_6–w_8) allow the option for an individual to be a case or a control, and all vertices lead to w_∞, representing the leaf vertices. Vertices of situations in the same stage are joined with a dotted line and their edges coloured using the colours from the staged tree. In Figure 5.3 this applies to positions w_3 and w_4 which show their high blood pressure edges in red and low blood pressure edges in green, showing the distribution of blood pressure is indistinguishable for w_3 and w_4. Of course, these findings are conditional on the thresholds chosen for the categories of continuous variables.

From Figure 5.3, birth weight appears to have an association with kidney disease since high and low lead to different positions, $w_0 \to w_1$ and $w_0 \to w_2$, hence groups of individuals with high and low birth weight contain different proportions of cases. Height is not associated with the outcome when the birth weight is high, $w_1 \to w_3$ but height is associated with the outcome when birth weight is low, $w_2 \to w_4$ and $w_2 \to w_5$. Low birth weight, tall height and high blood pressure ($w_0 \to w_2 \to w_4 \to w_8$) lead to the same vertex as low birth weight, short height and high blood pressure ($w_0 \to w_2 \to w_5 \to w_8$), suggesting height is irrelevant when birth weight is low and blood pressure is high, in terms of association with the outcome. In addition, high birth weight and high blood pressure ($w_0 \to w_1 \to w_3 \to w_6$) lead to the same vertex (w_6) as low birth weight, tall

height and low blood pressure ($w_0 \to w_2 \to w_4 \to w_6$), and low birth weight, short height and low blood pressure ($w_0 \to w_2 \to w_5 \to w_7$) lead to the same vertex (w_7) as high birth weight and low blood pressure ($w_0 \to w_1 \to w_3 \to w_7$). Findings such as these can be referred to clinical experts to discuss the biological mechanisms acting if known, or to form hypotheses to test.

Dynamic CEGs[381] can be used when there are an infinite number of edges or vertices, which may be useful when dealing with longitudinal studies which develop over a (possibly undefined) period of time. Since case-control studies are retrospective and all outcomes and available options are know at data collection, these extensions are not required. An advantage of case-control studies is that the outcome is binary and this allows ordinal chain event graphs to be used as will be discussed in §5.2.7.

5.2.5 Chain Event Graphs Where Variables Have Additional Categories

Additional categories can be added and the same method applied. For example, let birth weight now have three categories as shown by the staged tree in Figure 5.4, comprising of stages

$$u_0 = \{s_0\}, u_1 = \{s_1\}, u_2 = \{s_2, s_3\}, u_3 = \{s_4, s_5, s_6, s_8\}, u_4 = \{s_7, s_9\}, u_5 = \{s_{10}, s_{12}, s_{15}, s_{19}\},$$
$$u_6 = \{s_{11}, s_{13}, s_{17}, s_{21}\}, u_7 = \{s_{14}, s_{16}, s_{18}, s_{20}\},$$

and positions

$$w_0 = \{s_0\}, w_1 = \{s_1\}, w_2 = \{s_2, s_3\}, w_3 = \{s_4, s_5\}, w_4 = \{s_6, s_8\}, w_5 = \{s_7, s_9\},$$
$$w_6 = \{s_{10}, s_{12}, s_{15}, s_{19}\}, w_7 = \{s_{11}, s_{13}, s_{17}, s_{21}\}, w_8 = \{s_{14}, s_{16}, s_{18}, s_{20}\}.$$

Figure 5.5 shows the corresponding CEG. Low and average birth weight are similar to one another, but differ from high birth weight. When birth weight is high, height is irrelevant in terms of the outcome ($w_0 \to w_1 \to w_3$), but when birth weight is low or average, height dictates the next vertex ($w_0 \to w_2 \to w_4$ or $w_0 \to w_2 \to w_5$). Vertex w_6 can be reached using (at least) two different paths; high birth weight and high blood pressure ($w_0 \to w_1 \to w_3 \to w_6$), or low/average birth weight, tall height and low blood pressure ($w_0 \to w_2 \to w_4 \to w_6$). Vertex w_7 can also be reached by (at least) two paths; high birth weight and low blood pressure ($w_0 \to w_1 \to w_3 \to w_7$), or low/average birth weight, short height and low blood pressure ($w_0 \to w_2 \to w_5 \to w_7$), suggesting that when blood pressure is low, high birth weight has the same association with kidney

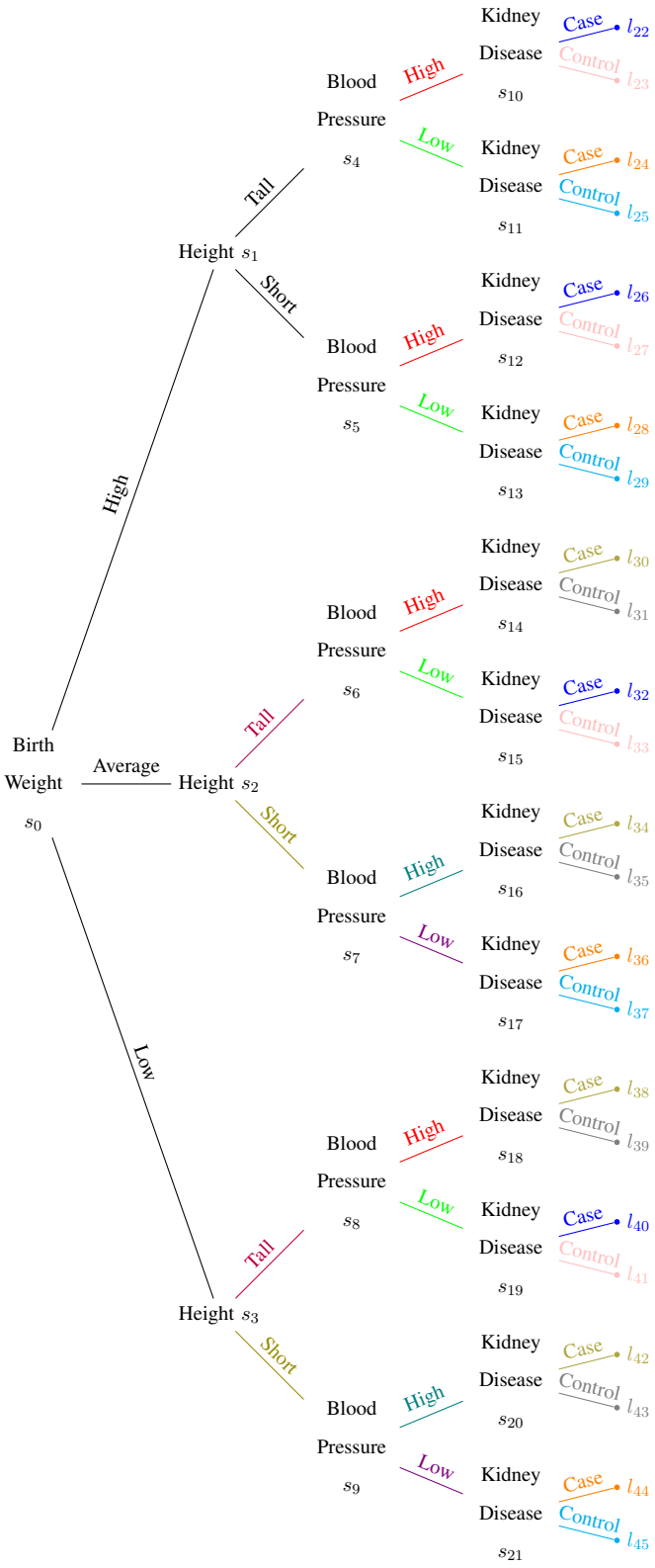

Figure 5.4: A staged tree including a variable which has more than two categories.

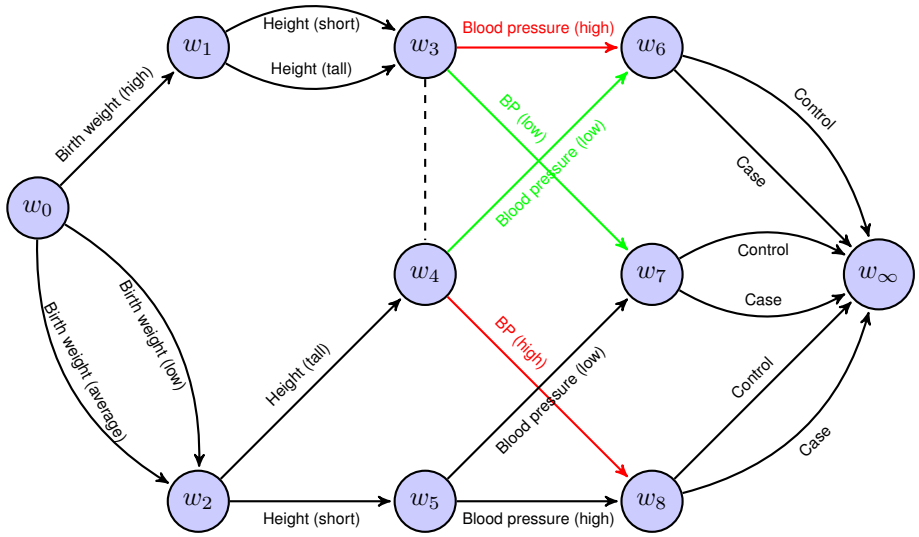

Figure 5.5: Chain event graph for variables with extra categories, with corresponding staged tree in Figure 5.4. BP = blood pressure.

disease as low/average birth weight and short height. Vertex w_8 can be reached by (at least) two different paths; low/average birth weight, tall height and high blood pressure ($w_0 \to w_2 \to w_4 \to w_8$), or low/average birth weight, short height and high blood pressure ($w_0 \to w_2 \to w_5 \to w_8$), implying that when birth weight is low/average and blood pressure is high, height is irrelevant with respect to the outcome. Again, these equivalences are conditional on the parameterisation, which may affect the stages found and hence positions. Therefore, there should be clinical reasoning for the chosen categories.

5.2.6 Chain Event Graphs With Additional Variables

The same method is also valid when more variables are present. Assume diabetic status is known before height and blood pressure, and hence let the tree now read as birth weight, diabetes, height, blood pressure, kidney disease status. A staged tree for this is given in Figure 5.6, with stages,

$$u_0 = \{s_0\}, u_1 = \{s_1\}, u_2 = \{s_2\}, u_3 = \{s_3, s_4, s_5\}, u_4 = \{s_6\}, u_5 = \{s_7, s_9, s_{12}\},$$
$$u_6 = \{s_8, s_{10}, s_{14}\}, u_7 = \{s_{11}, s_{13}\}, u_8 = \{s_{15}, s_{19}, s_{23}\}, u_9 = \{s_{16}, s_{20}, s_{26}\},$$
$$u_{10} = \{s_{17}, s_{21}, s_{25}, s_{28}, s_{30}\}, u_{11} = \{s_{18}, s_{20}, s_{26}\},$$

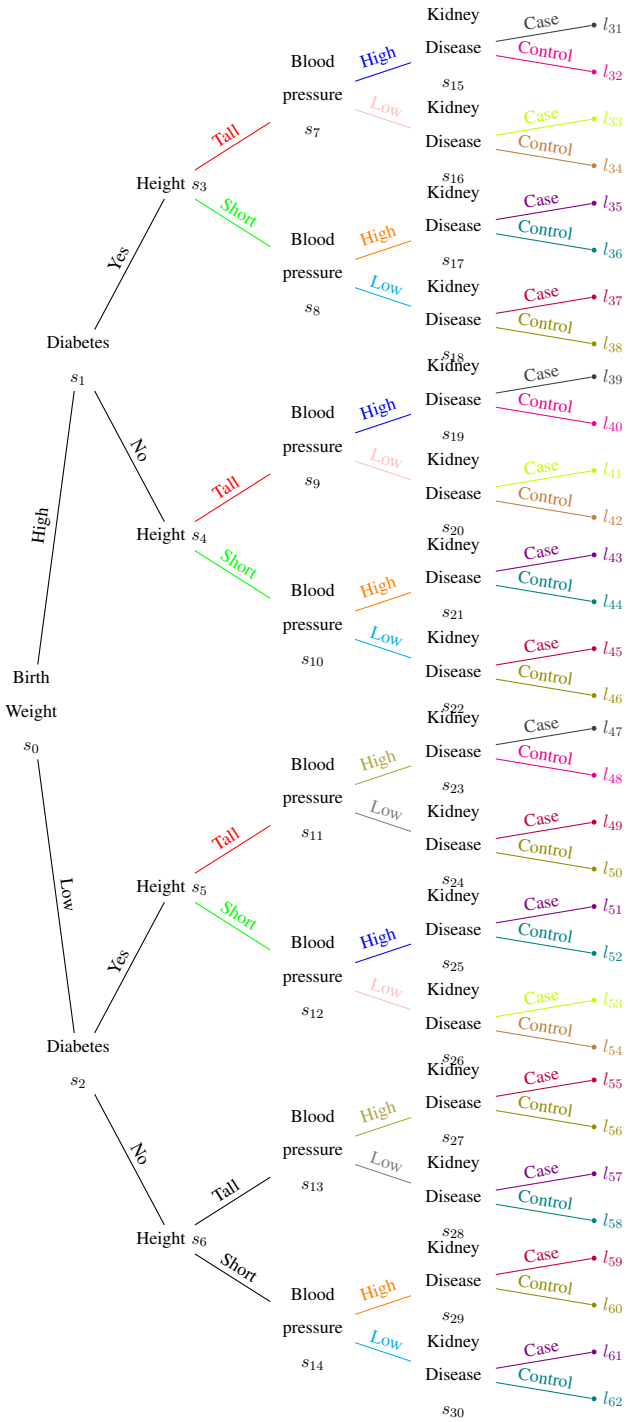

Figure 5.6: Staged tree for example with five variables.

and positions,

$$w_0 = \{s_0\}, w_1 = \{s_1\}, w_2 = \{s_2\}, w_3 = \{s_3, s_4\}, w_4 = \{s_5\}, w_5 = \{s_6\}, w_6 = \{s_7, s_9\},$$
$$w_7 = \{s_8, s_{10}\}, w_8 = \{s_{11}\}, w_9 = \{s_{12}\}, w_{10} = \{s_{13}\}, w_{11} = \{s_{14}\}, w_{12} = \{s_{15}, s_{19}, s_{23}\},$$
$$w_{13} = \{s_{16}, s_{20}, s_{26}\}, w_{14} = \{s_{17}, s_{21}, s_{25}, s_{28}, s_{30}\}, w_{15} = \{s_{18}, s_{22}, s_{24}, s_{27}, s_{29}\}.$$

The resulting, more complicated CEG can be found in Figure 5.7. When birth weight is high, w_1, diabetes status is irrelevant with respect to the outcome (w_3), but when birth weight is low, w_2, diabetes status determines whether the path continues to w_4 or w_5. Each of the height options lead to different vertices (w_6–w_{11}) depending upon the path taken thus far. Vertex w_{12} can be reached by the following paths; high birth weight, tall height and blood pressure high ($w_0 \to w_1 \to w_3 \to w_6 \to w_{12}$), or low birth weight, diabetes, tall height and blood pressure high ($w_0 \to w_2 \to w_4 \to w_8 \to w_{12}$), suggesting that high birth weight is equivalently associated with kidney disease to having low birth weight and diabetes, when the individuals are tall and has high blood pressure. Vertex w_{13} can be reached by the following paths; high birth weight, tall height and low blood pressure ($w_0 \to w_1 \to w_3 \to w_6 \to w_{13}$), or low birth weight, diabetes, short height and low blood pressure ($w_0 \to w_2 \to w_4 \to w_9 \to w_{13}$), suggesting that low blood pressure, high birth weight and tall height, have a similar association with kidney disease to low birth weight, diabetes and short height.

Vertex w_{14} can be reached by four paths, high birth weight, short height and high blood pressure ($w_0 \to w_1 \to w_3 \to w_7 \to w_{14}$), low birth weight, diabetes, short height and high blood pressure ($w_0 \to w_2 \to w_4 \to w_9 \to w_{14}$), low birth weight, no diabetes, tall height and low blood pressure ($w_0 \to w_2 \to w_5 \to w_{10} \to w_{14}$), or low birth weight, no diabetes, short height and low blood pressure ($w_0 \to w_2 \to w_5 \to w_{11} \to w_{14}$), implying that when birth weight and blood pressure are low and there is no diabetes, then height is irrelevant for the outcome of interest. Additionally, when height is short and blood pressure is high, the association with kidney disease is similar whether there is a high birth weight, or a low birth weight and diabetes.

Vertex w_{15} can be reached using one of four paths; high birth weight, short height and low blood pressure ($w_0 \to w_1 \to w_3 \to w_7 \to w_{15}$), low birth weight, diabetes, tall height and low blood pressure ($w_0 \to w_2 \to w_4 \to w_8 \to w_{15}$), low birth weight, no diabetes, tall height and high blood pressure ($w_0 \to w_2 \to w_5 \to w_{10} \to w_{15}$), or low birth weight, no diabetes, short height

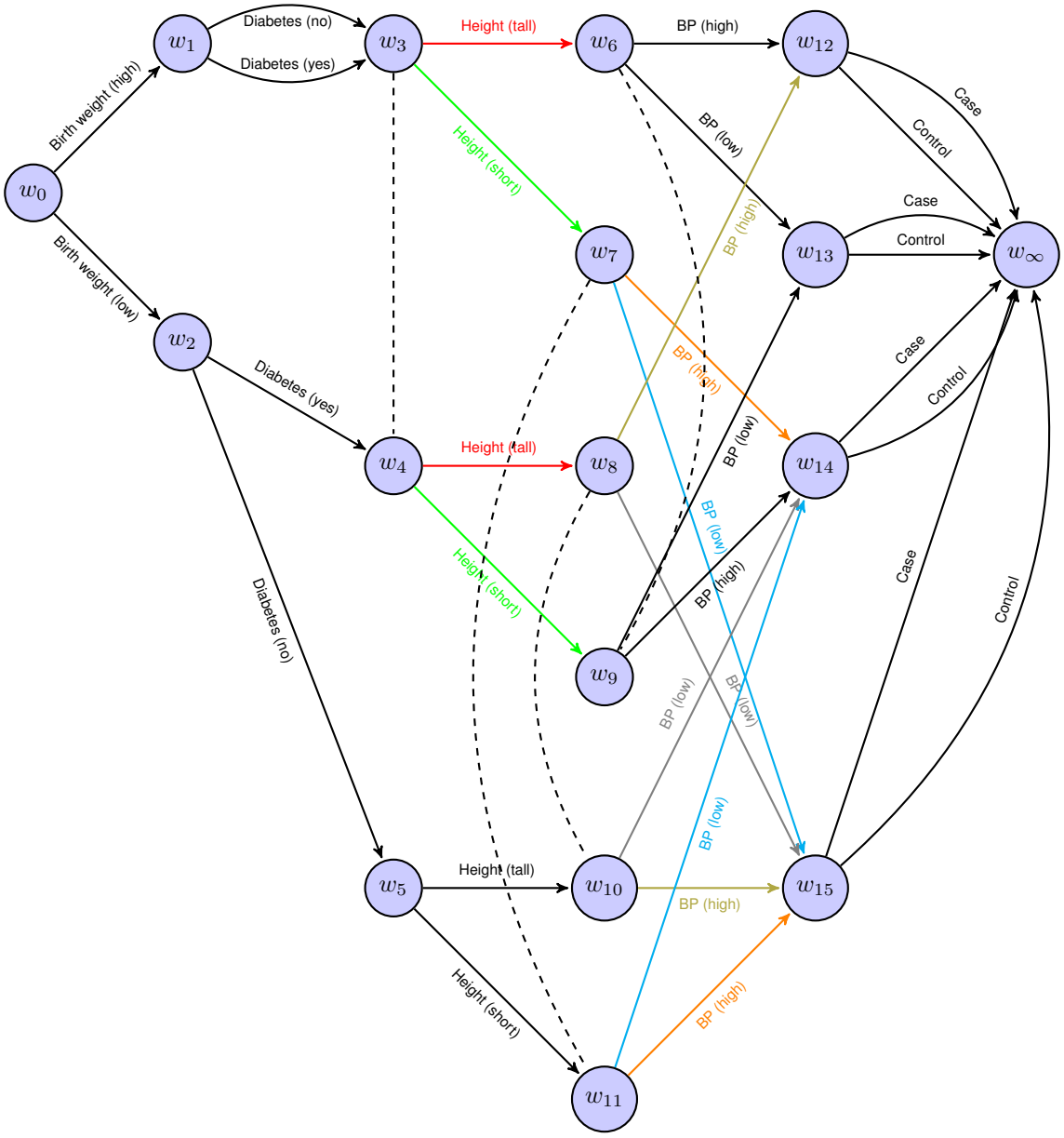

Figure 5.7: Chain event graph for a dataset with five variables, corresponding to the staged tree in Figure 5.6. BP = blood pressure.

and high blood pressure ($w_0 \to w_2 \to w_5 \to w_{11} \to w_{15}$), suggesting that when birth weight is low, there is no diabetes and blood pressure is high, then height is irrelevant for the outcome of interest. Also, when birth weight is low and height is tall, then having diabetes and low blood pressure has the same association with kidney disease as not having diabetes but having high blood pressure.

5.2.7 Ordinal Chain Event Graphs

The ordinal CEG[8] is an extension of the CEG which is specifically designed for binary outcomes and which allows more conclusions to be drawn by imposing extra conditions. Ordinal CEGs are CEGs which order the vertices within each variable, with respect to the outcome. To construct an ordinal CEG, first partition the situations in the tree into subsets such that each subset contains all vertices whose associated florets show the same variable. There will be one subset for each variable in the tree. Next, the vertices within a subset may be grouped into stages so a CEG can be formed. When constructing the ordinal CEG, the positions within a given vertex subset are aligned vertically. This results in each variable from the data representing one of the 'columns' in the ordinal CEG. Finally, these positions are reordered such that they are in descending order with respect to the probability of the outcome not occurring given the CEG, that is the probability that the outcome has a value of zero. So for case-control studies, the lower down the graph the combination of the variables are, the more likely the individuals are to be a case.

An example of an ordinal CEG is given in Figure 5.8, where the percentage of individuals with the outcome of interest have been included at each vertex. In this example, there are 50% cases as shown in w_0 and w_∞, then for variables 1 and 2, the vertices are ordered such that the highest percentage of cases within the variable are positioned towards the bottom of the graph (75% is positioned lower than 33%, and 91% is positioned lower than 38%).

When the paths in the ordinal CEG are all of the same length, which can occur frequently, the CEG can be used at different stages to draw conclusions on the variables encountered thus far, rather than using the entire ordinal CEG. This may be particularly useful for applications to cohort studies to consider different points in time, or for survival analysis to investigate time since diagnosis. Ordinal CEGs are simple to construct since they involve just the reordering of the vertices within each variable, and are suitable for case-control studies since they just require a binary outcome.

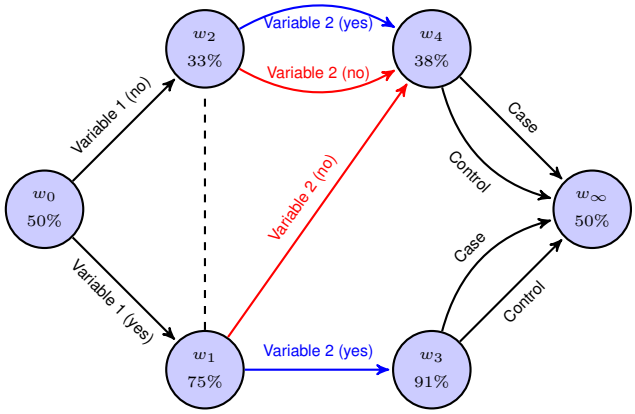

Figure 5.8: An example of an ordinal chain event graph, where the vertices within a variable are ordered with respect to the outcome.

5.3 Chain Event Graphs to Explore Missingness Mechanisms in Case-Control Studies

Exploring missingness is the main focus for CEGs in this thesis. This chapter has introduced CEGs and shown their current use. This section will describe how CEGs can be used to investigate missingness mechanisms and demonstrate this using the diabetes dataset in §5.4. Chapter 6 will then extend the ideas here, for use specifically with non-participation in case-control studies.

CEGs can be used to systematically represent and explore the missingness in a dataset, and draw conclusions[365] using the three standard missingness categories defined by Rubin; missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR).[289] However, further assumptions are required to confirm data which are thought to be MAR.[365] When the data are MNAR, ordinal CEGs can help to define the missingness mechanism further, including how the variables involved are combined. The missingness categories can be defined using conditional independence statements as in CEGs. For MCAR, the missingness is independent of the observed variables and the partially observed variables. MAR is when the missingness is conditionally independent of the variables with unobserved values given those with completely observed values. Finally, MNAR is when the missingness is not independent of the variables with observed values or those with unobserved values. Chapter 4 discussed the assumptions associated with the current methods used to reduce participation bias, some of which

required the data to be MAR. CEGs can be used to help determine which missingness mechanism is operating, and hence the appropriateness of the methods can be assessed. One way in which to display missing data is by using a missingness indicator, which then acts as an additional variable in the event tree and CEG, as will be demonstrated in this section.

5.3.1 An Illustrative Dataset With Missing Data in One Variable

An illustrative dataset is used to demonstrate data which are MCAR, MAR and MNAR. Assume that blood pressure is unknown for some individuals due to non-participation, while birth weight, height and kidney disease are known from health records, or that blood pressure is missing as the individuals were willing to participate in a survey, but not in a physical examination. The tree can be adapted as shown in Figure 5.9, where extra vertices are added before the blood pressure variable. If blood pressure is missing, the path is shortened and the next vertex relates to kidney disease. Figure 5.9 can be converted to an ordinal CEG and conclusions drawn as to whether an assumption of MAR is plausible, as required by some of the methods described in Chapter 4. Typical CEGs for when data are MCAR, MAR and MNAR follow.

5.3.2 Chain Event Graphs: Missing at Random

Under the MAR assumption, when the outcome is fully observed as in case-control studies, there is the additional assumption that the outcome is independent of the missingness process, given the observed variables.[365] Under this assumption, the probability of the outcome given the observed variables and no missingness, is equal to the probability of the outcome given the observed variables with missingness. The probability of the outcome is consequently a weighted average of the outcome given the observed variables along the given path, hence the probability of the outcome when data are missing should lie between the categories of the variable when it is observed. If using an ordinal CEG, the vertical alignment aids the visual interpretation of these probabilities and the likelihood that the data are MAR. If the graphical layout is not satisfied, the data are MNAR. In case-control studies the outcome may not always be independent of the missingness process, as participants are selected on the outcome and data are collected retrospectively. This assumption, which may also not hold for some cohort studies, underlies the ability of CEGs to differentiate MAR from MNAR.

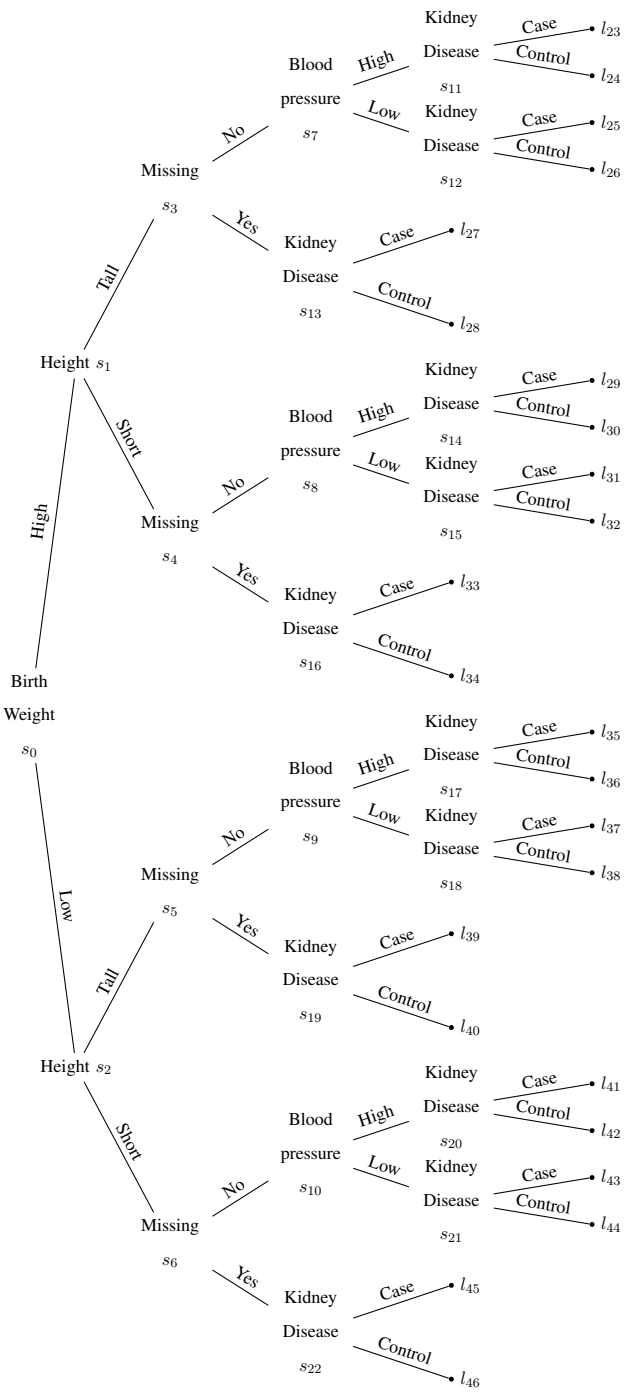

Figure 5.9: Tree showing missingness in the blood pressure variable.

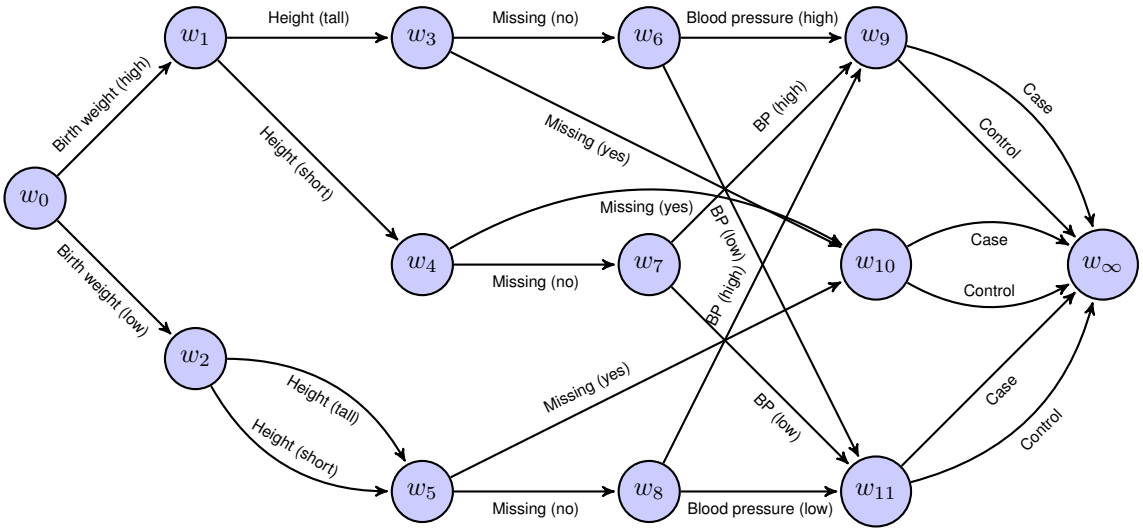

Figure 5.10: Chain event graph example for data which are missing at random (MAR). BP = blood pressure.

An example of an ordinal CEG corresponding to Figure 5.9 for when the data are MAR is given in Figure 5.10. Using vertices w_9, w_{10} and w_{11}, it can be seen that all high blood pressures go to w_9 ($w_6 \to w_9$, $w_7 \to w_9$ and $w_8 \to w_9$), all low blood pressures go to w_{11} ($w_6 \to w_{11}$, $w_7 \to w_{11}$, and $w_8 \to w_{11}$), and all missing blood pressures go to w_{10} ($w_3 \to w_{10}$, $w_4 \to w_{10}$, and $w_5 \to w_{10}$). Therefore, the missingness category leads to a vertex between those for high and low blood pressure, meaning the data could be MAR. Given this CEG is ordinal, it also shows that high birth weight is associated with a lower probability of being a case than low birth weight, and that high blood pressure is associated with a lower probability of being a case than low blood pressure. When birth weight is high there is also a greater association with the outcome when height is short compared with tall.

To ascertain whether the data are truly MAR, the probability of the outcome given the observed variables must be equivalent with or without missingness, and this must be checked. It is possible to obtain a CEG structure such as that in Figure 5.10 without the probabilities being equal. This can occur when the probabilities for high and low blood pressure are very unbalanced for example, leading to one probability near one and the other close to zero, hence the probability with missingness is likely to fall between the two. Instances such as this are more likely to occur when the missing variable is something considered to be rare and less likely when the missing variable is something such as height, where tall and short both occur frequently. The graph hence gives

the first step towards satisfying the MAR assumption and the calculation using the probabilities provides the second and final step. To clarify, the position of the edge depicting missingness relative to the other edges representing known categories is important, but the angle between these edges is unimportant and does not display the proportions of categories in the missing category. Therefore, CEGs can be displayed as clearly as possibly by amending the angles between edges, provided the ordering of the edges leaving a vertex is unchanged.

The identities of the missing values in Figure 5.10 cannot be stated with great confidence, but are likely to be a combination of the recorded categories; high and low blood pressure. To reduce the bias from non-participation, methods such as multiple imputation (§4.4.2) and inverse probability weighting (§4.4.1) could be used, since the data are MAR.

5.3.3 Chain Event Graphs: Missing Completely at Random

If data are MAR, they may also be MCAR. For this, the vertices associated with the missingness indicator must be in the same stage, hence the probability of a missing value will be independent of the preceding variables as shown in Figure 5.11. A method to reduce any bias resulting from non-participation is unlikely to be required since although the sample size is reduced, the association between variables is not affected and any estimates or reported associations should remain valid.

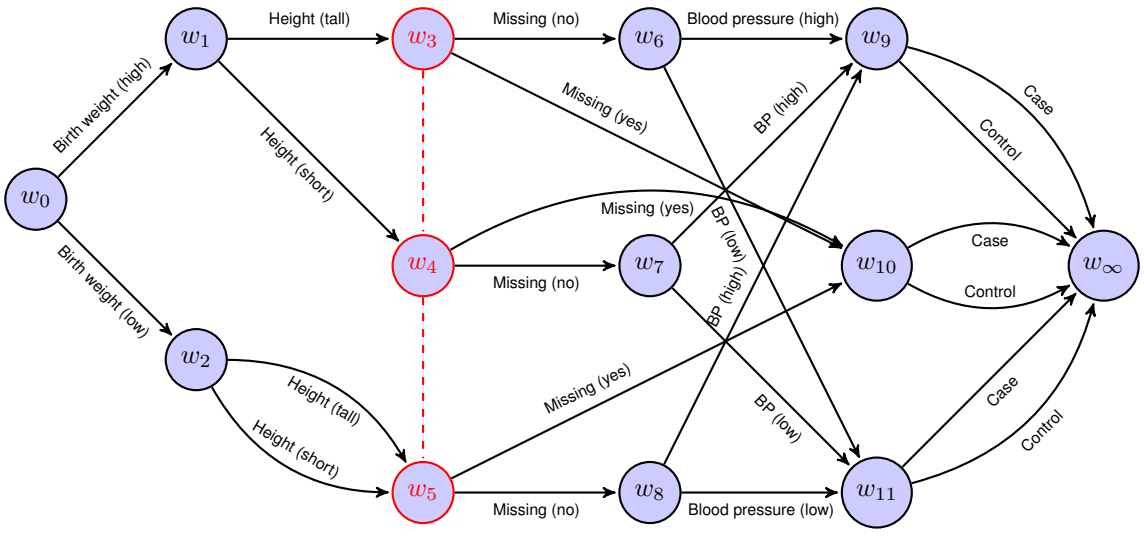

Figure 5.11: An extension of the CEG in Figure 5.10 showing data which are MCAR. BP = blood pressure.

5.3.4 Chain Event Graphs: Missing Not at Random

When data are MNAR, the missingness indicator is dependent on both the observed and unobserved values. This may be evidenced in a CEG by all missing categories having a lower/higher probability of being a case than the observed categories.[365] In instances such as this, the ordinal CEG alone is sufficient to declare the data are MNAR.

Data which are MNAR may provide an outcome worse than the poorest observed category as shown in Figure 5.12, where all high blood pressures lead to w_9 (the position least associated with being a case), all low blood pressures lead to w_{10} (the position second least associated with being a case) and the missing data lead to vertex w_{11} (the position most likely to result in a case). The converse may also be true, that the missing data provide a superior outcome, as shown in Figure 5.13. All high blood pressures now lead to the centre vertex w_{10}, while all low blood pressures lead to the bottom vertex w_{11}. Methods to reduce bias resulting from non-participation will need to be applicable to data which are MNAR, such as sensitivity analyses (§4.1), rather than multiple imputation or inverse probability weighting which usually require the data to be MAR.

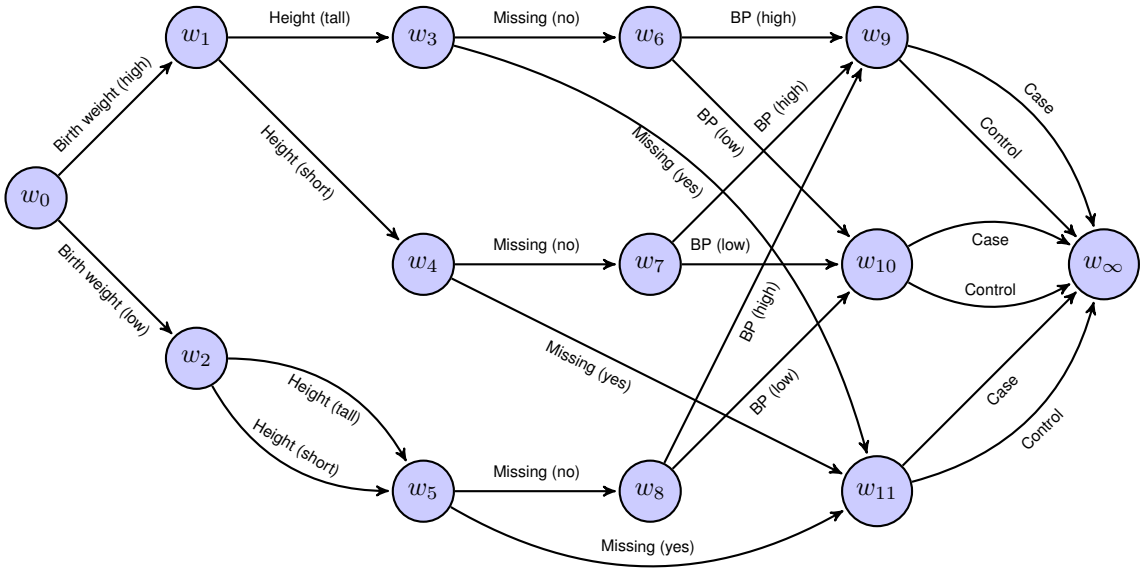

Figure 5.12: An example of when data are MNAR, with a poorer outcome than observed data. BP = blood pressure.

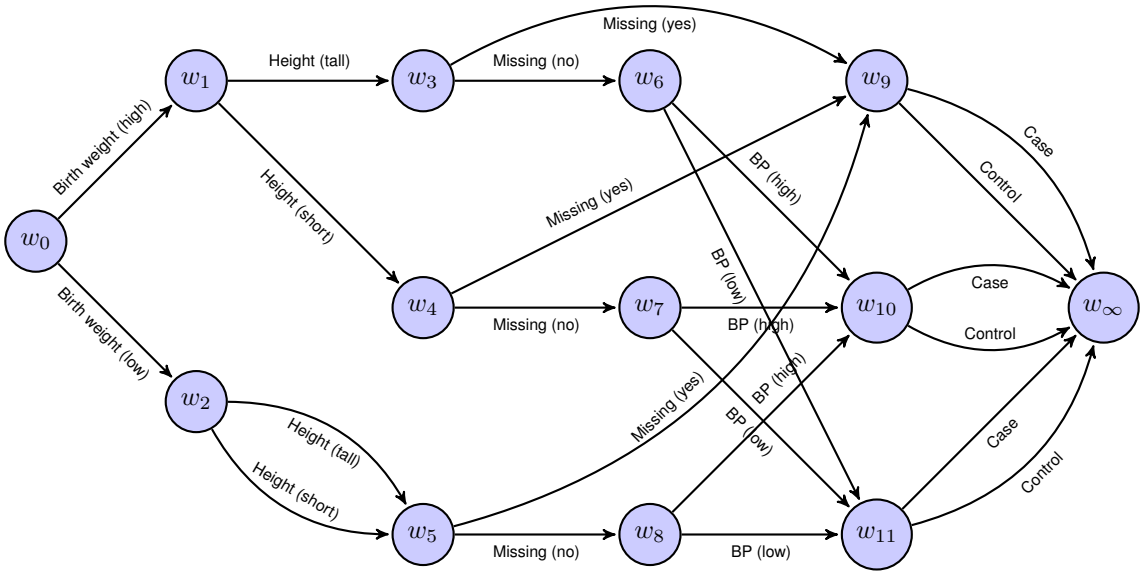

Figure 5.13: An example of when data are MNAR, with a superior outcome than observed data. BP = blood pressure.

5.3.4.1 Extra Level of Detail When Missing Not At Random

Missingness may depend upon a category from a previous variable, as shown in Figure 5.14, where missingness is MNAR when birth weight is low, but MAR (conditional on further checks on probability) when birth weight is high, regardless of height. When birth weight is high, the missingness in blood pressure leads to a position in the ordinal CEG which is between those of high and low blood pressure, yet when birth weight is low, the missingness in blood pressure leads to a position below that of both high and low blood pressure. Therefore, the missingness mechanism can be dependent upon categories within a variable and not necessarily generalised to the entire variable. This finer division of the MNAR mechanism has not before been applied to case-control studies, but in §5.4 it will be applied to the diabetes dataset.

One option here for methods to reduce the bias resulting from non-participation is to select a method which allows for the least random form of missingness, where MCAR is the 'most' random followed by MAR, and with MNAR as the 'least' random. This ensures all categories of the variable are accounted for and the method assumptions are valid.

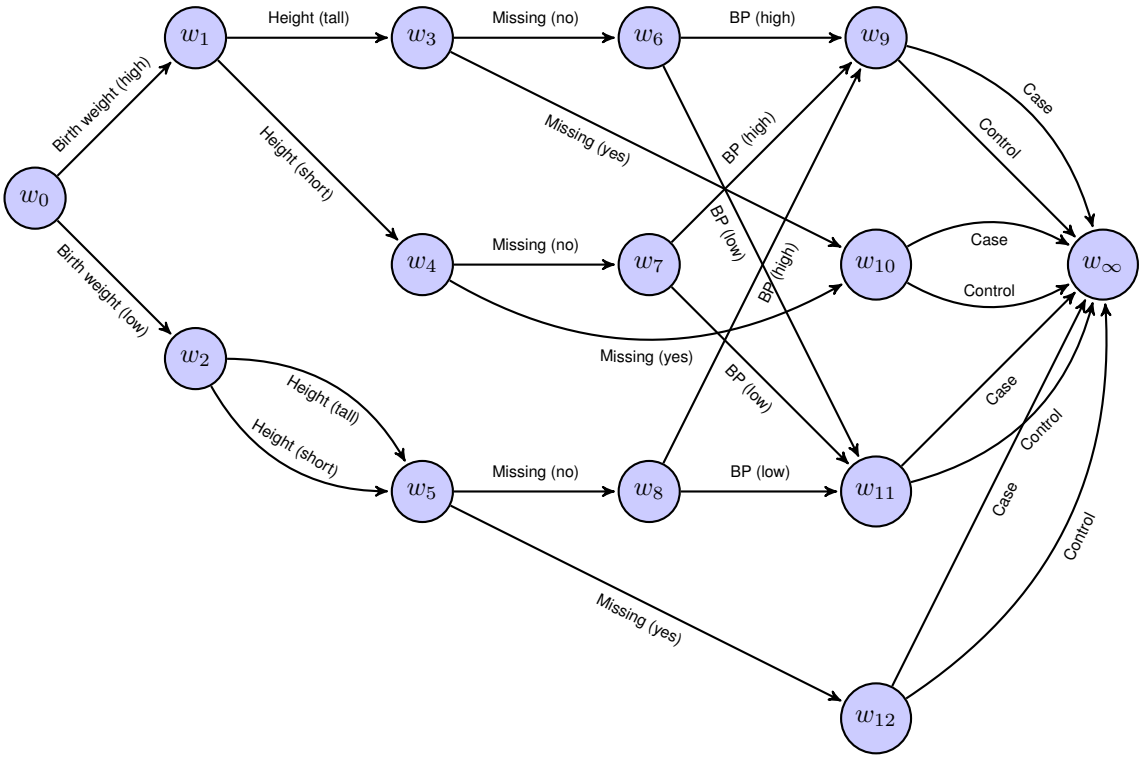

Figure 5.14: An example of when data are MNAR for some categories of a previous variable, yet MAR for others. BP = blood pressure.

5.3.4.2 Suggestion of the Missing Data Categories

The influence of the missing values can also be concluded from an ordinal CEG. For example, if the missing category always sits with a higher/lower probability of being a case than the observed categories, then the missingness appears to be influential.[365] However, if the missing category always joins to the same vertex as one of the observed categories and hence returns the same probability of being a case, then the missing values are less influential. When the missing category always sits above/below the observed categories on the ordinal CEG, it can be assumed (but not known) that the missing values are mainly those from the nearest observed category. Sensitivity analyses could be used to test the effect of this assumption.

The association of the outcome with the missing category may be similar to one of the observed categories as shown in Figure 5.15, where the missing category and low blood pressure lead to the same vertex and hence it can be assumed that many of the missing values are likely to be low. The influence of the missingness can also be judged, for example in Figure 5.15 where missingness

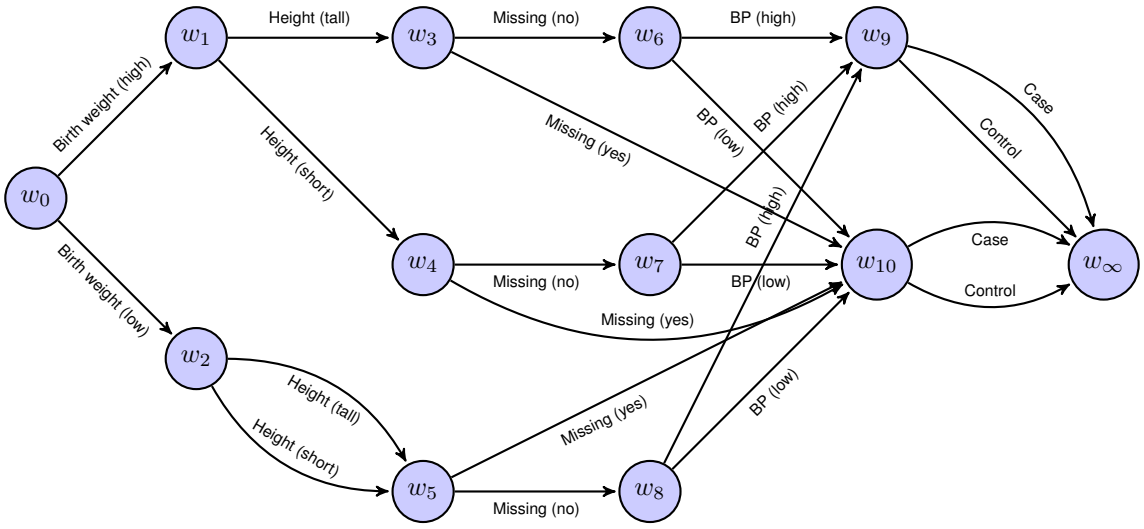

Figure 5.15: An example of when data are MNAR but similar to an observed category (low blood pressure). BP = blood pressure.

in blood pressure has a similar association with the outcome to low blood pressure, it may be considered to be less influential then in Figure 5.13 when the outcome was superior to that of the recorded blood pressure values. These data would also require a method which allows for data to be MNAR, since these data are likely to be mainly from one of the observed categories.

5.3.5 Missingness in Multiple Variables and Reduced Ordinal CEGs

Missingness in case-control studies can occur in more than one variable, whether it be due to non-participation or partial-participation. An example of missingness in two variables (height and blood pressure) is shown in Figure 5.16, and in this instance one approach used is to extend the missingness indicator to describe the number of missing variables.[8] For example for two variables, both are missing, neither are missing, the first only, or the second only. Another approach is to add 'missing' as an edge leaving any variable known to contain missing values; the CEG using this approach is given in Figure 5.17. This limits the ability to distinguish between data which are MCAR and MAR, since it is no longer clear whether the missingness indicator is independent of all the (non-outcome) variables, or just independent given the observed variables. However, it still allows valuable decisions about the relationships in the data to be drawn[8] and is sufficient for determining whether data are MAR as required by some methods in Chapter 4.

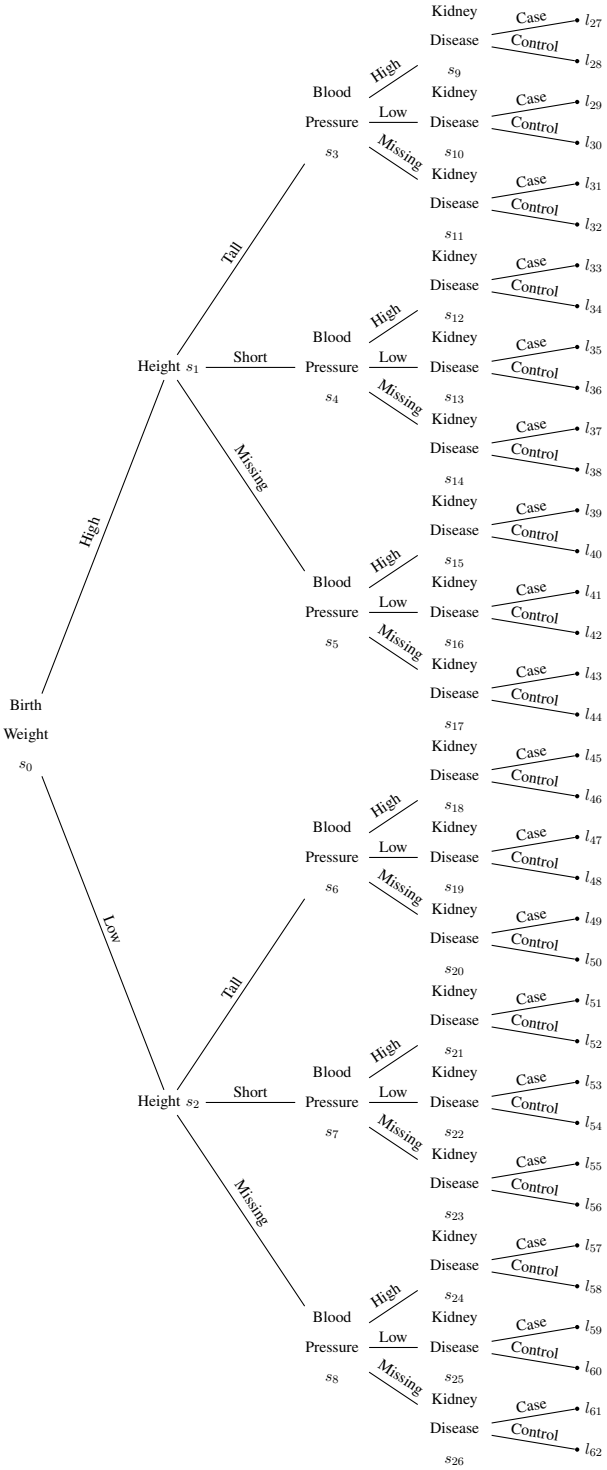

Figure 5.16: An example of a tree with four variables, two of which have missing values.

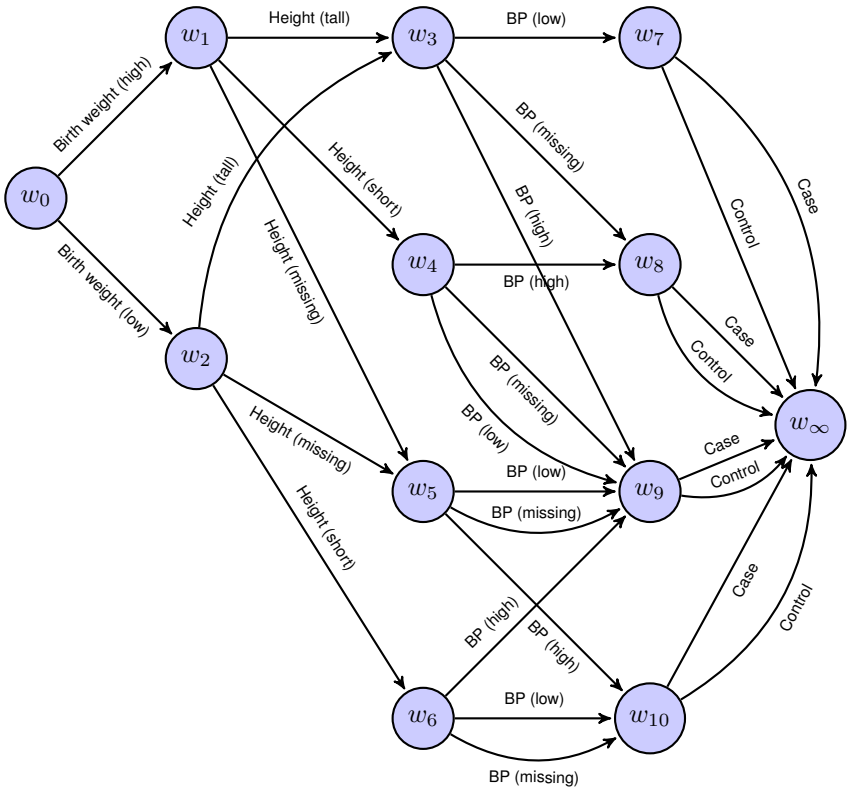

Figure 5.17: An example of a CEG where more than one variable has missing data. BP = blood pressure.

This section has already shown how CEGs can be used to investigate the missingness mechanisms where data are missing from a single variable. Where data are missing from multiple variables, the methods to reduce bias resulting from non-participation will depend upon the overall aim for the data. If only some variables are required, methods which suit the missingness mechanisms of those variables can be used. If all variables are required and there is a mix of missingness mechanisms, one approach could be to select a method to reduce bias resulting from non-participation which allows for the 'least' random form of missingness. For example, for combinations of MCAR and MAR a method for MAR data could be used, since MCAR would still be plausible. An option which should suit most scenarios is a sensitivity analysis.

If necessary, a *reduced ordinal CEG* can be formed, which aims to improve the clarity of the graph.[8] The reduced ordinal CEG works by retaining the root vertex and leaf vertex, plus the vertices leading to the leaf vertex. All other vertices are represented using 'intermediate' positions, which are groupings of the other variables, for example "the number of risk factors"; exactly one

high risk, one high risk plus one missing, one low risk plus one missing, etc. The next position may describe which of the variables has the missing data. As few edges are used as possible to describe the data, hence paths from the root vertex to the leaf vertex may be of different lengths. Re-plotting the ordinal CEG using these intermediate positions can result in a much simpler graph;[8] examples are available currently only in an online thesis.[8]

The intermediate vertices are labelled with an I superscript and the positions in the same stages are no longer joined using a dashed line.[8] There are only single edges between these vertices and hence one edge may describe multiple levels of a variable, unlike the standard ordinal CEG. Each edge is still labelled with the categories it represents. The ordinal feature of the CEG allows conclusions to be drawn regarding the number of risk factors and whether or not their values are missing. These risk factors can also be linked to other variables in the CEG which are fully observed. It is possible that comparisons between missing variables could be made, for example, missing variable X is generally associated with a poorer outcome than missing variable Y. However, full information regarding the variable is not available from a reduced ordinal CEG and so it must be used in conjunction with a standard ordinal CEG. For this reason, reduced ordinal CEGs will only be introduced here for comparison to other approaches later (§6.3.2), and not used during analysis. Further details are available in the thesis through which they were developed.[8]

5.4 Diabetes Dataset: Five Variables, Including Missing Data

The CEG framework has been explained, examples have been given, and the algorithm required for application to real data has been provided. CEGs will now be used with the diabetes data, to demonstrate that the application of CEGs with case-control data is possible (§D.4) since it has not before been achieved (see §5.1.2), primarily to explore the missingness produced through non-participation.

For completeness, the three variables used thus far (amniocentesis, caesarean delivery and diabetes) are analysed using CEGs in Appendix D.4, but to investigate missingness, two additional variables are added to the diabetes data; the fully-observed school-leaving-age of the mother, and the partially-observed rhesus factor category of the mother. The missingness in the rhesus factor variable is due to the category not being recorded in the medical notes, but the same analysis would

follow if the missingness had been due to non-participation in a given test. Unfortunately, data regarding non-participation was not collected for the diabetes data, hence the partially-observed rhesus factor variable is used as an example of a variable with missing data.

In this section, the analysis will be extended to incorporate prior knowledge for the paths through the event tree and will investigate the effect of swapping variables where their chronological ordering is unclear. Analyses will also be conducted which test the sensitivity of the results to the prior information, by changing the priors and varying the strength of the prior beliefs. Suitable methods to address the missingness in the rhesus factor variable will be discussed and details for how the missingness is structured will be provided.

5.4.1 Chain Event Graph Formation

The chronological ordering of the four categorical exposure variables and outcome is (i) rhesus factor of the mother, determined by the presence or absence of a protein in the blood (positive/negative/unknown), (ii) school-leaving-age of the mother, assuming the pregnancy begins after the mother has left school (16 or under/over 16) (iii) amniocentesis, usually during weeks 15–20 of the pregnancy[382] (yes - at least one with the study child/no - none), (iv) caesarean delivery at the end of the pregnancy (yes/no for the study child), and (v) diabetes status of the child, with type I diabetes diagnosis during childhood (case/control). Values were recorded for all participants for all variables, except for some missing rhesus factor values. Each variable here is categorical and the chronological ordering is apparent, but in other instances expert opinion may be required to determine a plausible ordering and sensible cut-off values, preferably with clinical meaning. Solutions are given in §5.4.3 for when the chronological ordering of the variables is not clear.

A strength of CEGs being a Bayesian approach is that prior information from previous studies can be incorporated,[6,383–387] as given in Table 5.1. Another approach would be to seek expert opinion for the probability of each edge in the tree. Table 5.1 shows that around 86% of the UK are rhesus positive,[383,384] but it cannot specify the expected percentage of unknown rhesus factor categories in a study. If the proportion of unknown rhesus factor from the data is used (3–4%) in conjunction with the data from Table 5.1, a split of negative:positive:unknown as 2:17:1 can be used as an approximation for the ratio of each category.

Variable	Categories	Ratios	Source and Assumptions
Rhesus factor	Negative:Positive :Unknown	–	Around 86% of the UK are rhesus positive.[383,384]
School-leaving-age	16 and under:Over 16	7:3	A parliamentary paper[385] assuming the majority of mothers left school around 1970–1985.
Amniocentesis	Yes:No	1:49	Around 15,000 amniocentesis in Britain each year[386] (about 2% of pregnancies).[6]
Caesarean section	Yes:No	1:9	Around the time the children were born, around 10% of births were by caesarean.[387]
Diabetes	Case:Control	1:2	Participants are in matched pairs ($\times$67) or triplets ($\times$129). Let us assume that controls are twice as common as cases in these data.

Table 5.1: Ratios of the variable categories in the diabetes data, provided for the time at which the study was conducted. The true case-control ratio could be used, but is simplified to 1:2 to reduce the equivalent sample size (see §5.2.2.2). This is also true for the rounded rhesus factor ratio.

The ratios in Table 5.1 can be used to assign values along each path to show their probability of being taken. This estimated probability from prior knowledge is used in conjunction with the sample data in the analysis. A common approach when using uniform priors for the probability along each path is to assign each leaf a value of one, and work backwards through the tree, so the root vertex starts with a value equal to the number of paths in the tree. This approach ensures each edge is assigned an integer. If integers are chosen while including the ratios, the equivalent sample size required is $20\ (2+17+1$ for rhesus factor$) \times 10\ ($school age$) \times 50\ ($amniocentesis$) \times 10\ ($caesarean$) \times 3\ ($diabetes$) = 300,000$. The larger this starting value, the more confidence there is in the values assigned from Table 5.1. This value of 300,000 is then divided at each vertex according to the ratios in Table 5.1 as shown in Figure 5.18.

The prior knowledge is incorporated into the analysis using the AHC algorithm (§5.2.2.1), with the amended R code for the algorithm as shown in Appendix D.3. Further details for the priors are available.[379] Incorporating prior information can have the effect of changing the stages which are reported by the algorithm, and so the sensitivity of the results with respect to both the prior

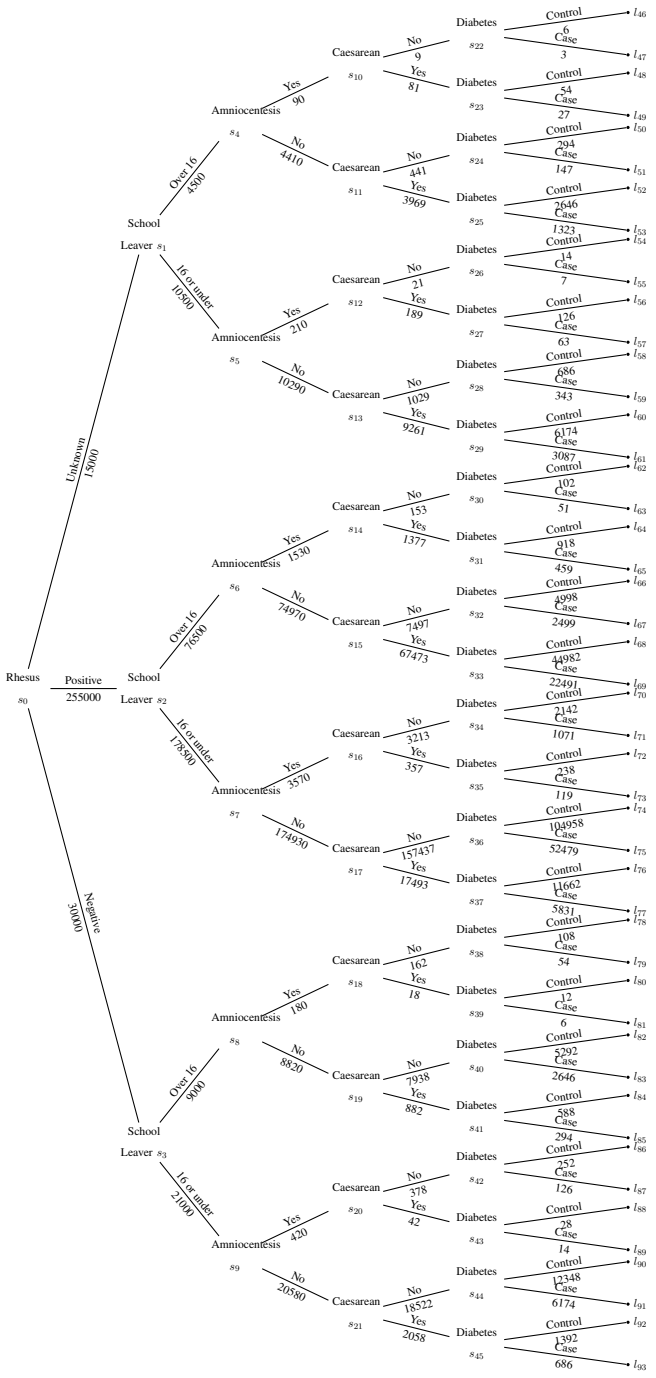

Figure 5.18: Five variable diabetes event tree, showing the ratios along each edge.

knowledge and the strength of the prior beliefs, should be tested (see §5.2.2.1 and §5.2.2.2). The algorithm output for the diabetes data was similar in format to that shown in Appendix D.5, but longer since the tree included additional variables and edges. The algorithm returned 15 stages after 32 iterations; the scores of which are shown in Figure 5.19, which shows the score used in the algorithm being maximised and stabilising. Each iteration results in a score with a greater value, until the score can no longer be maximised. The resulting (staged) tree is given in Figure 5.20, with the number of individuals taking each edge shown.

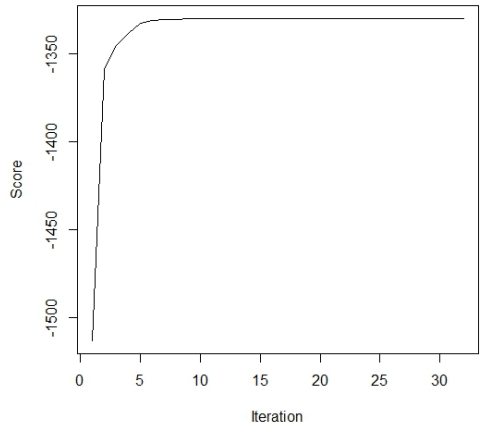

Figure 5.19: Plot of the scores generated during the AHC algorithm.

Trees and CEGs can be pruned by removing unused edges to reveal a simpler graph which is easier to read.[8] Figure 5.21 shows the pruned ordinal CEG resulting from collapsing Figure 5.20 over its positions, with the percentage of cases given at each vertex.

5.4.2 Interpretation

Figure 5.21 shows 38% of the individuals in the dataset are cases. There is little difference between the rhesus factor categories in the CEG, since around 40% of the individuals at each vertex are cases. In addition, the categories for school-leaving-age do not display any clear pattern, with the over 16 years category leading to both the highest (w_9 = 50%) and lowest (w_4 = 20%) proportion of cases. These findings suggest that the rhesus factor and school-leaving-age of the mother are not associated with the disease status of the child.

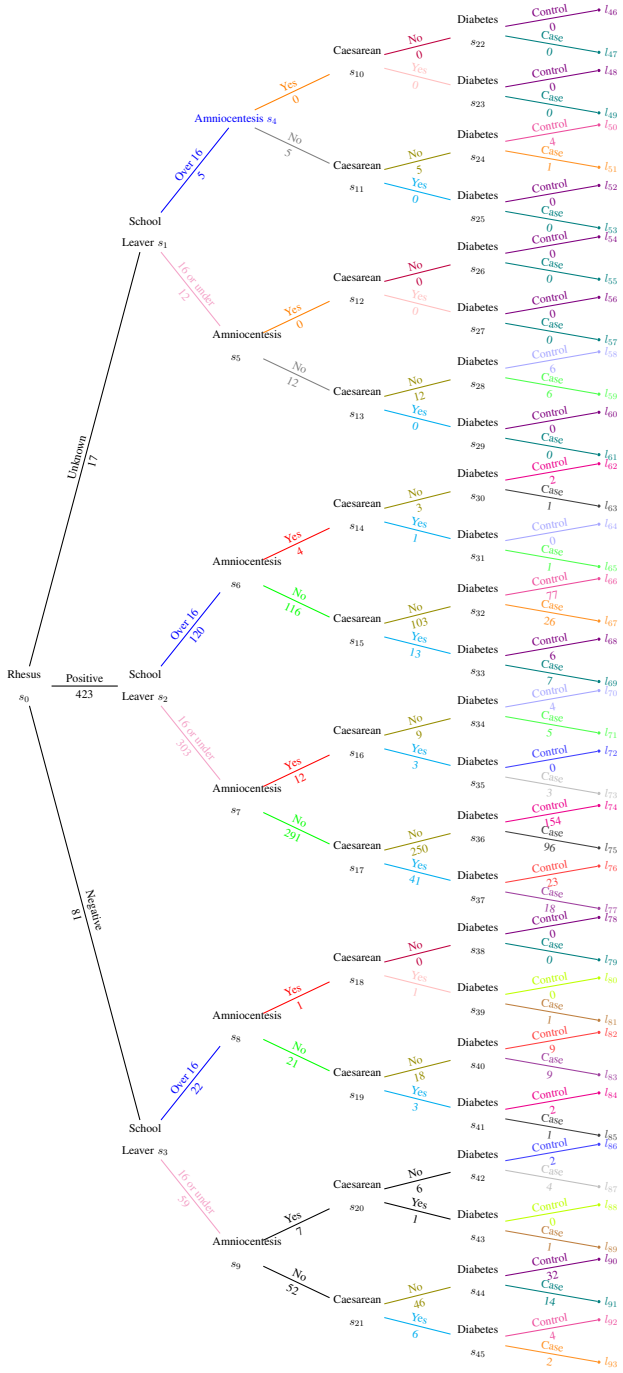

Figure 5.20: Staged tree for the four exposure and outcome variables; unequal probabilities along each path.

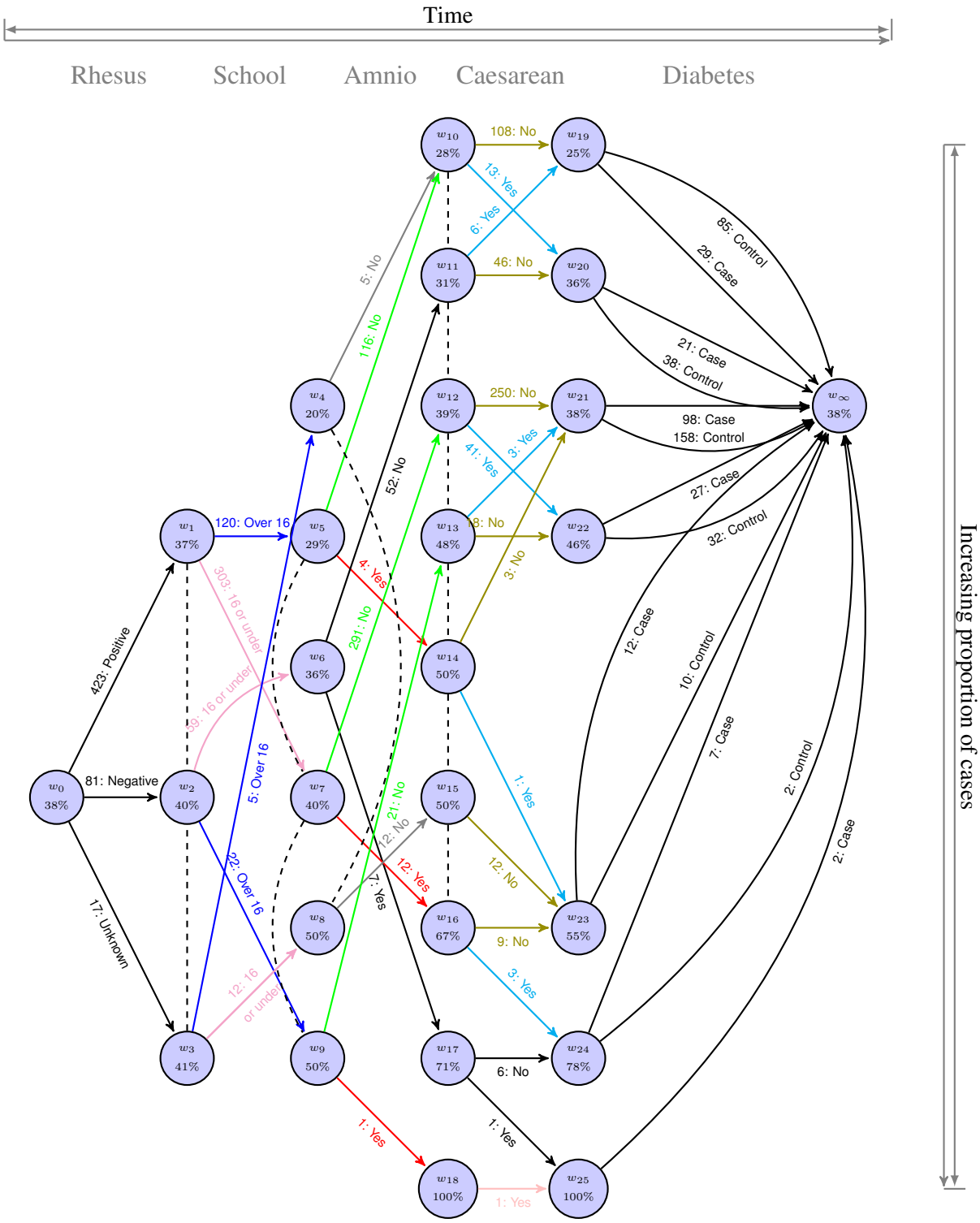

Figure 5.21: Pruned ordinal chain event graph for the five variables. Positions are labelled conventionally from left (w_0) to right (w_∞). Arrows show the chronological ordering and dashed lines show positions in the same stage. Numbers along edges illustrate the number of individuals taking the path and numbers in vertices show the percentage of cases.

The probability of disease in the case-control CEG is not the probability of disease in the population, but is instead the probability of disease in the sample given the path taken thus far. This is a result of the sample being selected by disease status, rather than randomly from the population. For example, of those who are rhesus positive, 37% have children who are cases, whereas for rhesus negative mothers, 40% have children who are cases. Rather than interpret this percentage directly, the value should be compared with the overall percentage of cases in the sample, which here is 38%. Therefore for rhesus factor, the three categories of positive, negative and unknown have 37%, 40% and 41% of the children being cases respectively. Since these values are similar to the starting percentage of cases (38%) then rhesus factor is concluded to have little association with diabetes in the child. However, it will be shown for the amniocentesis variable that no amniocenteses are generally associated with low percentages of cases, whereas at least one amniocentesis is associated with a high percentages of cases, relative to the 38% starting value. Therefore the vertices at the start and end of a CEG (w_0 and w_∞) will always display the same percentage and while it may seem unnecessary to include both vertices, the associated edges provide additional information. For example, whether or not both cases and controls are observed from a given position to w_∞.

Mothers with at least one amniocentesis are situated towards the bottom of the ordinal graph, suggesting a higher probability of their child being a case, whereas those with no amniocenteses are situated towards the top of the graph, suggesting a higher probability of their child being a control. There are just two vertices towards the centre of the graph which *appear* to disrupt this pattern (w_{14} and w_{15}), however, since their probabilities are the same, they could be switched. Therefore amniocentesis is clearly associated with the diabetes status of the child. For the delivery of the child, there is a less clear pattern. However, generally the children delivered by caesarean have a higher probability of being a case than those not. The edges from w_{10-18} to w_{19-25} are those which depict caesarean delivery, and all the 'yes' edges lead to lower positions in the ordinal graph than the 'no' edges, with only w_{11} and w_{13} as exceptions. For these exceptions, the edges for the two delivery options are only one vertex apart in the next variable, hence the difference in probability of disease is small. The combination of at least one amniocentesis and caesarean delivery can be found in one of the three bottom vertices (w_{23}, w_{24}, w_{25}), indicating a higher probability of case status, with 55–100% of the participants in these vertices having diabetes. The only other paths to lead to these bottom vertices were participants with mothers who left school

aged 16 or under. This interaction between amniocentesis and caesarean delivery is the most prominent in the CEG.

For vertices w_{19}–w_{25}, the paths containing no amniocentesis and delivery not by caesarean are positioned at least as high as those with at least one amniocentesis and caesarean delivery, showing the combination of these two variables to be associated with diabetes. Where there is only one of amniocentesis or caesarean, those with caesarean are generally positioned higher on the ordinal CEG than those with amniocentesis, suggesting at least one amniocentesis is more of a risk factor than caesarean delivery.

The vertex with the highest probability of being a case ($w_{25} = 100\%$) can be reached via two paths; both of which require only negative rhesus factor, at least one amniocentesis and caesarean delivery. This finding suggests the school-leaving-age is not strongly associated with the disease, while the other three categories may act as risk factors. The vertex with the lowest probability of being a case ($w_{19} = 25\%$) can be reached by three paths, all containing no amniocenteses, again suggesting amniocenteses are associated with the disease. Unpopulated paths also provide information, for example there are no paths with amniocenteses and unknown rhesus factor, which may suggest the rhesus factor category is recorded for an amniocentesis. Further conclusions for the rhesus factor variable will be provided in §5.4.4.1 and §5.4.4.2.

5.4.3 Sensitivity of the Findings

5.4.3.1 Sensitivity to Variable Ordering

In these data, the ordering of the rhesus factor and school-leaving-age variables were swapped and the same conclusions were drawn from the resulting CEG, as shown in Figure 5.22. This is to be expected since the rhesus factor and school-leaving-age of the mother are likely to act independently, hence their ordering is less important. Note that different colours have been assigned to Figure 5.22 so it is not confused with CEGs generated from the original ordering.

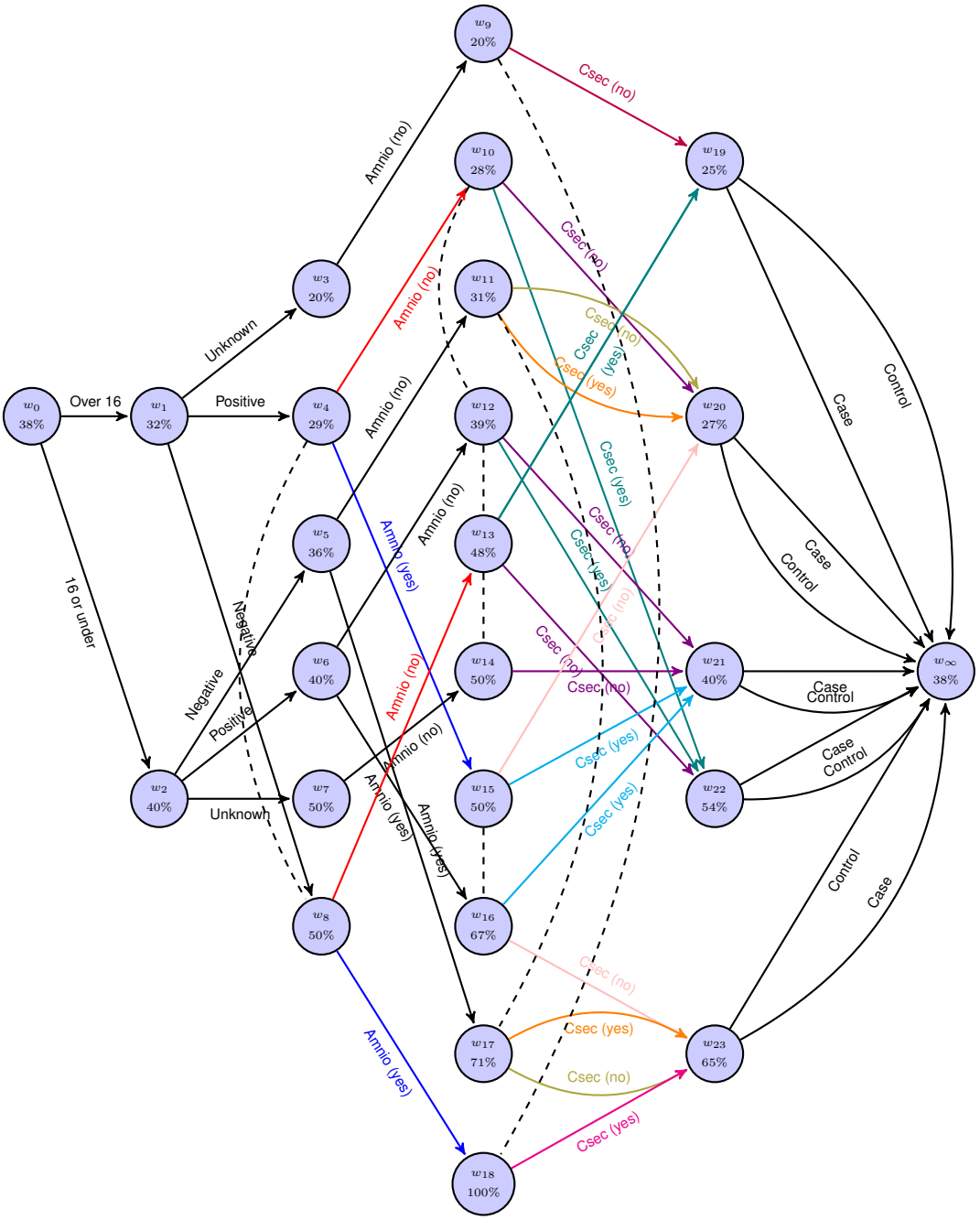

Figure 5.22: Pruned ordinal chain event graph for the diabetes dataset, five variables, missing data, unequal probabilities along each path, and the rhesus factor and school-leaving-age variables swapped. Csec = caesarean. Amnio = amniocentesis.

5.4.3.2 Sensitivity to the Equivalent Sample Size

When incorporating prior knowledge an equivalent sample size of 300,000 was adopted, which is larger than normally used since it assumes data from a comparatively large cohort study, but is appropriate since population data were used which incorporate a large proportion of the target population.[388]

However, the analysis was also conducted with an equivalent sample size of 30 and of 5, to test the sensitivity of the results to prior beliefs. The corresponding trees and CEGs are shown in Figures 5.23, 5.24, 5.25 and 5.26.

The three CEGs generated in Figures 5.21, 5.24 and 5.26 were identical for rhesus factor and school-leaving-age. The CEGs generated using the smaller equivalent sample sizes of 5 and 30 were also identical for amniocentesis. There were fewer positions when a smaller equivalent sample size was used; a total of 22 for Figures 5.24 and 5.26 compared with a total of 27 in Figure 5.21, with the allocation of the positions to the variables the same for Figures 5.24 and 5.26.

The smaller equivalent sample sizes therefore led to simpler graphs with fewer positions and hence fewer edges. However, all three CEGs drew very similar conclusions and hence the clinical interpretation was not altered by changes in the equivalent sample size.

5.4.3.3 Sensitivity to the Prior Knowledge

Also for comparison, the analysis was conducted using uniform priors, with the tree as shown in Figure 5.27 and the corresponding CEG as shown in Figure 5.28. The AHC algorithm used is shown in Appendix D.2, and the priors were chosen such that simple fractions were present in the calculations, as discussed in §5.2.2.2.

The CEG in Figure 5.28 again concluded little association with the rhesus factor and school-leaving-age of the mother, with diabetes in the child, but an association with amniocentesis and a less clear but possible association with caesarean delivery. This shows that the findings are not sensitive to the priors used.

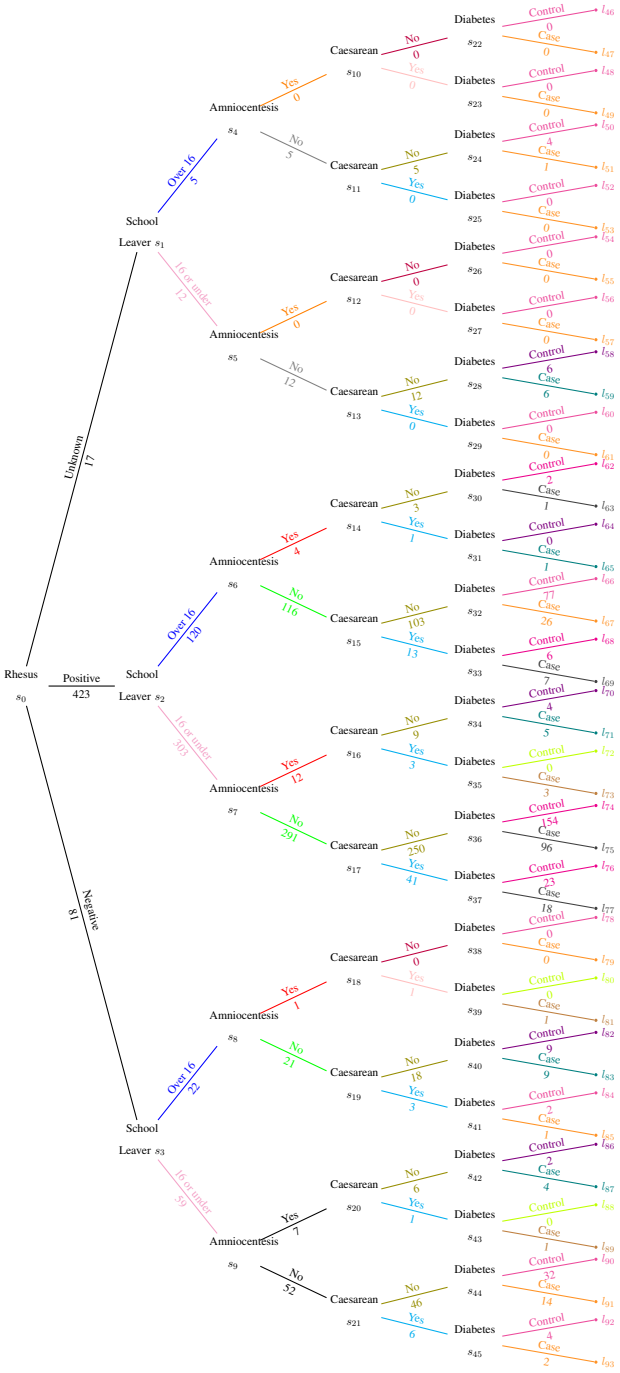

Figure 5.23: Staged tree for the four exposure and outcome variables; unequal probabilities along each path, equivalent sample size of 30.

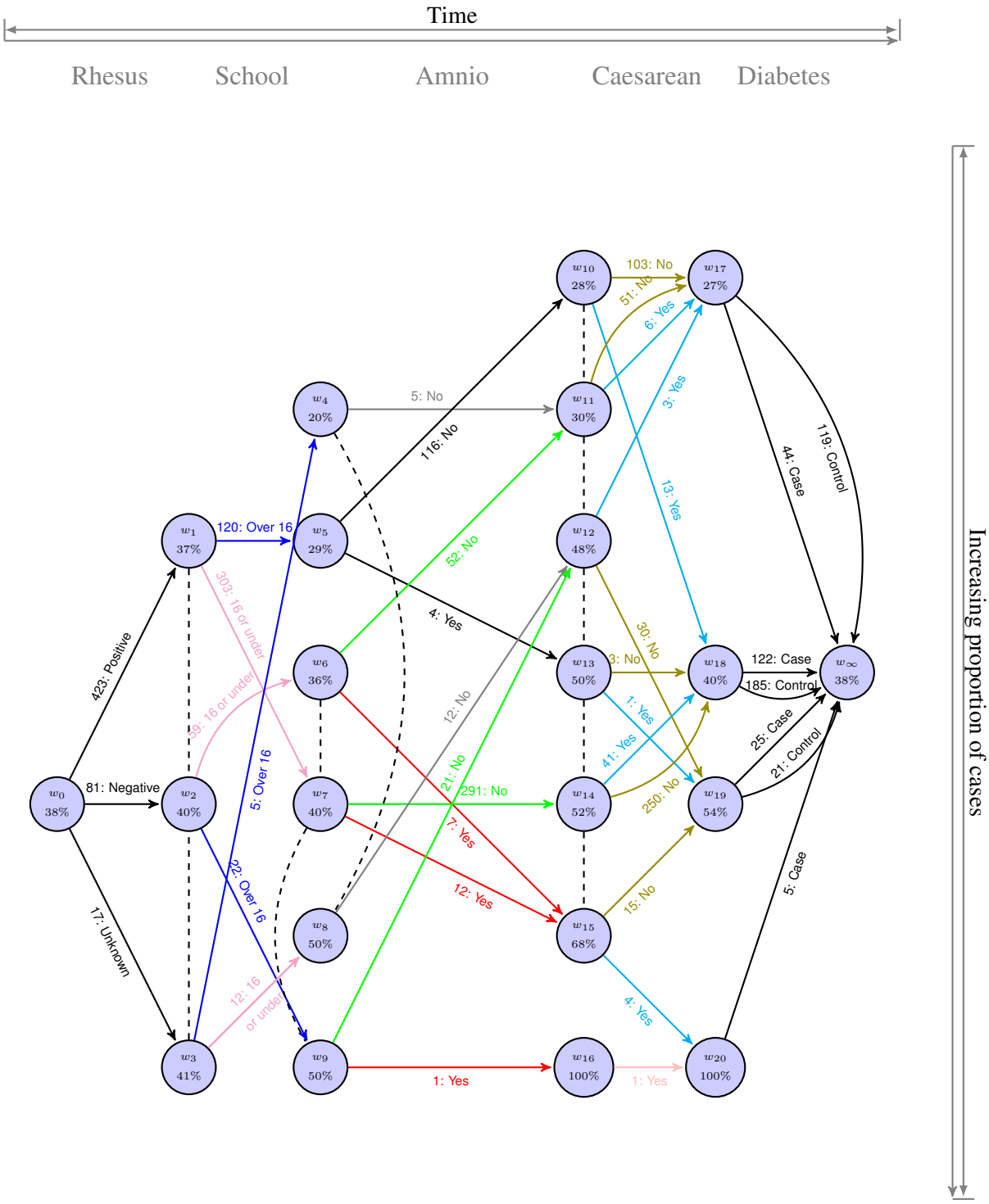

Figure 5.24: Pruned ordinal chain event graph for the five variables, generated using an equivalent sample size of 30. Positions are labelled conventionally from left (w_0) to right (w_∞). Arrows show the chronological ordering and dashed lines show positions in the same stage. Numbers along edges illustrate the number of individuals taking the path and numbers in vertices show the percentage of cases.

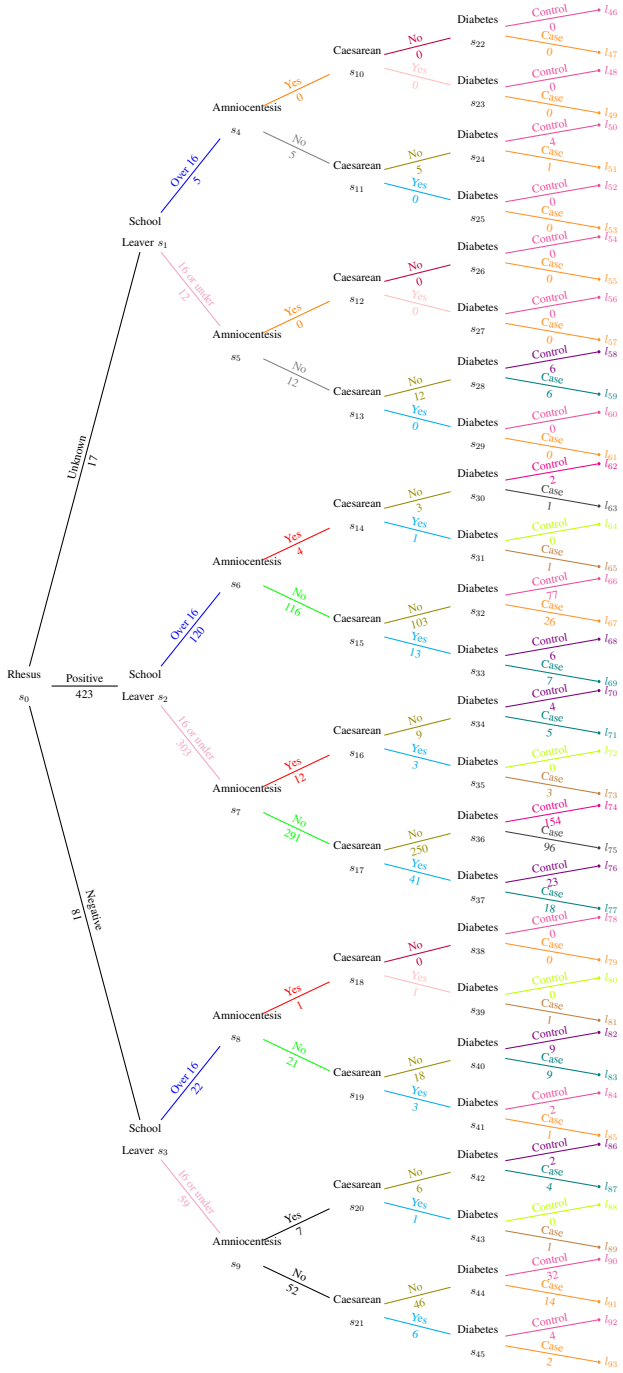

Figure 5.25: Staged tree for the four exposure and outcome variables; unequal probabilities along each path, equivalent sample size of 5.

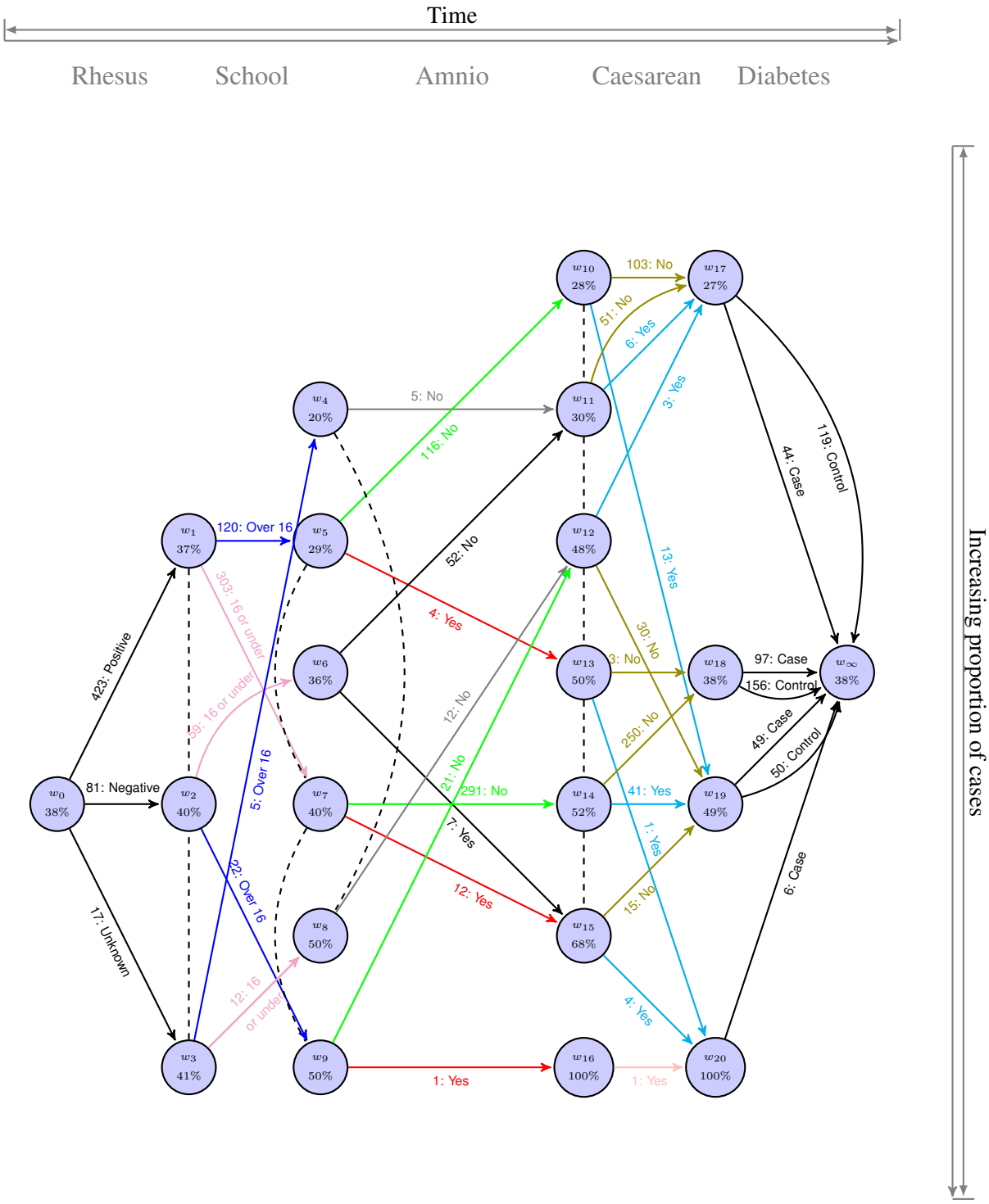

Figure 5.26: Pruned ordinal chain event graph for the five variables, generated using an equivalent sample size of 5. Positions are labelled conventionally from left (w_0) to right (w_∞). Arrows show the chronological ordering and dashed lines show positions in the same stage. Numbers along edges illustrate the number of individuals taking the path and numbers in vertices show the percentage of cases.

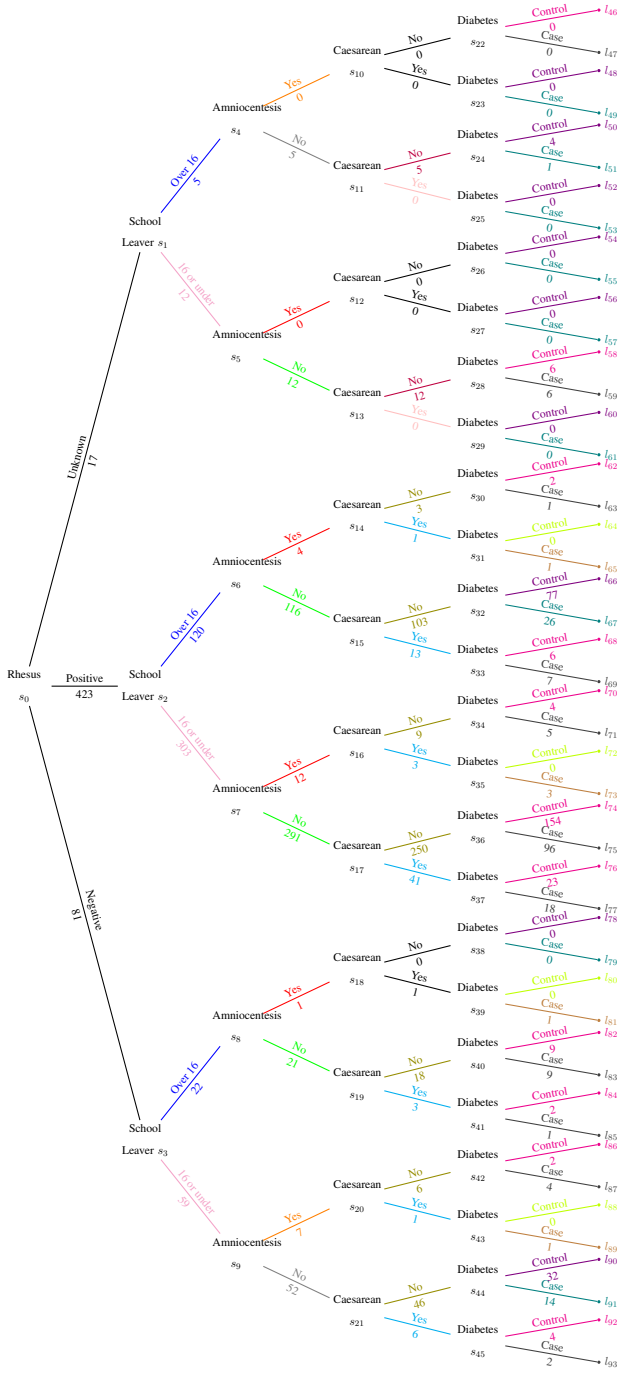

Figure 5.27: Staged tree for the four exposure and outcome variables; uniform priors.

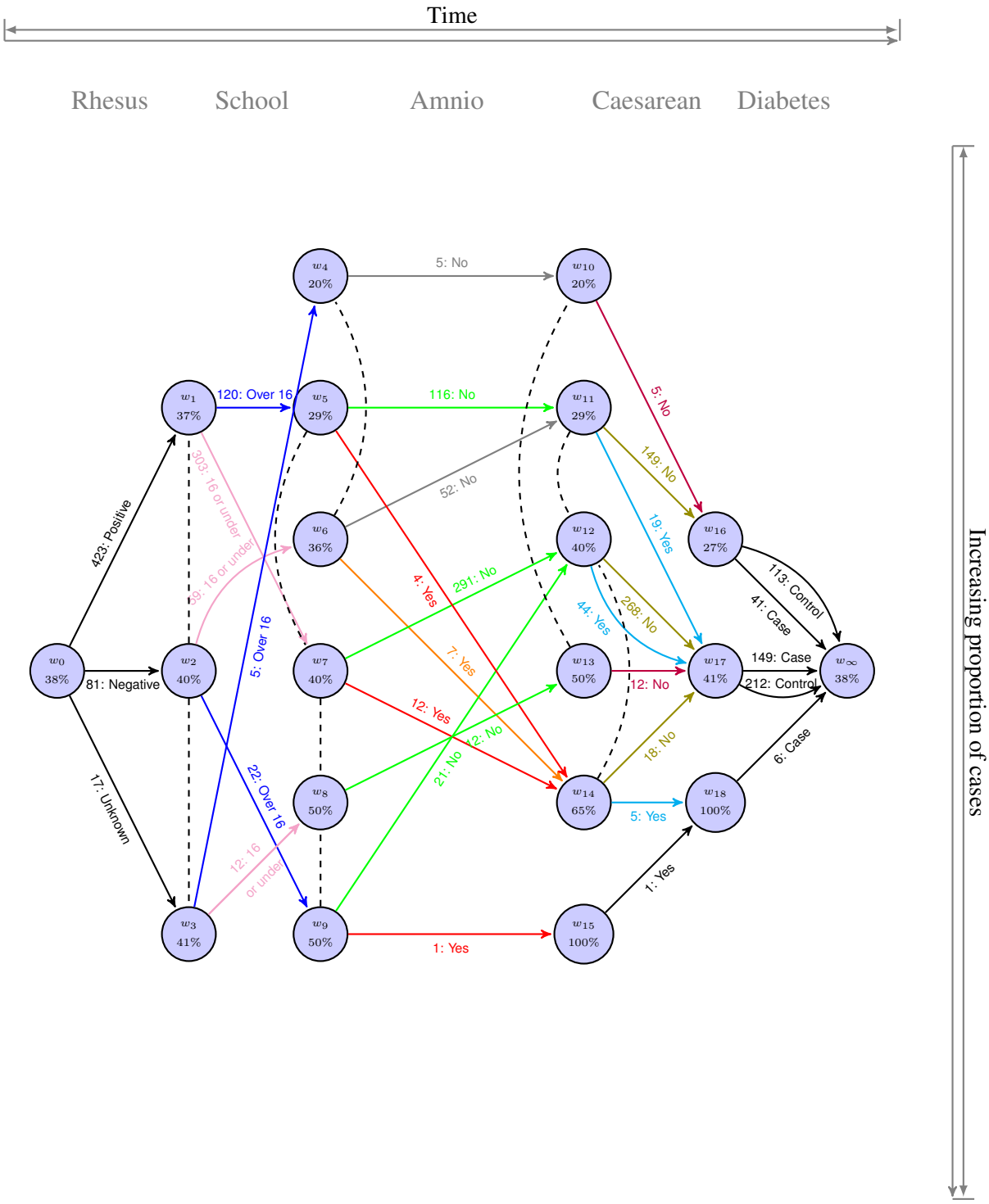

Figure 5.28: Pruned ordinal chain event graph for the five variables, generated using uniform priors. Positions are labelled conventionally from left (w_0) to right (w_∞). Arrows show the chronological ordering and dashed lines show positions in the same stage. Numbers along edges illustrate the number of individuals taking the path and numbers in vertices show the percentage of cases.

5.4.4 Missing Data Summary

5.4.4.1 Rhesus Factor Conclusions

The position of the unknown rhesus factor category in the ordinal CEG in Figure 5.21 can be used to draw conclusions about the missingness mechanism.[365] Since the unknown category (w_3) is positioned at the bottom of the ordinal CEG, underneath positions w_1 and w_2 which represent known rhesus factor, it is assumed that the rhesus factor data are MNAR, since those with missing values are associated with a (marginally) higher probability of being a case than either the rhesus positive or rhesus negative categories. If the data had been MAR, the missing values would be expected to be a combination of the recorded values in a proportion similar to those in the data, and hence the missing category would be positioned between the recorded categories on the ordinal CEG. However, the small percentage differences in positions w_1–w_3 should be noted (37%, 40%, 41%).

If the missing data in the rhesus factor variable are MNAR then an additional level of detail can be reported for how the missingness is structured. However, this additional level of detail relies upon there being at least one variable in the event tree before the variable which contains missing data. In Figures 5.21, 5.24, 5.26 and 5.28 rhesus factor is the first variable in the CEG and hence there are no variables preceding it to guide this extra level of detail. However, in Figure 5.22 the order of the school-leaving-age of the mother and her rhesus factor category were swapped, allowing this extra level of detail for the MNAR rhesus factor variable to be reported.

Figure 5.22 shows that rhesus factor is MNAR for mothers with either school-leaving-age category. When the school-leaving-age is over 16 years, the unknown rhesus factor category shows a lower percentage of cases ($w_3 = 20\%$) than either of the recorded categories (positive, $w_4 = 29\%$ and negative $w_8 = 50\%$). When the school-leaving-age is 16 years and under, the unknown rhesus category shows a higher percentage of cases ($w_7 = 50\%$) than either of the recorded categories (positive, $w_6 = 40\%$ and negative, $w_5 = 36\%$). This could suggest that for mothers with a school-leaving-age of over 16 years, the unknown rhesus category are MNAR and associated with fewer cases than those who are rhesus positive or negative, and suggest a protective effect, while that for mothers with a school-leaving-age of 16 years or under, the unknown category are MNAR and associated with more cases than those who are rhesus positive or negative, and hence suggest an

increased risk. In both categories of school-leaving-age, the CEG shows that the missing rhesus factor categories are likely to be mainly positive, since they offer a more extreme association with the outcome than the next closest category, with is positive rhesus factor. The structure of the missingness in the ordinal CEG also suggests that the missingness has a relatively strong influence, since the MNAR mechanism is not simply equal to one of the positive or negative categories, but is instead more extreme than the positive category.

There are instances under this parameterisation where paths are identical except for positive/unknown rhesus factor and lead to the same vertex, such as w_{10} and subsequent paths, which suggests that a large proportion of the unknowns may be positives. Since around 86% of the population are rhesus positive this would be a reasonable conclusion, and agrees with the conclusion of the rhesus factor variable being MNAR and likely positive, as stated above.

5.4.4.2 Suitable Methods to Reduce Bias When Rhesus Factor is Missing

As the rhesus factor variable had missing data, methods designed to reduce bias associated with these missing data can be used. Chapter 4 presented and discussed such methods and a flowchart was developed to aid the selection of a suitable approach. When using the flowchart in Figure 4.4, the first question is whether the study is potentially affected by participation bias. A DAG could be used as described in §2.3.4.1, to determine whether the study is likely to be affected by participation bias. The reason for unknown rhesus factor here is not recorded, however, it is possible that the requirements for participation bias are satisfied. Next, the flowchart tool asks whether the variable associated with participation is recorded. Since the diabetes dataset has no data available regarding non-participation, the answer to this question is no, the variable associated with participation is not recorded. The third question is whether relevant population data are available. Recall the population values which were used in Table 5.1 to calculate the priors used in the AHC algorithm. These values showed the rhesus positive and rhesus negative proportions in the population, hence sensitivity analyses and weighting may be suitable methods to adopt if these population values are sufficient. If the population data are not as required, the final question in the flowchart is whether non-participant data were collected. Since no information on participation was available for these data, the flowchart guides towards sensitivity analyses as a possible option.

The flowchart in Figure 4.4 has therefore eliminated all but weighting and sensitivity analyses.

Weighting was described in §4.4.1 where one assumption was that the data were MAR. The CEG analysis in this chapter has reported the rhesus factor missing values are unlikely to be MAR, hence a sensitivity analysis is the only method which remains. This example highlights the possible need for more methods which are suitable when data are MNAR, and this need will be explored further in Chapter 7.

While the conclusions drawn in §5.4.4.1 are a direct result of the structure of the ordinal CEG in Figure 5.22, it is possible that the data are not truly MNAR and that the structure resulting from the missingness is not informative, since the percentage of cases with positive/negative/unknown rhesus factor are all around 40%. If the data are not MNAR, there are additional choices for the methods to reduce participation bias if required, as shown §4.5.

5.4.5 Diabetes Data Summary

CEGs have been introduced in this chapter, not for their usual application of data analysis, but instead specifically for exploring missing data as a result of non-participation. CEGs have not been used before with case-control study data (§5.1.2) and therefore non-participation in case-control studies has not been investigated using CEGs. As demonstrated here, CEGs have the potential to be used with missing data and missingness resulting from non-participation in case-control studies, and if the missingness mechanism is investigated, this can subsequently assist with the selection of a method to reduce the effects of participation bias as discussed in Chapter 4. The unknown rhesus factor values in the diabetes data were likely to be MNAR, and mainly rhesus positive, and sensitivity analysis was concluded to be the most suitable method for these missing data.

All conclusions drawn here are based upon the 521 individuals in the case-control study conducted in Yorkshire,[7] which may not be representative of other areas. It is acknowledged that a different dataset may lead to a different CEG with different conclusions and hence analyses should be encouraged, prospectively or retrospectively, of other diabetes datasets as agreement between studies would strengthen findings. Associations are reported here and of course there may be unrecorded variables which are more closely associated with type I diabetes for which these recorded variables are acting as a proxy. These results nevertheless offer additional insight into the factors associated with type I diabetes.

The cases and controls in the diabetes data were matched by age and sex when recruited. Traditional analyses would be required to take this feature of the study design into account, using approaches such as conditional logistic regression. In CEGs, where cases and controls are matched by a given variable, there will be a predefined proportion of cases per vertex for the matched variables. For example, let there be a study with a 1:1 ratio of cases and controls, matched by gender. Let there be 20 females cases, 20 female controls, 10 males cases and 10 males controls. At w_0, 50% of the individuals will be cases. Since the cases and controls are matched by gender, there is also this percentage (50%) of cases in each of the male and female groups. Therefore, the AHC algorithm in §5.2.2.1 would group males and females into the same stage. Males and females may differ with respect to other variables included in the analysis, and hence the male and female categories may not be in the same position. In the resulting CEG, if males and females are in the same position, there would be $\frac{30}{60} = 50\%$ of the individuals who are cases. If males and females were in different positions, there would be a dashed line between the vertices denoting that they were in the same stage, and there would be $\frac{10}{20} = 50\%$ of the males and $\frac{20}{40} = 50\%$ of the females who were cases. Therefore, including matching variables in the CEG is not necessary. Similar to conditional logistic regression, it would result in there not being information available regarding the matched variable.

5.5 Summary

This chapter has demonstrated the application of an existing method to a new field, namely CEGs to case-control studies. The mathematics for CEGs was not provided here since it is available in the literature[8, 21, 365, 379] and repetition here would not add to current knowledge. Instead the theory for CEGs was applied to case-control data and used to drawn conclusions about the missingness. CEGs here are suggested as an exploratory rather than analytical tool, as an intermediary tool to guide the further analysis of a dataset containing missing data. The results from the diabetes CEG contributed towards the selection of a method suitable to reduce bias resulting from non-participation. The conclusions for both the variables and the missingness were found to be insensitive to changes in the priors, the strength of the priors, and the ordering of the first two variables.

5.5.1 Critical Evaluation of Chain Event Graphs: In General

5.5.1.1 Advantages

CEGs have advantages over traditional methods. For example, they allow prior information to be incorporated in the analyses, which approaches such as logistic regression do not as standard. While methods such as Bayesian logistic regression are available, they are not common practice in calculations following case-control studies. More generally, CEGs are a graphical approach, which may be preferable to numerical approaches for some researchers, and which offer an alternative means by which to communicate complex statistical models to clinical experts who may not be statisticians. CEGs may be particularly useful when discussing interactions between variables, which may be easier to follow on a tree than through terms in a regression model. However, CEGs are much more time-consuming to analyse and produce than logistic regression or similar methods.

For comparison, the logistic regression model for amniocentesis and caesarean delivery was produced, with each variable included plus their interaction. The resulting odds ratios are given in Table 5.2. Often logistic regression models do not include the interaction term, and in instances such as these where the interaction term is 3.10×10^{6}, with an implausible confidence interval, the interaction term would be removed from the analysis and not reported. However, there is a potentially important finding here that *all* mothers in the study who had at least one amniocentesis and who delivered by caesarean, had children who were cases. This finding is clear in CEGs but not clear from logistic regression, even when an interaction term is used. There is the limitation that there are *no* controls in the sample who were born to mothers who had at least one amniocentesis and who delivered by caesarean, and there are only six cases in this category, but these associations could have important clinical implications.

The non-parametric nature of CEGs can be advantageous. For example, CEGs could be used when assumptions for traditional methods are not met, such as the rare-disease assumption for odds ratios or regression assumptions in modelling. Sparsely populated categories can also be troublesome during numerical analyses, but there are procedures in place for CEGs such as pruning the tree, combining edges or representing sparse edges using dotted lines.[8]

Case-control studies are retrospective and this is often considered to be a negative feature of the

Variable	Odds ratio estimate	Lower CI	Upper CI
Intercept	0.54	0.44	0.66
Caesarean	1.48	0.865	2.53
Amniocentesis	2.32	0.90	6.19
Interaction Caesarean:Amniocentesis	3.10×10^{6}	1.30×10^{-18}	NA

Table 5.2: Logistic regression model with amniocentesis and caesarean delivery, plus their interaction term. CI = confidence interval.

study design, but one which may be advantageous for CEGs. Firstly, the number of variables and time period covered is known before analysis, and hence avoids the need for more complex graphs such as dynamic CEGs.[381] Expert knowledge gained over time can also be incorporated into the analysis and inform paths which are more likely, or eliminate any paths which are not clinically plausible, in the same way as the data in Table 5.1 were utilised. One disadvantage of the retrospective study design may be the unclear variable ordering, upon which case-control study CEGs depend. The ordering of some variables will be obvious, while others may have occurred at seemingly the same time. For example, two variables such as amniocentesis and the occurrence of x-rays during pregnancy may be difficult to order chronologically. One way to circumvent this issue could be to create a new variable combining the two; there could be categories of 'x-ray and amniocentesis', 'x-ray but no amniocentesis', 'no x-ray but amniocentesis' etc, covering all combinations of the two variables.

5.5.1.2 Limitations

One drawback of CEGs is the time required to calculate the stages and positions, and to draw the CEG before interpretation. To generate the CEG the study data must first be represented using a spreadsheet, with each column containing just one categorical variable, and each row containing just one individual. Next, the data must be read into a statistical software package (R[380] if the algorithm code from Appendix D.2 is to be used). If required, non-uniform priors need to be specified as a list (and the code from Appendix D.3 used instead). The algorithm is then run, which returns the stages (example output is given in Appendix D.5). It is necessary to use the stages generated, in conjunction with the more detailed aspects of the output such as the results

section, to ensure the correct vertices are identified. The remainder of the process is conducted by hand. First, it is useful to draw the event tree and colour edges from corresponding vertices with the same colours, which aids the identification of positions. Next, any vertices in the same stage must be checked to determine whether they are also in the same position. A list of vertices in the same stage and a list of vertices in the same position are recorded. A CEG can then be drawn by hand, where the list of positions forms the vertices and corresponding edges, and the list of stages enabled dashed lines to be drawn between vertices in the same stage but not the same position. Finally, the CEG can be interpreted. Since hand-drawn diagrams are not desirable, the trees and CEGs can be drawn in LaTeX for improved readability, although this is also time-consuming.

The majority of CEGs used in this thesis have used binary covariates, but examples have been given where covariates have more than two categories. The number of categories can be further increased to approximate continuous covariates, but CEGs cannot truly accommodate continuous data, leading to a loss of information in some examples. However, the application of the CEG will determine whether this loss is important for the conclusions being drawn.

The focus of CEGs here has not been for the analysis of a case-control study, but instead as a tool to explore the missingness resulting from non-participation. CEGs are used elsewhere (§5.1.2) to analyse data, but this use does not result in an odds ratio as may often be desired in case-control study reports. If CEGs were to be added to the flowchart in Figure 4.4 in §4.5, they could be positioned either with 'adjust for, stratification and propensity score' when the variable associated with participation is recorded, or with 'imputation, weighting and sensitivity analysis' when non-participant data are collected. CEGs can be used during analysis regardless of how much missing data there are. Often the outcome will be recorded, even if miscategorised, or it will be assumed that individuals not recorded as cases are controls, therefore conclusions regarding missingness by cases and controls can be drawn. If *all* data are missing from a group of individuals, these individuals can still be added to the CEG, to provide information about the patterns of missingness and summarise the amount of missingness, although will be less informative than if at least one variable was recorded. Here it is suggested that CEGs are used to investigate missingness, before an analytical method is applied, such as those in Chapter 4 or standard case-control analyses as in §2.2.5 if appropriate. CEGs can complement traditional analyses, and the combination of the model-based and graphical methods can give an extensive analysis for a case-control study.

5.5.2 Critical Evaluation of Chain Event Graphs: For Investigating Missingness

An advantage of CEGs is that individuals with missing data need not be excluded from the analysis, as a 'missing' category is included in the tree as an edge and therefore classed as an informative category. This provides information about the individuals with a missing variable compared with the individuals who have a recorded category. Although this can be achieved with logistic regression, it is not always practiced, and approaches such as complete case analysis are often chosen. This information regarding the missingness can be used to determine whether the data are likely to be MAR or not, and thus whether methods such as multiple imputation are possible, as shown in §5.4.4.2. The information can also be used to estimate the likely category of the missing data. However, it must be noted that no conclusions are drawn about the *individuals* within the missing category, but rather that conclusions are drawn about different subgroups of the dataset.

Logistic regression could be used to explore the association between the missing values and recorded variables, by specifying binary missingness as the outcome and the recorded covariates as the explanatory variables, in an approach similar to the calculation of the propensity score (see §4.4.3). This approach would identify variables associated with missingness, but would not directly incorporate the missingness into the analysis. The advantage of the CEG approach as described here, is that the missingness forms an informative category during the analysis, and also considers the missingness in each variable in conjunction with the recorded and missing values in other variables. For example, CEGs can highlight that females are less likely to record an older age, whereas males are just as likely to have a missing age value at any age. This level of detail regarding missingness would not be identified using the logistic regression approach, without including interaction terms. The use of CEGs for investigating missingness as a dependent variable is considered further in Chapter 6. The advantage of the logistic regression approach, is the ability to perform the analysis quickly and easily, and using statistical software.

5.5.3 Chain Event Graphs in the Identification of an Appropriate Method to Reduce Bias

The main focus for CEGs in this chapter has been their ability to identify whether data are likely to be MAR, which is a difficult task to perform from the data alone. Exploration of the missingness mechanism in the rhesus factor variable of the diabetes data allowed a suggestion of the likely category for the missing data, a conclusion that the data were unlikely to be MAR, and a suggestion that a sensitivity analysis was a sensible method to adopt to reduce the effects of any bias resulting through these missing data.

Therefore CEGs, in conjunction with the flowchart tool in §4.5, can be used to identify methods which may be suitable for bias reduction. The CEG can also provide analysis of the dataset in addition to conclusions regarding the missingness mechanism. While identifying a method suitable for use with the rhesus factor variable, it was highlighted how few methods are suitable when the data are MNAR. Chapter 7 therefore aims to develop a new method suitable for these instances, which can be added as an alternative method to the flowchart tool in §4.5.

5.5.4 Further Work

Further work could include the application of CEGs to other datasets with missing data to investigate the missingness mechanisms, or to additional variables in the diabetes dataset which have missing data. As shown with the diabetes dataset, the sensitivity of CEGs can be tested with respect to changes in the priors or strength of the prior beliefs. Where conclusions regarding missingness are robust, the selected method to reduce participation bias should be presented with increased confidence.

CEGs have thus far been used in the literature (§5.1.2) with a disease of interest or similar as the final variable. However, in studies which suffer from non-participation, CEGs may be more useful as a means by which to investigate non-participation. This application will be developed and demonstrated in Chapter 6.

Chapter 6

Chain Event Graph Adaptations for Use With Case-Control Data

Chain event graphs (CEGs) were introduced in Chapter 5 as a method to explore the missingness in case-control studies, including missingness resulting from non-participation. In this chapter, CEGs will be adapted for use specifically with case-control data to further investigate non-participation. Seven adaptations to the graphs will be presented; (i) to see how missingness varies with the severity of a disease (§6.1.1), (ii) to see how recruitment varies with the data collection approach used (§6.1.2), (iii) to report the characteristics of those who participate (§6.2.1), (iv) how these characteristics differ between cases and controls (§6.2.2), (v) how a form of meta-analysis can be conducted using data for similar (but not identical) studies (§6.2.3) regardless of data missing from non-participation or differing recorded variables, (vi) how the analysis can be adapted to incorporate the reliability of different data sources (§6.3.1) which may be affected by non-participation, and (vii) how subsets of the data can be analysed separately depending on the outcome of interest (§6.3.2).

These adaptations extend the ideas of the characteristics and methods which are associated with participation from §2.3.2 and §2.3.3, and propose new ways to investigate the factors associated with participation. These adaptations are designed to assist with the understanding of participation in case-control studies and identification of where bias may occur, using a graphical approach which has not been used before with either case-control studies or for the investigation of participation. The findings from these adapted CEGs may also assist with recruitment in future

case-control studies; not to increase participation as such, but to balance the characteristics between groups of individuals with the aim of reducing bias.

An example is provided for each of the adaptations to illustrate the steps from the event tree to the CEG, provide interpretation of the graph and explain its use in case-control studies. The diabetes dataset from Appendix A unfortunately does not have information available regarding non-participants and other aspects of the study required in this section, hence small hypothetical examples are used but the application would not change for real data.

6.1 Study Design Adaptations

6.1.1 Missingness by Disease Severity

Case-control studies usually record a binary outcome of case or control. However, diseases often have different severities such as terminal or not, and this level of detail is likely to be clinically useful and hence recorded in the medical notes. The tree and corresponding CEGs can therefore have additional edges denoting the possible severities of the disease such as control, mild case or severe case.

The CEG can be formed in the usual manner as described in Chapter 5, and as before conclusions can be drawn regarding the combinations of variables associated with the range of severity outcomes. It is possible that only individuals with a particular characteristic or combination of characteristics are able to possess the most severe category of the disease and this information may otherwise be hidden in a standard case-control analysis. The case categories can of course be collapsed and the data analysed with a binary outcome for comparison with previous analyses.

It was demonstrated in Chapter 5 that missingness can be investigated in CEGs. It may be that the missingness mechanism differs for each severity category. For instance, there may be more missingness amongst cases who have the more severe version of the disease and the missingness may be in variables which require input from the patient. In some studies these differences may be useful to highlight where missingness is occurring and in subsequent studies, different data collection strategies may be adopted, such as data collection for all participants only through medical records.

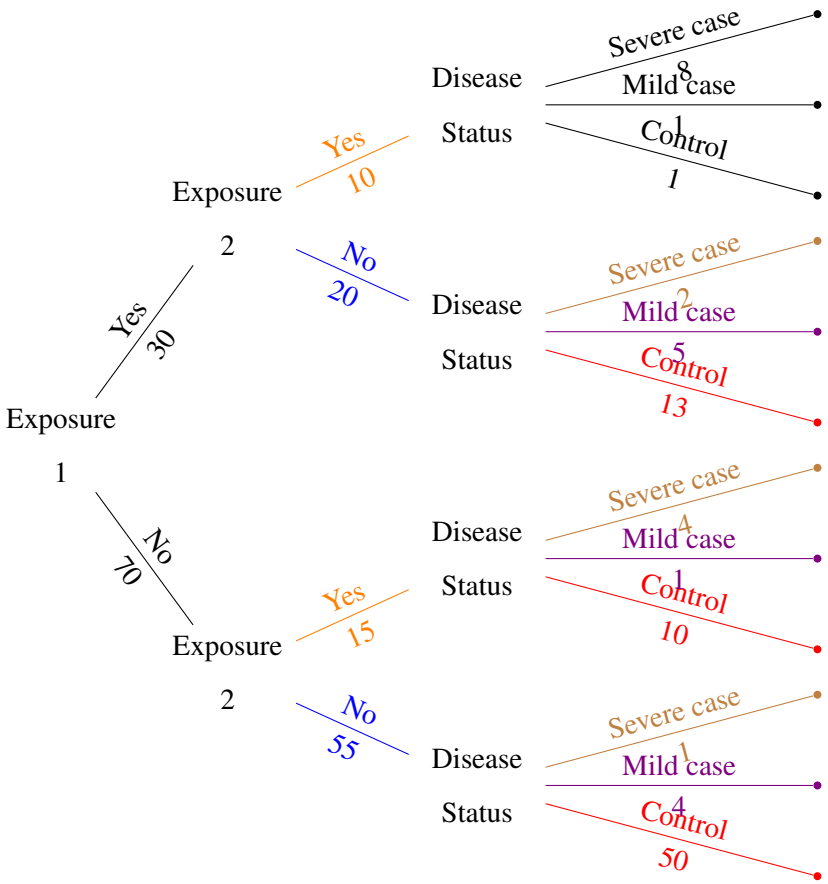

Figure 6.1: Severity staged tree.

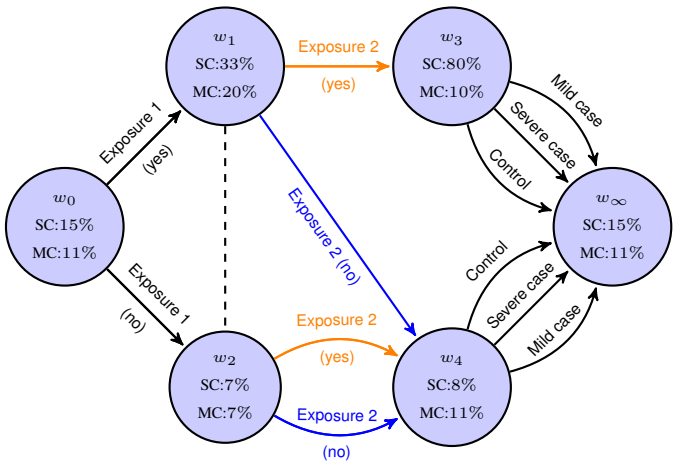

Figure 6.2: Chain event graph for severity. Percentage of severe case (SC) and mild case (MC) individuals shown at each position.

Hypothetical Example

Let there be a case-control severity study which consists of a control group plus two categories of cases; mild case and severe case. Let there be two independent exposures of interest, each of which is binary. The staged tree for the data is shown in Figure 6.1 and the corresponding CEG is given in Figure 6.2. The CEG shows that when only one exposure is present (w_4), the individuals have generally the same probability of the three disease categories as when no exposures are present (SC: 8%, MC: 11%). However when both exposures are present (w_3), the individual is associated with an increased probability of being a severe case (SC: 80%). Exposure 1 alone (w_1) increases the probability of an individual being a case, for both severities (SC: 33%, MC: 20% compared with SC: 7% , MC: 7%). A similar CEG could be constructed with missing values as shown in §5.3, to investigate missingness with respect to the disease severities. For example, missing edges may only lead to a severe disease status, while recorded edges may lead to any of the three disease categories.

6.1.2 Recruitment by Data Collection Method

The effectiveness of different recruitment techniques or data collection strategies (such as web surveys, postal surveys, and electronic reminders) may be of interest. Rather than use a CEG for case-control data which shows personal characteristics along each path, a CEG can be developed which contains solely information about the data collection approaches adopted. With the binary disease status forming the final vertices in the event tree, the CEG can be used to determine which approaches are more associated with cases and which are more associated with controls.

Individuals who participate by face-to-face interview may differ from those who participate using an online survey. Therefore the findings from these CEGs could be used to recognise where differences in the characteristics of cases and controls may occur, and to balance the two groups if necessary. Although this approach could highlight methods which are most successful for recruiting case and controls respectively, adopting these approaches would firstly assume a causal effect rather than an association, and secondly could result in bias by recruiting the disease groups in different ways.

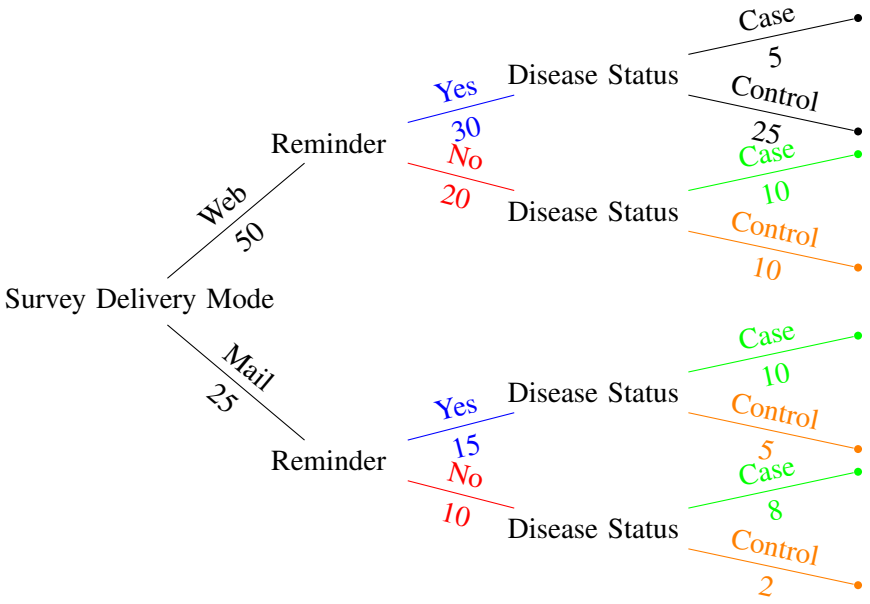

Figure 6.3: Data collection staged tree.

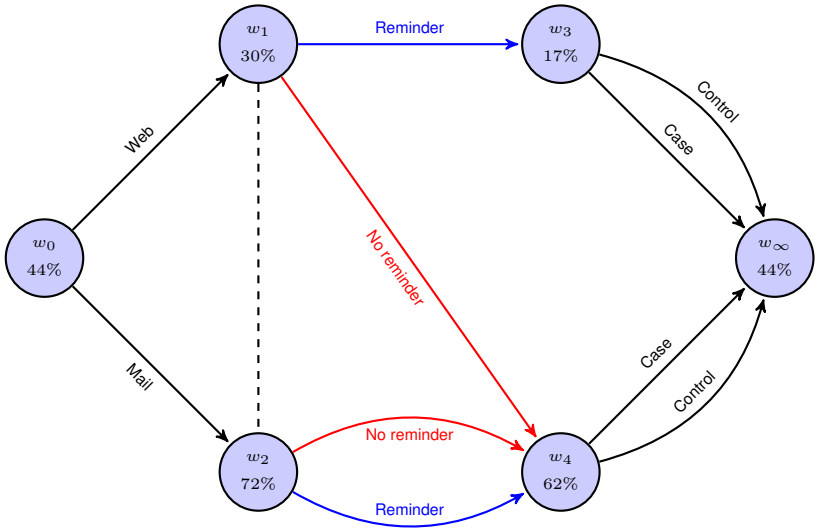

Figure 6.4: Chain event graph for data collection. Percentage of cases shown at each position.

Hypothetical Example

Let there be a hypothetical survey conducted. Figure 6.3 shows that 25 of the participants were recruited by mail, and 50 were recruited by a web survey. The quantities of reminders required and the disease status recorded from the survey are also shown. This tree can be used to summarise which data collection techniques are more associated with case recruitment and which techniques are more associated with control recruitment.

The CEG formed using this information is given in Figure 6.4. Mailed surveys, or web surveys without reminders, recruited a group consisting of around 62% cases and 38% controls. Web surveys with reminders recruited a greater proportion of controls (around 83% controls and 17% cases). Mailing alone was more successful at recruiting cases (around 72%), while web surveys were more successful at recruiting controls (around 70%). These percentages can be compared directly, since the study consisted of approximately half cases and half controls.

6.2 Participation Adaptations

Thus far, in this thesis and more widely, CEGs have been used with a disease status or similar as the outcome of interest. CEGs have not before been used to investigate participation as the outcome of interest. Therefore, this section proposes adaptations to the graphs where the final variable represents participation.

6.2.1 Participation as the Outcome of Interest

Here it is demonstrated how CEGs can be used to further investigate non-participation in case-control studies. Typically in CEGs the final vertices of the event tree represent the outcome of interest. For case-control studies this is the disease status of the individuals and thus presents two options; case or control. If the outcome of interest is instead non-participation, the event tree can be restructured such that the final vertices represent participation (yes/no). Data collection techniques or individual characteristics then form the paths in the tree. The CEG highlights which combinations of techniques or characteristics result in comparable probabilities of participants, and an ordinal CEG can be used to order the combinations of categories from those associated

with the lowest probability of participation to the highest. This approach provides a summary of the participation rates as well as giving information on which factors or their proxies are associated with higher participation.

This use of CEGs could be extended to more than two participation categories. For example the final vertices could have vertices representing "no participation", "partial participation" and "full participation", where partial participation relates to those willing to give demographic data but no sensitive data, or for those willing to participate in a questionnaire but not in a subsequent interview. A similar approach could be used for recruitment phases, where each variable represents a study phase such as first contact, reminder, second reminder and so on, with the outcome of interest being whether the individual participated, refused or ignored, along with their reason for non-participation if given.

Hypothetical Example

Let there be 50 copies of a survey distributed by mail and 100 distributed using a cheaper web option. Reminders are sent to 40% of the mail recipients and 50% of the web recipients, since electronic means are cheaper than postage costs. Some individuals return a completed survey, while others do not. Figure 6.5 shows these variables in a tree along with the number of individuals taking each edge; the corresponding CEG is shown in Figure 6.6.

The CEG shows that distribution by web and mail generally result in the same probability of receiving a completed survey (50%), but reminders are associated with increased participation rates. Those designing the survey may consequently choose to distribute web rather than mail surveys to save costs, but to include reminders to those who do not participate in the first phase. This is important, since often those who respond to reminders differ from those who responded to the initial survey and differ again to those who do not respond at all (see Chapter 2). Therefore, while this tactic appears to increase equality and reduce bias by increasing participation rates, it is possible that this may in fact lead to increased bias by recruiting different participant characteristics, possibly in each disease group. Therefore, this CEG may need to be used in conjunction with a CEG similar to that introduced in the next section (in §6.9), to compare the characteristics of the individuals.

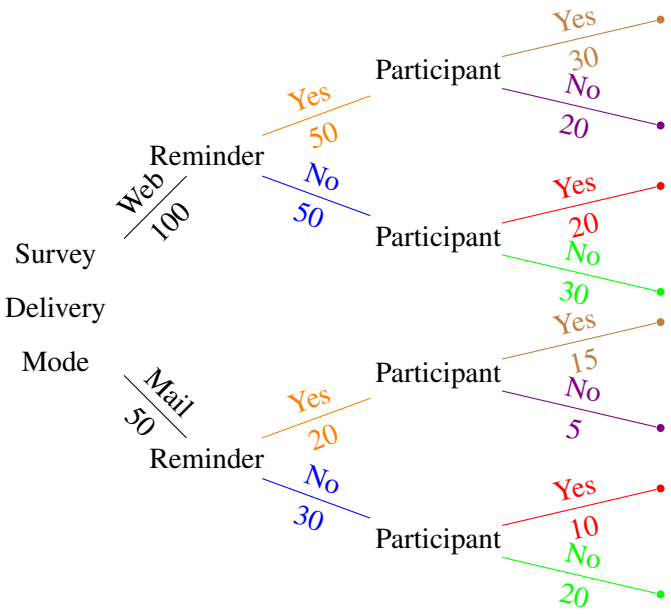

Figure 6.5: Participation staged tree.

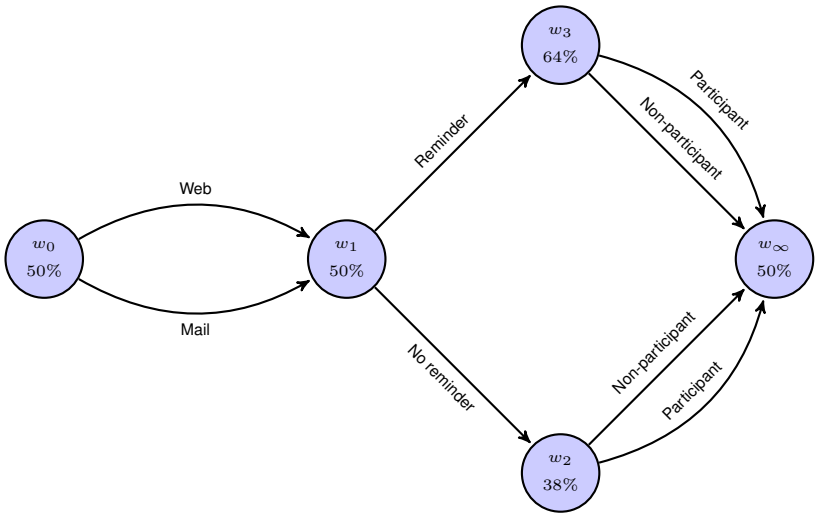

Figure 6.6: Chain event graph for participation. Percentage of participating individuals shown at each position.

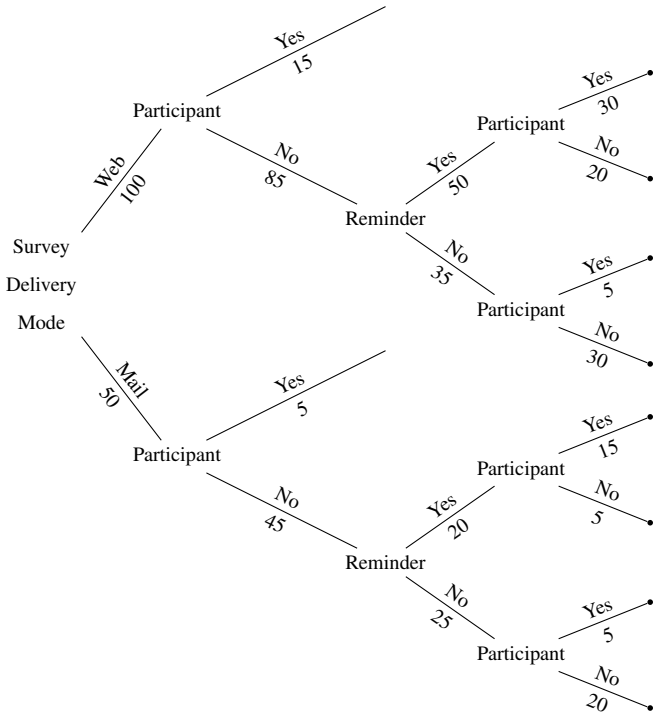

Figure 6.7: An example of an asymmetric tree.

In addition, CEGs can be used for asymmetric problems, meaning that a particular decision at one variable can affect the choices available for later variables. This allows the tree here to be restructured such that the chronological ordering is web/mail, participant/non-participant, reminder/no reminder (only for those who were non-participants after the first phase) and finally participant/non-participant. Therefore the paths through the tree would be of varying lengths, as shown in Figure 6.7. It may be important to distinguish between participants from the first phase and participants recruited after the reminder phase, and this can be achieved by using different category names for the two groups of participants.

Another example of an asymmetric tree may be where controls are offered face-to-face interviews or web surveys whereas cases are only offered face-to-face interviews, as reported in Chapter 2, but which is also known to contribute to differences between the disease groups.

6.2.2 Participation by Disease Group

This adaptation investigates how participation differs between the case and control groups, and builds on the findings of factors associated with participation discussed in §2.3.2.

It is of interest to learn about the factors associated with participation for the case and control group separately, since it is expected that the different disease groups will have different reasons for choosing to participate. The CEGs used in this and the previous chapter have ordered the variables chronologically. However, different orderings can be adopted by CEGs depending upon their use.[389] If knowledge regarding the factors associated with participation for the cases and controls is required separately, disease status can be placed as the first variable in the tree, and participation as the final variable. This ensures the cases and controls are reported separately regarding participation. Here the CEG shows which variables are associated with a higher or lower probability of participation whilst considering the disease status of the individuals, which is expected to affect participation.

Hypothetical Example

Let there be 100 cases and 200 controls who are asked to participate in a hypothetical study, where the variables of interest are gender (male or female) and age (under 50 years, or 50 years and over). The staged tree is shown in Figure 6.8 and the corresponding CEG is given in Figure 6.9.

Figure 6.9 shows cases are more likely to participate (80%) than controls (37%) regardless of gender or age. There are gender differences in the control group, with females (44%) participating more than males (30%), and age group differences with older males ($\frac{20}{50}$=40%) participating more than younger males (20%). Older male controls have a similar probability of participating as female controls of any age (43%). This information incorporates the disease status with other individual characteristics to summarise the participants. The findings can be used to explore differences between the disease groups for the consideration of methods to reduce participation bias.

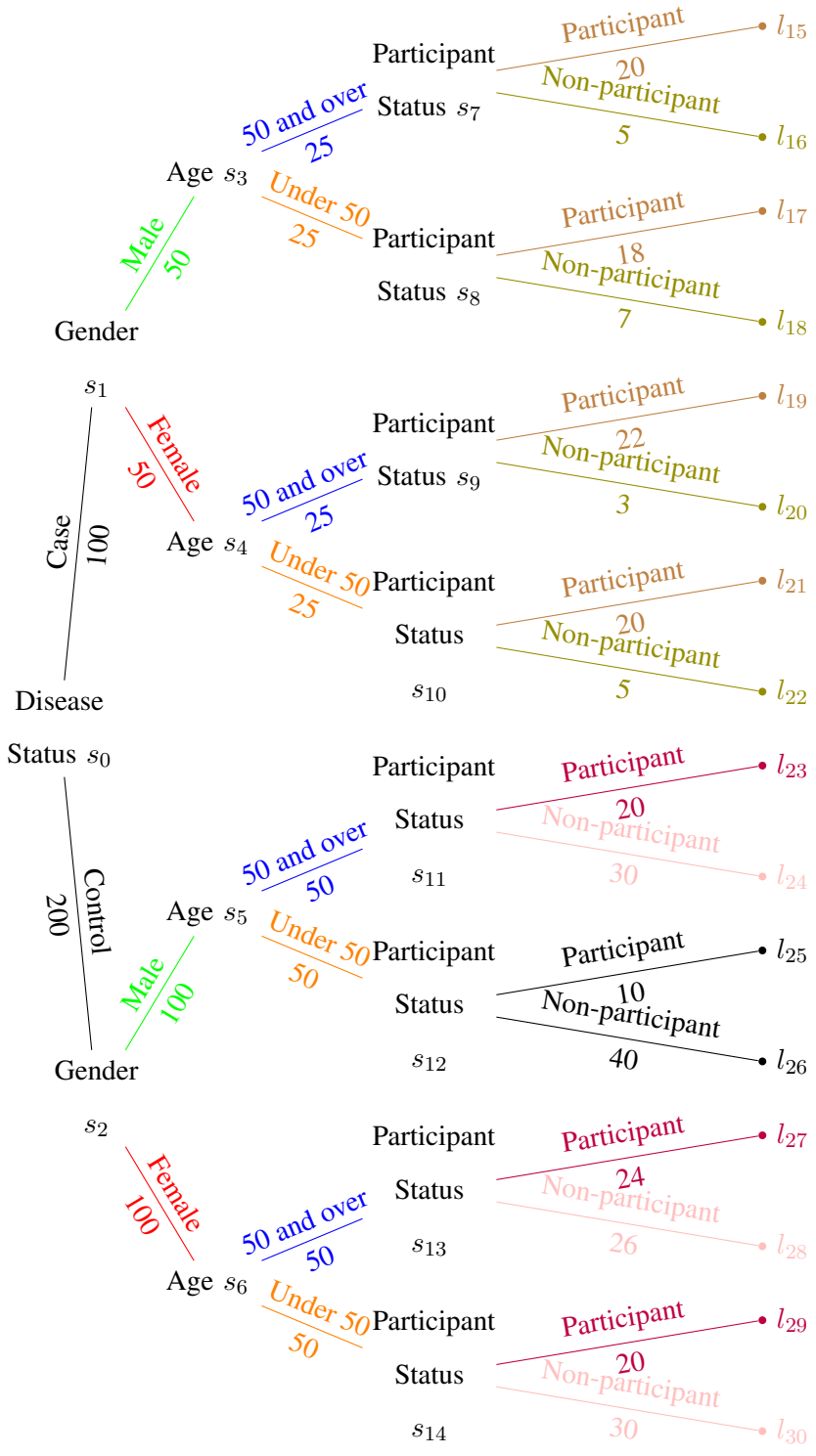

Figure 6.8: Tree by disease group. s denotes a situation and l denotes a leaf.

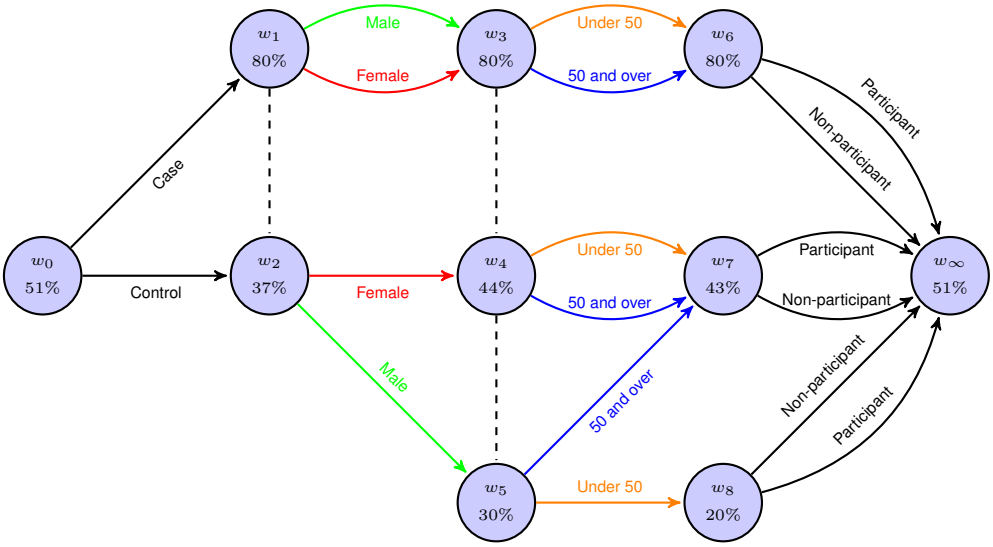

Figure 6.9: Chain event graph for participation by disease group. Percentage of participating individuals with given characteristics are shown at each position.

6.2.3 Amalgamated Case-Control Participation Data

If there are a series of case-control studies with similar variables recorded, the data from each study can be combined into one larger analysis using CEGs, where the outcome would be the disease status. Alternatively, the characteristics of cases or controls who are more likely to participate in a study of a particular topic may be of interest, or the most successful recruitment techniques may be sought. These findings can be achieved by having participation as the outcome variable, rather than disease status, as was demonstrated in Figures 6.6 and 6.9.

CEGs can be used in the same way as in §6.2.1 but the data are fed in from several studies. This may be particularly useful for studies of sensitive topics or those investigating very rare diseases, where the number of participants may be smaller. Conclusions can be drawn about the combination of factors associated with participation as demonstrated in §6.2.1. Patterns in the data can be used to form hypotheses regarding ways in which to increase participation in under-represented categories, or to inform future studies, although increased participation would not necessarily result in reduced bias. As CEGs can incorporate missing data,[365] studies which record similar but not identical variables can be combined directly, with unrecorded variables included as an additional edge labelled ‘unrecorded’.

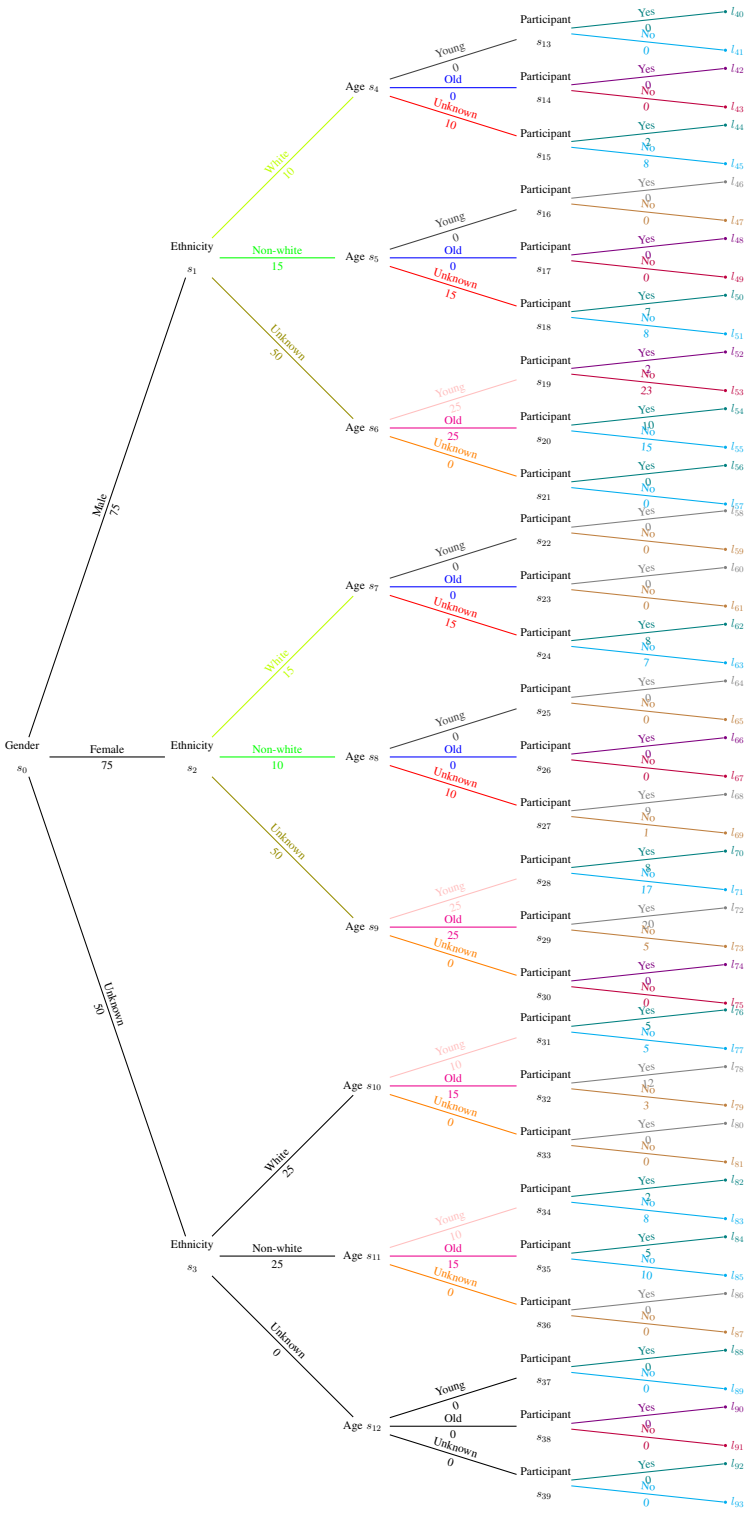

Figure 6.10: Staged tree formed from amalgamated data. s denotes a situation and l denotes a leaf.

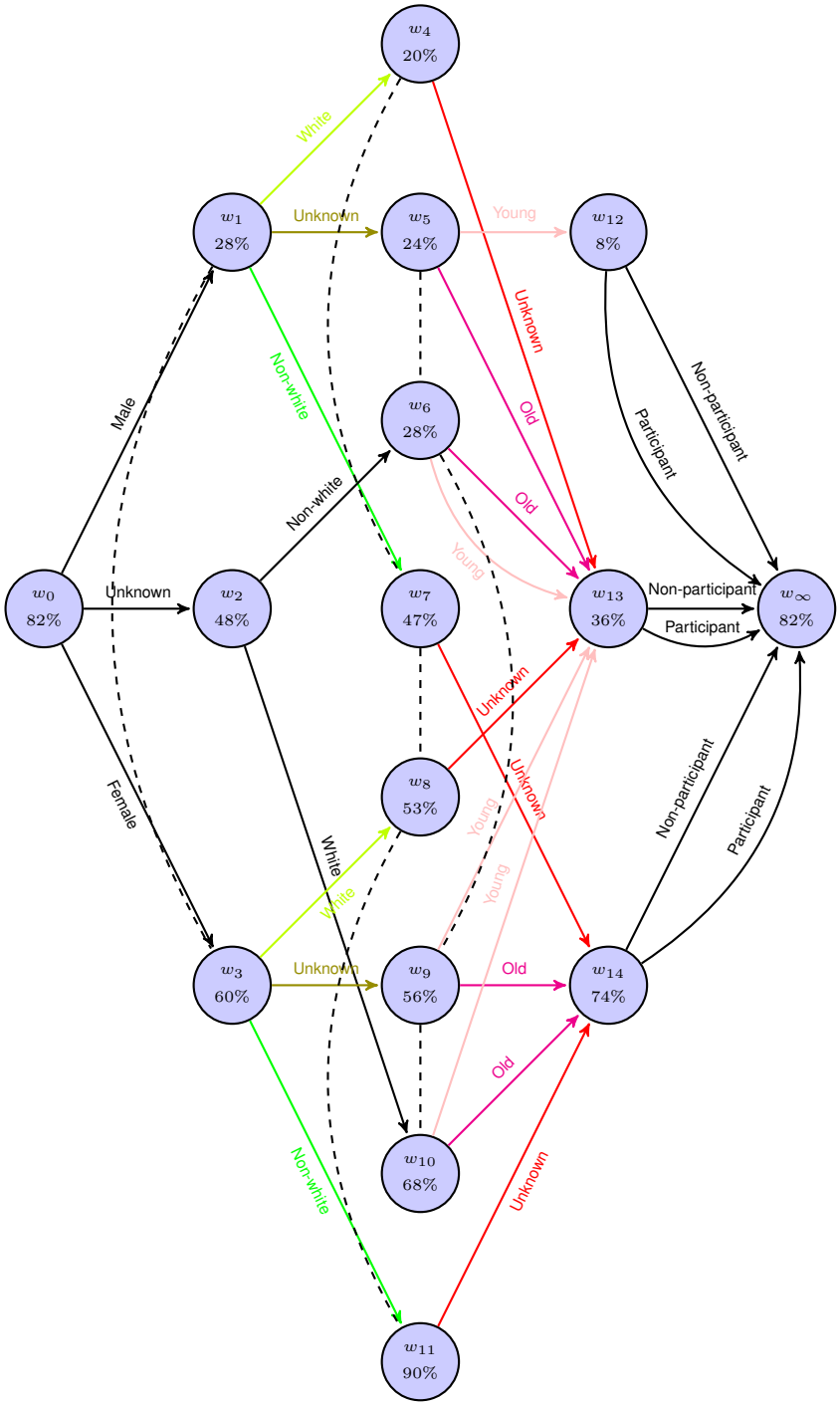

Figure 6.11: Chain event graph for the amalgamated data. Percentage of participants shown at each position.

This approach could also be used with the disease status as the outcome of interest, and with missing data as a result of non-participation or otherwise, included as extra edges. If desired, data missing from non-participation and data missing for other reasons, could be given separate edges in the event tree, such that differences between the types of missingness can be represented.

Hypothetical Example

Let there be three hypothetical studies; one which recorded age and gender, another which recorded age and ethnicity, and a third which recorded ethnicity and gender. These data could be used to investigate the general characteristics of those more likely to participate in a case-control study by having the final vertices showing the participation status of the individuals.

Assume these variables are non-sensitive and hence were available to the researchers within a given study whether the individuals chose to participate or not. Sensitive data would only have been recorded for participants but could be investigated by including 'missing' edges for non-participants. Depending upon the purpose of investigating participation, the tree could be constructed for the entire study group, just controls, or separate trees could be constructed for cases and controls for comparison.

Let the event tree be as in Figure 6.10, with the corresponding CEG as in Figure 6.11. For each of the studies, there are two variables recorded and one variable considered to be missing.

The CEG shows that males are less likely to participate than females, and the unknown gender category may be MAR as it is positioned between males and females (see §5.3 for further details on missingness in CEGs). If the data are MAR with respect to this larger sample, this could suggest that methods such as multiple imputation could be adopted[292] if participation bias is a concern.

The distribution of ethnicity is shown to be indistinguishable given known gender, since the same green colours are assigned to edges emanating situations s_1 and s_2 in Figure 6.10, hence ethnicity is distributed similarly amongst males as it is females, as would be expected in the population. When gender is known, white participants are more likely to participate than non-white, since the edges representing white participants lead to w_4 and w_8 which are positioned higher in the ordinal CEG in Figure 6.11 than positions w_7 and w_{10}, which the edges representing non-white participants lead to. The unknown ethnicity edges lie between the two known ethnic groups and

hence it is possible that the unknown category consists of both white and non-white individuals, suggesting that the unknown ethnicity values may be MAR.

The distribution of age is indistinguishable given known gender and ethnicity, and it is also indistinguishable given unknown gender or ethnicity, as indicated by the corresponding colours and dashed lines between positions $\{w_4, w_7, w_8, w_{11}\}$ and $\{w_5, w_6, w_9, w_{10}\}$ in Figure 6.11 respectively. If gender is unknown, non-white individuals are more likely to participate than white, since the non-white edge leads to position w_6, which is positioned higher than the white edge which leads to w_{10}.

The lowest probability of participation in the ordinal CEG ($w_{12} = 8\%$) can be reached only by one path; young males with unknown ethnicity. The highest probability of participation in the CEG ($w_{11} = 90\%$) is reached only by non-white females. Those with older age are generally positioned lower in the ordinal CEG than those with younger age, suggesting older individuals are more likely to participate. Overall, age and gender appear to be associated with participation, with females and older individuals more likely to participate. There is no such association for ethnicity.

Initially there were three studies, one of which showed older females were more likely to participate, the second showed older white individuals were more likely to participate and the third showed non-white females were most likely to participate. Combining these studies into one overarching study allows for a larger sample size, since there are more participants, and a more generalisable conclusion, since these studies may have been located in different areas and with different research questions, and been affected by non-participation in different ways. Of course if one of these studies already covers the research question and location of interest, it would be preferable to focus on that particular study. This approach of combining data may be useful in case-control studies which collect information regarding rare diseases and where participation rates have declined in recent years, to increase the overall sample size.

6.3 Analysis Adaptations

6.3.1 Data Reliability

CEGs have been used previously with prospective data but rarely with retrospective data, which can have limitations such as recall bias. This feature of the case-control study data can be incorporated into the CEG framework to enhance the authenticity of the analysis and to include additional information about the reliability of the data obtained. This approach may also be applied when data are missing and the reliability of the recorded data are altered.

Data in retrospective studies may be recorded using a variety of means such as medical records, through interview or using national databases. Some sources may be cross-checked and verified with other authorities, while other sources may depend upon a single handwritten report, such as in older medical records or in areas without electronic databases. In some instances the only source will be the memory of those present and will require the individual to recall specific details. Recall bias is known to differ between participants of different disease groups in case-control studies[390] and hence data reliability may differ by both source and disease status.

One way in which to allow for potentially less reliable study data is to form a CEG which has a greater dependency on prior knowledge, provided these data are collected from a more reliable source. Since CEG learning is Bayesian and combines prior knowledge with data, non-uniform priors can be specified during the AHC algorithm phase to achieve this, as discussed in §5.2.2.1. The equivalent sample size[361] is also specified and if a large equivalent sample size is used this suggests stronger prior beliefs and hence allows the priors to play a more dominant role, rather than depending strongly on the data.

The resulting tree will be structurally identical, but labelled with different priors (or left unlabelled). The CEG may differ according to whether uniform or non-uniform priors are used, and this will depend upon the priors assigned and the data collected. This approach could be particularly useful in studies which suffer from non-participation, since the true population distributions of variables may not be apparent from the study data and this could potentially affect the conclusions generated.

Hypothetical Example

Let there be a hypothetical study with two binary exposures, one of which is gender, plus a binary disease. First the analysis will be conducted using uniform priors and then with non-uniform priors constructed using hypothetical expert knowledge. Figure 6.12 shows the staged tree formed with uniform priors and Figure 6.13 shows the corresponding CEG. Figure 6.13 shows that males are more associated with case status than females (80% compared with 27%) and that for males, being exposed or not leads to the same position in the CEG (w_3) with 80% of the individuals being cases. However females differ by exposure, with 13% of exposed females (w_4) being cases and 80% of non-exposed females being cases (w_3).

Figure 6.14 shows the staged tree which uses the same data, but where non-uniform priors are allocated. The priors are the smallest possible such that each value is integer as discussed in §5.2.2.2 and these priors are shown along the edges of the tree in Figure 6.14. The priors have been assigned using the hypothetical prior knowledge of half males and half females in the population, with the exposure being as common amongst males as it is females, and with around 20% of the population being exposed. The ratio of cases to controls in the study has been maintained. The same data as in Figure 6.12 but with the priors in Figure 6.14, results in the CEG in Figure 6.15.

Figure 6.15 differs from Figure 6.13 in that position w_3 in Figure 6.13 is split into two positions (w_3 and w_4) in Figure 6.15, hence priors can affect the CEG produced. This split separates unexposed males, from exposed males and unexposed females, but returns two positions (w_3 and w_4) with the same proportion of cases (80%). Otherwise the CEGs are comparable and similar conclusions can be drawn. The effects of changes in prior information were seen more clearly in §5.4.3 where the sensitivity of the results from the diabetes data to changes in priors were tested.

In this example, the vertical ordering in the CEG is unchanged since the original position (Figure 6.13, w_3) and the two new positions (Figure 6.15, w_3 and w_4) each have 80% of individuals who are cases. However it is possible that in some instances the splitting of positions could result in the reordering of the variables associated with these positions, especially if the percentage of (in this example) cases is similar amongst the positions. The equivalent sample size corresponds to the strength of the prior beliefs and could also affect the CEG, hence it is advised to check the robustness of the CEG with respect to changes in the equivalent sample size.[8]

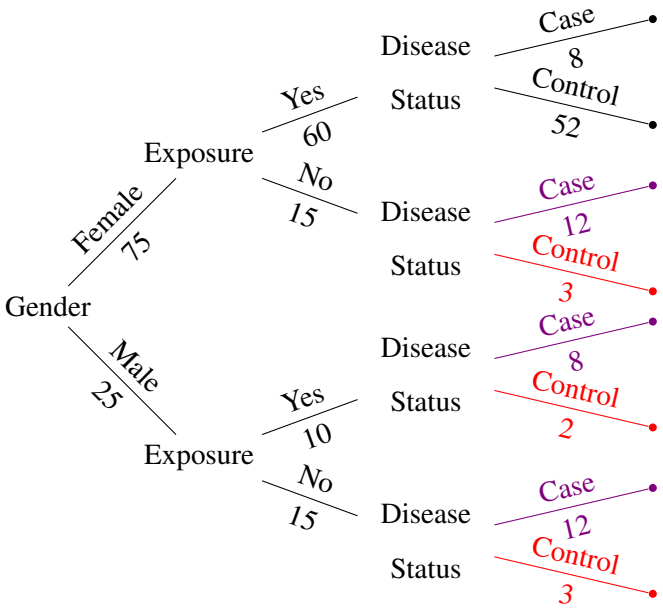

Figure 6.12: Data reliability: Staged tree with uniform priors.

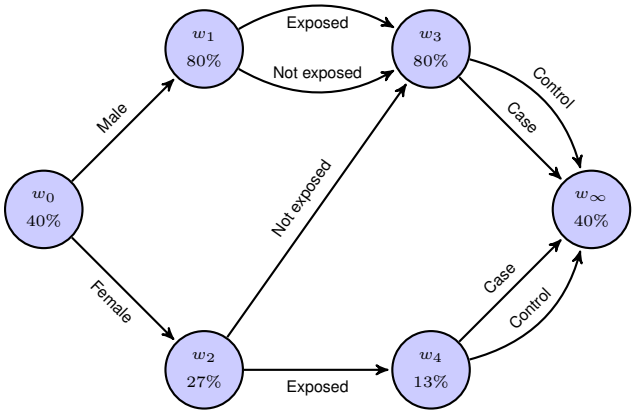

Figure 6.13: Data reliability: Chain event graph formed from uniform priors.

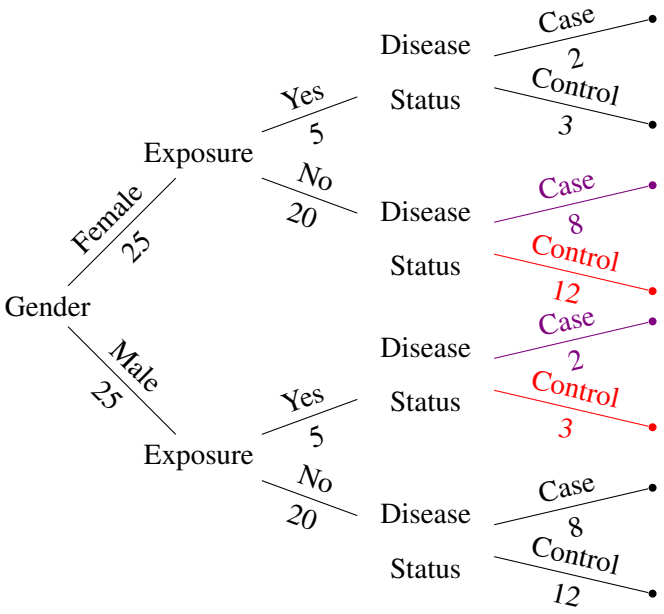

Figure 6.14: Data reliability: Staged tree with non-uniform priors. The numbers indicate priors rather than individuals; the number of individuals are shown in Figure 6.12.

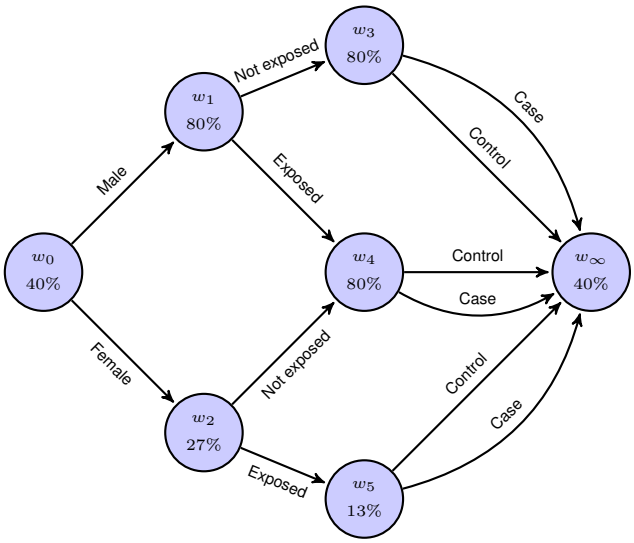

Figure 6.15: Data reliability: Chain event graph formed from non-uniform priors.

6.3.2 Subset-Chain Event Graphs

CEGs are usually simple for a small number of variables, but quickly become complicated and more difficult to read when there are a large number of variables and/or each variable has a large number of categories. This includes where there are a large number of variables with missing data, where each variable includes an additional edge denoting missingness. Here, for these instances which can occur in case-control studies, subset-chain event graphs (subset-CEGs) are proposed as a new variant of CEGs.

A subset-CEG is simply a subset of variables displayed in a CEG which relate to a particular aspect of the data, which can later be interpreted alongside other subset-CEGs. One such CEG could be constructed for individual characteristics and another for environmental factors, with the number of subset-CEGs dictated by the number of variables and categories. If desired, one final CEG can be constructed at the end of the analysis which contains all variables found to be important in the subset-CEGs.

Hypothetical Example

Let there be a study where a total of 150 male and female individuals, who can be classed as either old or young by a given cut-off age, are asked to participate in a study by web or mail, with some receiving a reminder to participate and others not. The characteristics of the individuals are given in the staged tree in Figure 6.16 and the corresponding CEG is shown in Figure 6.17. The recruitment details for the study are as were shown in Figures 6.5 and 6.6.

Figure 6.17 shows that participation does not differ between males and females, but is more likely from those who are old than those who are young. Figure 6.6 showed reminders to be associated with participation, but not the survey delivery mode and, if desired, these variables can be used to construct a final CEG for the dataset. Figure 6.18 shows the staged tree for the variables of age and reminders on participation, and Figure 6.19 shows the corresponding CEG. The CEG suggests that old individuals are more likely to participate than young, and that reminders are associated with increased participation in old individuals, but are not as effective for young individuals.

Since the natural ordering of age and reminders is unclear, the analysis was also rerun with the reminder variable before age. This resulted in the CEG shown in Figure 6.20 which shows

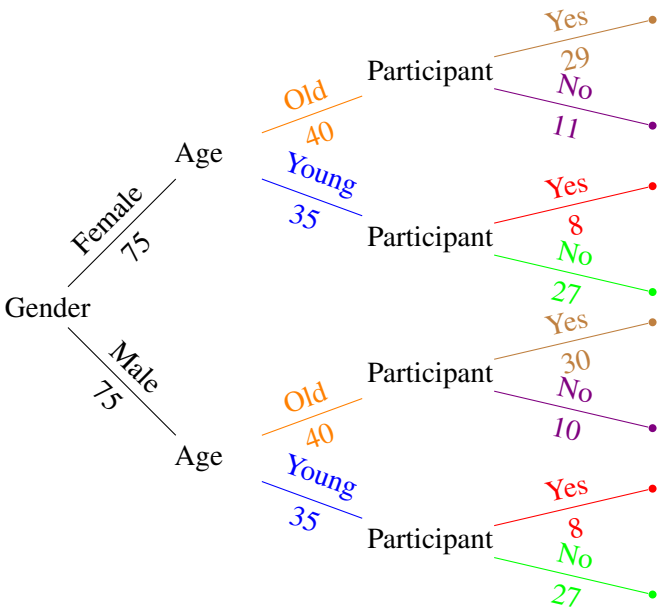

Figure 6.16: A staged tree used to form a subset-chain event graph.

reminders to be associated with increased participation and old individuals to be more likely to participate than young individuals. Again reminders are not as effective for young individuals as they are old individuals. Therefore the same conclusions are drawn, regardless of the ordering of the age and reminder variables. This should be expected, since age is not affected by reminders and the allocation of reminders is not determined by age, hence the ordering of these two variables is less important than in other scenarios.

Here, subset-CEGs have been used to simplify the analysis into smaller steps and use variables thought to be associated with the outcome to form a final CEG. Each subset-CEG shows a different aspect of the study, which might have been missed in a full CEG. This also improves the readability of the CEG. Another approach to improve readability, may be to present the CEGs against a grid representing the percentage of individuals at each vertex who have the outcome of interest (participation or disease status here). Rather than display the percentages within the vertices, here it is proposed that the vertices could be placed vertically against the grid to show their relative positioning. Figure 6.20 has been redrawn using a grid and is shown in Figure 6.21 as an example. This improves readability for spatial readers, but may cause the edges to be less clear and result in fewer planar graphs and hence a graph which is more difficult to interpret.

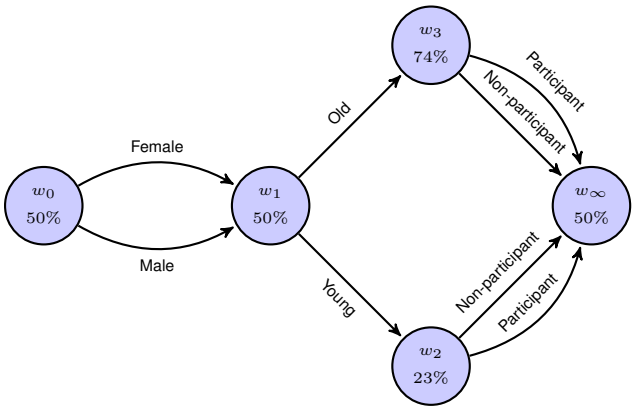

Figure 6.17: An example of a subset-chain event graph. Percentage of participating individuals shown at each position. Colouring is not required since stages and positions are equal.

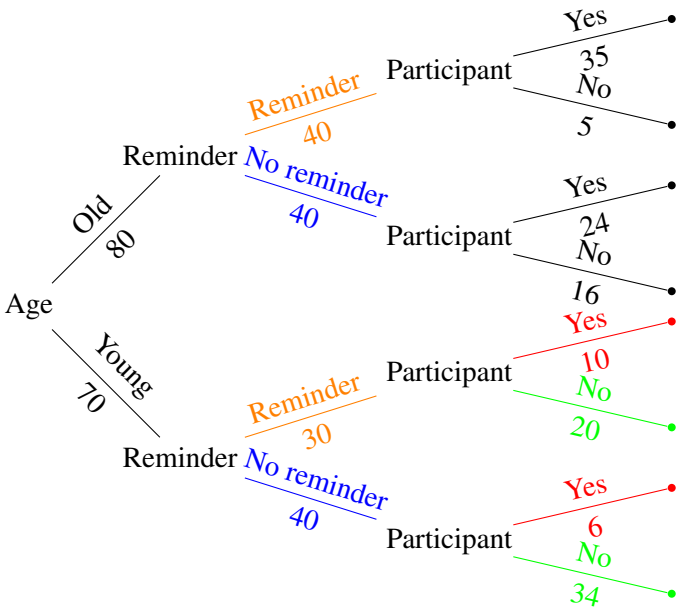

Figure 6.18: A staged tree with variables selected using subset-chain event graphs.

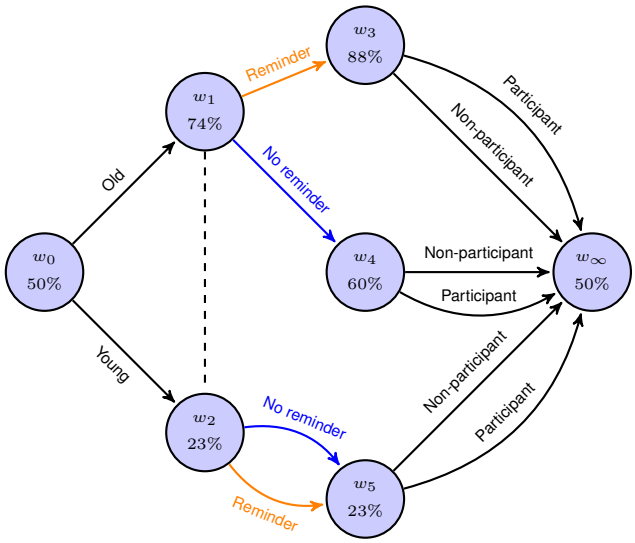

Figure 6.19: A final CEG with variables selected using subset-chain event graphs. Percentage of participating individuals shown at each position.

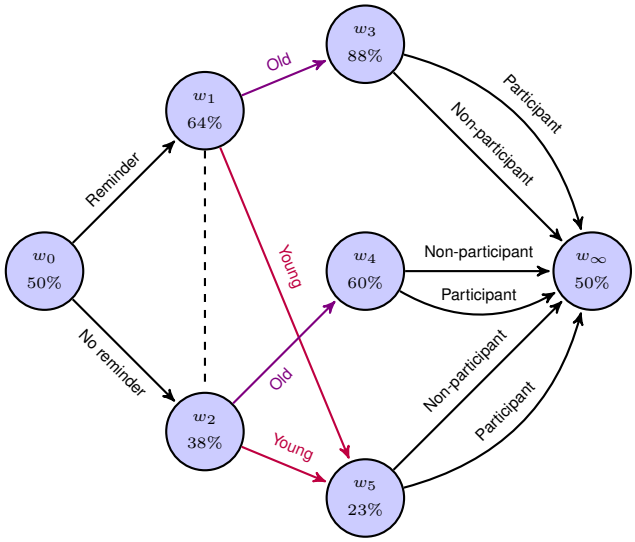

Figure 6.20: A final CEG with variables selected using subset-chain event graphs, age and reminder variables swapped. Percentage of participating individuals shown at each position. Example colouring has been used to highlight which positions were in the same stage, as the staged tree is not shown.

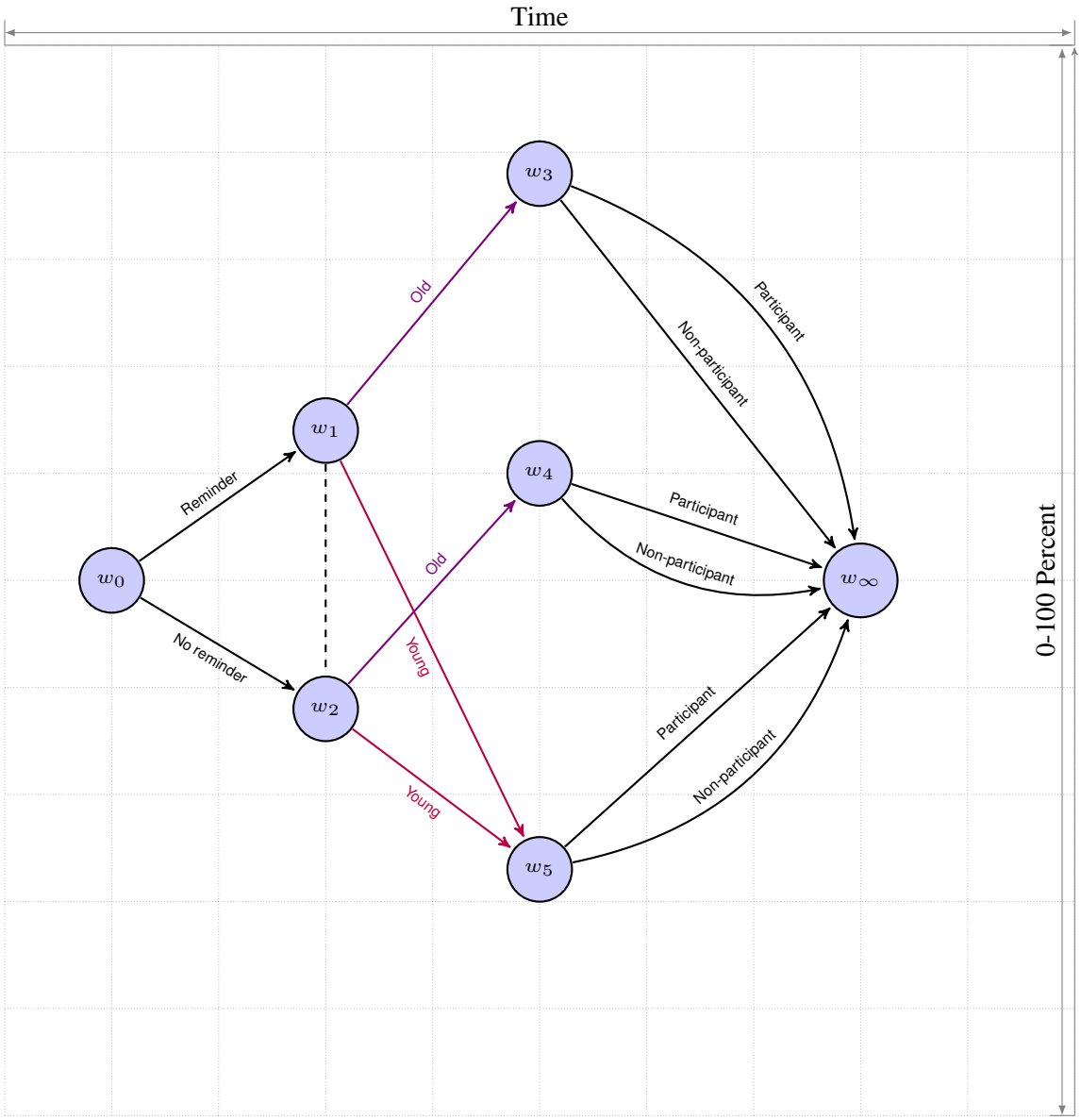

Figure 6.21: Example of a grid to position vertices vertically with respect to their percentage in an ordinal CEG. Each vertical line in the grid represents 10%.

6.4 Summary of the Adaptations to Chain Event Graphs for Case-Control Data

In this chapter, the CEG framework has been adapted to further explore missingness in case-control studies, in particular for missingness resulting from non-participation. Seven adaptations have been proposed, relating to the study design, participation, or analysis of a study.

Here the structure of the CEGs, the ordering of variables, or the outcome of interest have been changed, but the underlying methodology is the same as in previous publications.[21,365] CEGs have therefore been applied to a new epidemiological area and used for new purposes such as data collection and non-participation summaries, which are applicable to case-control studies.

6.4.1 Conclusions for Study Design Adaptations

CEGs have been suggested here to investigate the missingness resulting from different disease severities, which extends the usual binary outcome of case-control studies to more categories. This additional information may be useful for understanding where missingness occurs in a case-control study, and does not prevent the categories from being collapsed so that traditional (binary outcome) analyses can be conducted.

CEGs have also been used here to investigate the different recruitment techniques adopted within a case-control study. CEGs have not been used before to investigate data collection techniques (see §5.1.2), but these findings may highlight differences between the disease groups being compared, particularly if cases and controls engage with the study in different ways.

6.4.2 Conclusions for Participation Adaptations

CEGs have been used here with participation as the outcome of interest; an outcome which has not before been suggested (see §5.1.2). Information regarding the characteristics of those who do and do not participate, and comparisons between participating cases and controls, can be informative for determining whether participation bias is likely to have occurred in a study (see §2.3.4.1).

The basic characteristics of those who have declined to participate have been included in some

of the hypothetical examples given here. This approach could be viewed as unethical, since they have not consented for their data to be analysed. However, the data included in the CEGs could be publicly available data, and is similar to the level of detail which could be used to conduct a sensitivity analysis, a participation flowchart, or a comparison between the characteristics of participants and non-participants. Distinction may be required between observational studies which do not require participant consent, and scenarios where individuals have refused to participate. However, it could also be seen as unethical *not* to consider the non-participants and their characteristics, which can be used to assess the possibility of participation bias. Ethical concerns have been touched upon in the literature,[391] and ultimately referral to an appropriate ethical committee for guidance may be required.

6.4.3 Conclusions for Analysis Adaptations

Reduced ordinal CEGs have been introduced previously[365] but here subset-CEGs have been suggested to simplify the graphs and improve readability. Subset-CEGs can be used for any number of outcome categories and hence have an advantage over reduced ordinal CEGs which require the outcome to be binary. A grid-based background has also been suggested here for use with CEGs, such that the percentages do not need to be included in the vertices and instead the spatial layout of the CEG indicates these percentages.

6.4.4 Overview

6.4.4.1 Additional Suggestions

Sensitivity analyses can be performed for the CEGs given in this chapter, and would use the same methods as those demonstrated in §5.4.3. This includes changing the ordering of the variables in the event tree, using different priors, or altering the strength of the prior beliefs. Of course, there may be instances where reordering the variables is not sensible, such as where the tree includes initial requests to participate and then reminders. Testing the sensitivity of these CEGs relating to study design can conclude whether the findings regarding participation are robust or consistent.

6.4.4.2 Critical Evaluation

While the current CEG framework was suitable for straightforward case-control study data as shown in Chapter 5, adaptations to the CEG framework enabled analyses specific to the case-control study design. Several aspects from the study design can be incorporated into the analysis, such as their retrospective nature, information from previous studies or experts, implausible variable combinations and asymmetric problems. These features allow the analyst to throughly address the missingness resulting from non-participation while taking into account context-specific information. CEGs have not before been used to summarise non-participation, data collection techniques, data reliability or disease severity in case-control studies as shown in §5.1.2.

The lack of data recorded in the diabetes study for non-participants demonstrated the potential disadvantage of using these CEG adaptations with case-control data. Therefore, to be able to use the adaptations suggested here, the relevant data regarding non-participants, recruitment methods or case severity would need to be available.

Each hypothetical example given is relatively simple to demonstrate the new idea, but can of course be extended to include more variables and more variable categories as would be found in real data. However, these smaller graphs demonstrate how simple CEG structures can be while describing interesting associations between variables. As demonstrated here, it is possible to include prior knowledge in the analysis, rather than assuming each path is equally likely.

As discussed in §5.5, a limitation of CEGs is that they require the data to be categorical. While this may be problematic for continuous variables which do not have a sensible cut-off values (clinical in most instances), it may be less problematic here since many of the variables considered will be naturally categorical, such as the data collection method being face-to-face interviews or postal questionnaires. Where categories are defined, they may be more functional; for example dictated by cost per participant or by the speed of data collection.

6.4.5 Further Work

6.4.5.1 Agglomerative Hierarchical Algorithm Further Work

The strengths of the AHC algorithm are recognised[379] and some advantages were briefly discussed in §5.2.2.1. However, the algorithm also has a limitation. Currently learning CEGs assume that the variables are random and observed, even when they are deterministic in reality. If some variables are allocated, such as the number of reminders following a survey, then these variables are no longer random. The AHC algorithm runs with this random assumption and hence may be grouping vertices into the same stage which should not be grouped. In reality these variables are from different distributions and hence should be in different stages and positions. A constraint could be added to the algorithm to ensure particular vertices are not grouped to the same position.

Further work could also convert the AHC algorithm R code into an R package for use more universally and to encourage applications of CEGs in areas such as medicine and social sciences. Ideally, the package could also generate figures for the trees and CEGs, which would be particularly helpful for larger examples.

6.4.5.2 Further Chain Event Graph Adaptations

Further work could include simulations to demonstrate participation bias as a result of different recruitment approaches or participant characteristics, and could use the information obtained through CEGs to see where the bias occurs. Rather than use the CEGs to identify successful recruitment techniques to increase participation rates, CEGs could be used to identify where the bias is occurring and whether a subset of the sample would be preferable to reduce bias.

Further adaptations to CEGs for use with case-control studies could include the reason for missingness, such as whether the potential participants were uncontactable, unwilling, or unable to participate. Other adaptations could also include the stage at which the refusal reason was provided, such as after the first request or following a reminder.

Chapter 7

A Method to Reduce Participation Bias Using Population Data

This chapter proposes a new solution to reduce participation bias using population data, as a solution for when data are MNAR since limited options were found following §5.4.4.2 for the diabetes MNAR rhesus factor variable. In recent years, extensive population data have become more widely available; partially due to advances in technology, increased routine data collection and emphasis on data sharing, along with the recent move towards, and focus on, Big Data.[392] Linked data (which can be connected to different sources) such as hospital episode statistics (HES),[393] the clinical practice research database (CPRD)[394] and ResearchOne,[395] allow information to be shared more easily and for research to be conducted. Often these databases hold much more information, on a greater number of individuals, than could easily be collected through a study. Some census databases also contain information relating to every member in a population, such as in Denmark or Sweden.[7,59,60]

In this chapter population data are used in place of control data, and in conjunction with case data, in a case-control framework. A literature search is conducted in §7.1 which confirms that this method has not been developed previously. The method is explained in §7.2 and demonstrated using simulated data (where the true odds ratio is known) in §7.2.5 and §7.2.6. The method is applied to the diabetes data[7] in §7.3 and a study of stroke in Bijapur, India[396] in §7.4. An overview of the method is provided in §7.5, which includes a critical evaluation in §7.5.1, possible extensions in §7.5.2, suggestions for amendments in §7.5.3, how the method could be used as a sensitivity

analysis in §7.5.4, and comparisons with other methods in §7.5.5, with a final summary in §7.6.

7.1 Literature Search

A Web of Science[130] literature search was conducted on 16th March 2016 to try to uncover any existing use of population data with case data for reducing participation bias. The `topic` was searched using the terms `("case-control" OR "case control") AND "population data" AND "bias"`. The first term was to search for the study design of interest, the second term included the data source and the third term was to link the results to methods to reduce bias. The bias type was not specified since participation bias can take other names, such as selection bias or non-response bias. There were no further restrictions on the search, and nine results were returned; seven from the Web of Science core collection and two from Medline. These were:

1. A theoretical publication considering selection, recall and interviewer bias when case-control studies of cancer use individuals with other cancers as controls.[397] Population data are discussed, but not used as in this chapter.
2. An article about small area data and links to population data, but not using the method proposed in this chapter.[398]
3. The results from a tuberculosis case-control study[399] following standard analyses.
4. A method to reduce sampling bias using weighted logistic regression to adjust the ORs[400] (weighting was described in §4.4.1).
5. A study of the effects of missing data in genetic epidemiology when the assumption of MAR does not hold, and the application of a missing data model to characterise missing data patterns in two or more genetic markers.[401]
6. A review of publication bias and the association between smokeless tobacco and oral cancer.[402]
7. An article written in French, but the English abstract showed a discussion between case-crossover designs and time series designs, which did not refer to participation bias in relation to case-control studies.[403]
8. Self-selection bias and breast cancer screening in the Netherlands, which concluded the bias to be minor and which did not use population data to reduce bias.[404]

9. Self-reported survey data in population-based studies of hormones and breast cancer, where population data were used to validate exposure prevalences[405] but not to reduce bias.

The brief summaries of the nine articles show that none are similar to the approach used here and hence the proposed method is assumed to be original. However, a secondary search was conducted, again on 16th March 2016 and using Web of Science, to try to ensure no relevant articles had been missed. This time the term `bias` was used in the `topic` search, along with `"case-population"` which would be a sensible description for when control data in a case-control study are replaced with population data. Seven results were returned as shown below. Again, none used population data to reduce participation bias in case-control studies.

1. Application of the capture-recapture method to historical epidemiology.[406]
2. A pregnancy and diazepam case-control study, investigating the possible human teratogenicity of diazepam during pregnancy,[407] which used standard analyses.
3. A theoretical paper for spatial clustering, not related to participation bias in case-control studies.[408]
4. A paleopathology review article which does not consider case-control studies nor participation bias.[409]
5. A simulation study for the effect of survival bias on case-control genetic association studies of highly lethal diseases,[410] which is unrelated to population data or participation bias.
6. An article about reconstructing populations, which is unrelated to case-control studies or participation bias.[411]
7. A simulation study to evaluate stratified random sampling of screening mammograms,[412] which did not consider participation bias in case-control studies.

Finally, a third search was conducted on 29th April 2016 which replicated the first search but included `"census data"` as a synonym for population data, and five additional results were returned. These five articles used census data to; compare participants with non-participants[413] (without adjustment), re-weight case and control groups,[414] collect additional data on occupational groups as the exposure of interest,[415] reconstruct historical data[416] and select controls.[417] However none of these additional five articles used census or population data as part of a new method to reduce bias as will be described in this chapter. This new method was also not encountered when the literature was searched to gather information for Chapter 2.

7.2 The Method

7.2.1 Method Development

The development of the method was through a thought experiment; the outline of which follows. In case-control studies, the case group is usually relatively unbiased, since cases are generally willing to participate. Participation bias is therefore more likely to be caused through the control group (see Chapter 2). It seems sensible to try to replace the possibly biased control data with unbiased, or less biased data, while utilising the existing, less biased case data. A possible source may be population lists, such as census data, which aim to measure the entire population and therefore the data should not be biased in same the way as study controls.

Population data contain information on both cases and controls, but only the control data are required since case data are available from the study. The study case data could be used to identify which information from the population data relates to cases, by scaling the case study data up to the population level. The data relating to the cases can then be subtracted from the population values, leaving solely the data relating to controls. Although, in many instances the control and population data will be similar since the case data form only a small proportion of the population in rare diseases. The remaining data can then be used in place of the control data from a case-control study, hence replacing the possibly biased control data with often widely available, less biased 'control' data from the population, for use in the analysis of the study.

7.2.2 Required Population Data

For any particular case-control study, a range of exposures may be investigated. There are three values required from the population for each exposure considered, which must be correct for the time and location of the original study:

1. The number of exposed individuals in the population;
2. The size of the population;
3. The number of cases in the population.

Various sources can be used to obtain these data; some publicly accessible. There follows two examples in §7.3 and §7.4 which use publicly accessible information to demonstrate the ease of

obtaining these required data, but more recent or detailed data could be obtained from other studies or databases if available, and which would likely improve the accuracy of the results.

7.2.3 Implementing the Method

The steps required to use population data in place of control data are;

1. Subtract the (scaled if necessary) case numbers from the population numbers to calculate the control numbers.
2. Use the exposed population and exposed case data to calculate the number of exposed controls.
3. Use the previous steps to calculate the remaining number of unexposed population individuals, cases and controls.
4. Use these values to calculate odds ratios from a contingency table or using logistic regression as shown in §2.2.5.

These steps are implemented in §7.3 and §7.4 with two different datasets for illustration, and the findings compared with the published results.

7.2.4 Mathematical Notation

This method can also be written mathematically; let P be the number of individuals in the population of interest, D be the binary disease of interest, E be the binary exposure of interest, a be the number of exposed study cases and c be the number of unexposed study cases. The presence of a binary variable is indicated using a value of one, and its absence is indicated using a zero.

Values from the population can then be substituted into the following equations. All stages are shown, with the necessary steps in bold.

$$P = P_{D=1} + P_{D=0}$$

$$\mathbf{P_{D=0}} = \mathbf{P} - \mathbf{P_{D=1}}$$

$$P_{E=1} = P_{D=1,E=1} + P_{D=0,E=1}$$

$$P_{D=1,E=1} = \frac{a}{(a+c)} \times P_{D=1}$$

$$\mathbf{P_{D=0,E=1}} = \mathbf{P_{E=1}} - \mathbf{P_{D=1,E=1}}$$

$$P_{E=0} = P_{D=1,E=0} + P_{D=0,E=0}$$

$$\mathbf{P_{E=0}} = \mathbf{P} - \mathbf{P_{E=1}}$$

$$\mathbf{P_{D=1,E=0}} = \mathbf{P_{D=1}} - \mathbf{P_{D=1,E=1}}$$

$$\mathbf{P_{D=0,E=0}} = \mathbf{P_{D=0}} - \mathbf{P_{D=0,E=1}}$$

7.2.5 Simulated Example: Data Missing at Random

Definitions for MCAR, MAR and MNAR were given in §2.3.7, and here simulations are used to demonstrate this method with data which are MAR. Let there be a hypothetical dataset of 120 cases and 240 controls taken from a population of 1000 individuals. Let there be the exposure, an auxiliary variable, the disease status, a correlation between the exposure and auxiliary variable, and let the auxiliary variable affect participation. The DAG for this setup is shown in Figure 7.1 and participation bias occurs due to conditioning on participation into the study.

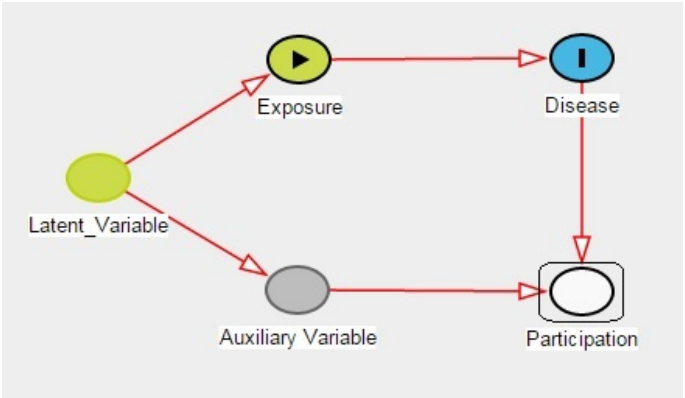

Figure 7.1: A directed acyclic graph showing the variables in the study. The latent variable allows the graph to represent the correlation between the exposure and auxiliary variable.

The population was simulated according to the following constraints. Each member of the population was classed as exposed/unexposed, did/did not possess the auxiliary variable, was/was not a participant and as a case/control; an entirely binary setup for simplicity. For reduced complexity and to demonstrate an extreme example of participation bias, participation occurred only in cases, or in controls who possessed the auxiliary variable. Although highly unlikely, this may arise in controls if, for example, a particular group chooses not to participate on religious grounds, leaving an entire category unrepresented. Also, if the method can recover the true odds ratio for extreme examples, it should also be applicable for more realistic scenarios. Finally, the exposure and auxiliary variable were strongly negatively correlated, those exposed were more likely to have the disease of interest, and there was no confounding. These factors resulted in the population shown in Tables 7.1 and 7.2.

	Not Exposed	Exposed
Not Diseased	685	166
Diseased	31	118

Table 7.1: Simulated binary population: Exposure and disease status.

	Not Exposed	Exposed
Auxiliary Variable (No)	22	262
Auxiliary Variable (Yes)	694	22

Table 7.2: Simulated binary population: Exposure and auxiliary variable.

The resulting correlation between the exposure and auxiliary variable was ρ = -0.89 (2dp) and for reference the correlation between the outcome and exposure was 0.47 (2dp). Summaries were generated from the population, which confirmed the setup was as intended (results not shown). Recall that participation occurred only in cases or if the control possessed the auxiliary variable, and that 120 cases and 240 controls were selected. Therefore there were sample cases, who are more likely to be exposed and less likely to have the auxiliary variable. There were also controls, all of whom had the auxiliary variable and who were less likely to be exposed. The resulting sample group is shown in Tables 7.3 and 7.4, with the correlation between the exposure and auxiliary variable now ρ = -0.92 and the correlation between the outcome and exposure now 0.84. The population and sample odds ratios (after sampling with participation bias) were calculated;

Table 7.5 shows the sample odds ratio is biased.

	Not Exposed	Exposed
Not Diseased	238	2
Diseased	24	96

Table 7.3: Simulated binary sample: Exposure and disease status.

	Not Exposed	Exposed
Auxiliary Variable (No)	1	87
Auxiliary Variable (Yes)	261	11

Table 7.4: Simulated binary sample: Exposure and auxiliary variable.

	Odds Ratio (95% Confidence Interval) (2dp)
Population Odds Ratios	15.71 (10.21, 24.16)
Estimated Sample Odds Ratios	476.00 (110.34, 2053.39)

Table 7.5: Simulated binary odds ratios with 95% confidence intervals.

The proposed method can be applied after gathering the required population data;

1. The exposure in the population was $\frac{(166+118)}{1000} = \frac{284}{1000} = \frac{71}{250} = 0.284$.
2. The size of the population was 1000.
3. The number of cases in the population was 149.

The resulting contingency table after applying the method is shown in Table 7.6, where the odds ratio was $\frac{(685\times118)}{(31\times166)} = \frac{80830}{5146} = 15.71(2\text{dp})$, agreeing with the population exposure odds ratio in Table 7.5. The steps used to achieve the values contained within Table 7.6 are given below.

$$\begin{aligned} \text{population} &= \text{cases + controls} \\ 1,000 &= 149 + \text{controls} \\ 1,000 &= 149 + 851 \end{aligned}$$

$$\begin{aligned} \text{exposed population} &= \text{exposed cases + exposed controls} \\ 284 &= 118 + \text{exposed controls} \\ 284 &= 118 + 166 \end{aligned}$$

$$\begin{aligned} \text{not exposed population} &= \text{not exposed cases + not exposed controls} \\ (1,000 - 284) &= (149 - 118) + (851 - 166) \\ 716 &= 31 + 685 \end{aligned}$$

	Not Exposed	Exposed
Not Diseased	685	166
Diseased	31	118

Table 7.6: Contingency table formed for the simulation example, using population data.

Therefore in this simulated population, the proposed method has been shown to recover the true odds ratio when using population data. This is to be expected, since the population values used here are known exactly, solely to demonstrate the validity of the method. Of course this assumes that accurate population values are available and case data are unbiased, which may not be true when using real data. The 95% confidence interval for this estimate can be calculated using logistic regression in R,[303] while confirming the point estimate calculated by hand (see §2.2.5). The interval was found to be (10.21, 24.16) (2dp), which is relatively wide due to the small population size of just 1000 members. Table 7.7 summarises these findings.

Further simulation studies could investigate how different correlations result in different amounts of bias, or how different from the truth the population data can be while still recovering a suitably accurate odds ratio estimate. A sensitivity analysis (see §4.1 for the method) could be conducted

Exposure of interest	Population (true) OR (95% CI)	Sample OR (95% CI)	Population data OR
E	15.71 (10.21, 24.16)	476.00 (110.34, 2053.39)	15.71 (10.21, 24.16)

Table 7.7: Simulations: Odds ratios and 95% confidence intervals (2dp) comparing the true and sample odds ratios, with those generated using population data.

to investigate this effect on the odds ratio of changes in the population data estimates (a related discussion is given in §7.5.3). Since case-control studies usually study rare diseases, the OR estimates may be sensitive to small changes in the population values due to the relatively small number of exposed and unexposed cases compared with controls. However, the purpose of this simulation is to demonstrate that the method can recover the true odds ratio, since the true odds ratio is unknown for the real data examples which will be shown in §7.3 and §7.4.

7.2.6 Example: Data Missing Not at Random

Chapter 4 showed that many of the methods to reduce participation bias require the data to be MAR. This section shows that the true odds ratio can be recovered using the population data method even when the data are MNAR. Using hypothetical data collected in a blinded survey, let the auxiliary variable be age (with two general categories; older/younger), with older controls being more likely to participate in the study and with all participants happy to record their age category. Let the exposure and auxiliary variable be positively correlated, so those who are older are more likely to participate and are more likely to have the exposure of interest. Also let participants be selected based upon their disease status and let cases be more likely to participate in the study than controls. Let the exposure be sensitive, such that those who are exposed are less likely to respond to the survey question regarding exposure. A population of 1000 adults was created, as in Table 7.8, where the true odds ratio is 2.29 (1.22, 4.29) (2dp), hence the exposure is a significant risk factor for the disease.

Let the population be sampled for the study and assume all 50 cases participate, since participation rates for cases tend to be high. Let there be a 1:1 ratio between cases and controls, hence 50 consenting controls are required. The initial 50 controls are randomly selected from the controls in the population, with some participating and others declining, partly as they are controls who

	Not Exposed	Exposed
Not Diseased	800	150
Diseased	35	15

Table 7.8: Missing not at random example: The true population values.

are less interested in the study results, and partly because they are younger and less willing to participate. Sampling continues until 50 controls are recruited. The resulting 50 controls who agree to participate in the blinded study are therefore generally older and more likely to have the exposure of interest.

Recall that the exposure was sensitive, so those who are exposed may be less willing to reveal this information. If the exposure question is hidden amongst other questions, the exposure of interest may not be obvious to the participant. The control may willingly answer questions such as whether they are 'older' or 'younger' based on a given cut-off age, but they may not wish to record their exposure details. Once the data are collected, it may be that some controls cannot be used as they have not answered the question(s) relating to the exposure of interest.

Let further controls be recruited, to ensure there is the required 1:1 ratio between cases and controls. To recruit the full 50 controls needed, it may be that those willing to answer the necessary exposure question, are unexposed. This may result in a control sample which is showing a higher proportion of unexposed individuals than in the control population. Table 7.9 shows a possible sample, which has an odds ratio of 10.29 (2.21, 47.90) (2dp), which still shows the exposure to be a significant risk factor, but with a higher odds ratio estimate. Note there are 48 and 2 unexposed and exposed controls in the sample respectively. Had the sample been representative of the population, this would have been (scaled and rounded to the nearest individual) 42 and 8 for unexposed and exposed controls respectively.

	Not Exposed	Exposed
Not Diseased	48	2
Diseased	35	15

Table 7.9: Missing not at random example: The biased sample values.

Now let the population method be applied. There is a sample with participation bias and exposure

values which were missing due the value of the exposure, hence that data are MNAR. The population data required are;

- Population size; 1000 individuals.
- Exposure in the population; $\frac{165}{1000}$.
- Cases in the population; 50.

The same calculations as in §7.2.3 can be applied as shown below, which result in the same values as Table 7.8, and hence the true population odds ratio is recovered, despite the data being MNAR.

$$\begin{aligned} \text{population} &= \text{cases + controls} \\ 1,000 &= 50 + \text{controls} \\ 1,000 &= 50 + 950 \end{aligned}$$

$$\begin{aligned} \text{exposed population} &= \text{exposed cases + exposed controls} \\ 165 &= 15 + \text{exposed controls} \\ 165 &= 15 + 150 \end{aligned}$$

$$\begin{aligned} \text{not exposed population} &= \text{not exposed cases + not exposed controls} \\ (1,000 - 165) &= (50 - 15) + (950 - 150) \\ 835 &= 35 + 800 \end{aligned}$$

7.3 Example: Type I Diabetes Case-Control Study

This dataset has been used in previous chapters and further details can be found in Appendix A.

7.3.1 Incorporating the Population Data

Population data from Yorkshire which were collected at the same time as the study took place (1990s) can be used, as given in Table 7.10. The caesarean calculation used time-specific data and the resulting contingency table can be found in Table 7.11. Amniocentesis data from the required time-point were unavailable, so more recent British (rather than Yorkshire) data were used, and scaled, with the corresponding contingency table shown in Table 7.12.

Population data	Specific requirement	Value collected	Source
1. Exposure	Caesarean births	9% of births[a]	BirthChoiceUK website[387]
	Amniocenteses	15,000 in Britain each year[b]	CambridgeFetalCare[386]
2. Size	Number of children	774,840[a]	Office of population censuses
		1,064,157[b]	and surveys[418,419]
3. Cases	Diabetes cases	248[a]	Yorkshire Childhood
			Diabetes Register[7]

[a] time-points relevant to the study.

[b] more recent data (2010s).

Table 7.10: Diabetes data: Population data used.

7.3.1.1 Caesarean

Case data are extracted directly from the article.[7] The necessary calculations to determine control data for caesarean birth exposures are,

$$\begin{aligned} \text{population} &= \text{cases + controls} \\ 774,840 &= 248 + \text{controls} \\ 774,840 &= 248 + 774,592 \end{aligned}$$

$$\begin{aligned} \text{exposed population} &= \text{exposed cases + exposed controls} \\ (0.09 \times 774,840) &= \left(\frac{34}{196} \times 248\right) + \text{exposed controls} \\ 69,736 &= 43 + 69,693 \end{aligned}$$

$$\begin{aligned} \text{not exposed population} &= \text{not exposed cases + not exposed controls} \\ (774,840 - 69,736) &= (248 - 43) + (774,592 - 69,693) \\ 705,104 &= 205 + 704,899 \end{aligned}$$

Delivery by Caesarean	No	Yes
Controls	704,899	69,693
Cases	205	43

Table 7.11: Cases and controls in the diabetes data: Type of delivery.

7.3.1.2 Amniocentesis

The necessary calculations to determine control data for amniocentesis exposures are,

$$\begin{aligned} \text{population} &= \text{cases + controls} \\ 1,064,157 &= 248 + \text{controls} \\ 1,064,157 &= 248 + 1,063,909 \end{aligned}$$

$$\begin{aligned} \text{exposed population} &= \text{exposed cases + exposed controls} \\ 24,068 &= \left(\frac{14}{196} \times 248\right) + \text{exposed controls} \\ 24,068 &= 18 + 24,050 \end{aligned}$$

$$\begin{aligned} \text{not exposed population} &= \text{not exposed cases + not exposed controls} \\ (1,064,157 - 24,068) &= (248 - 18) + (1,063,909 - 24,050) \\ 1,040,089 &= 230 + 1,039,859 \end{aligned}$$

Amniocentesis	No	Yes
Controls	1,039,859	24,050
Cases	230	18

Table 7.12: Cases and controls in the diabetes data: Amniocentesis.

7.3.2 Results

Table 7.13 shows the odds ratios and 95% confidence intervals generated using the population data method, along with the published odds ratios from the original study.[7] The population results support the findings of significantly raised odds ratios for birth by caesarean and amniocentesis, although the OR estimates differ somewhat. Table 7.13 also shows the population odds ratios have much narrower confidence intervals, which correspond to the increase in the number of individuals considered in the population compared with the case-control study.

Exposure of interest	Published OR (95% CI)	Population data OR (95% CI)
Caesarean	1.84 (1.09, 3.10)	2.12 (1.53, 2.95)
Amniocentesis	3.85 (1.34, 11.04)	3.38 (2.09, 5.47)

Table 7.13: Diabetes data: Odds ratios and 95% confidence intervals (2dp) comparing the published odds ratios with those generated using the population data method.

7.4 Example: Stroke Case-Control Study

The article for this next example was selected as it was a newly-published case-control article at the time of method development (2014), which gave sufficient details of the dataset, and for which population data were likely to be available. The study also used participants from a location

which was not in the UK, and so allowed the method to be demonstrated in a population which differed from that of the Yorkshire childhood diabetes data. The raw data in the diabetes example were available, however this method of using population data can also be applied without raw data (as in the stroke example), provided a given published article contains sufficiently detailed summaries. This approach of verifying published results is demonstrated using a study of 100 computed tomography (CT) confirmed cases of stroke, with age and sex matched controls, from hospital attendees in India.[396] The population data from India are shown in Table 7.14.

7.4.1 Incorporating the Population Data

Despite using only information obtained from the published article, the same calculations were applied as in the diabetes example in §7.3. The resulting contingency tables are shown in Tables 7.15–7.17, which can be used to produce odds ratios and confidence intervals. Checks should be performed to ensure that each contingency table obtained is reasonable and that mistakes have not been made during the calculation stage. For example, Table 7.15 shows stroke to be a rare disease as required by a case-control study, and that around 23% of the population have hypertension, while Table 7.16 shows stroke to be a rare disease, and that around 5% of the population have diabetes. These summaries agree with the data in Table 7.14.

Population data	Specific requirement	Value collected	Source
1. Exposure	Hypertension	23%	World Health Statistics[420]
	Diabetes	65.1 million	International Diabetes Federation[421]
	Smoking	14.925%	World Bank[422,423]
2. Size	Population size	1.237×10^9	World Bank[424]
3. Cases	Stroke cases	18,012,222	Rightdiagnosis.com[425]

Table 7.14: Stroke data: Population data used, from India, during the study period.

7.4.1.1 Hypertension

Case data are extracted directly from the article.[396] The necessary calculations to determine control data for hypertension as an exposure are,

$$
\begin{aligned}
\text{population} &= \text{cases + controls} \\
1,237,000,000 &= 18,012,222 + \text{controls} \\
1,237,000,000 &= 18,012,222 + 1,218,987,778
\end{aligned}
$$

$$
\begin{aligned}
\text{exposed population} &= \text{exposed cases + exposed controls} \\
(0.23 \times 1,237,000,000) &= \left(\frac{62}{100} \times 18,012,222\right) + \text{exposed controls} \\
284,510,000 &= 11,167,578 + 273,342,422
\end{aligned}
$$

$$
\begin{aligned}
\text{not exposed population} &= \text{not exposed cases + not exposed controls} \\
(1.237 \times 10^9 - 284,510,000) &= (18,012,222 - 11,167,578) + (1,218,987,778 - 273,342,422) \\
952,490,000 &= 6,844,644 + 945,645,356
\end{aligned}
$$

Hypertensive	No	Yes
Controls	945,645,356	273,342,422
Cases	6,844,644	11,167,578

Table 7.15: Cases and controls in the stroke data: Hypertension.

7.4.1.2 Diabetes

The necessary calculations to determine control data for diabetes as an exposure are,

$$
\begin{aligned}
\text{population} &= \text{cases + controls} \\
1,237,000,000 &= 18,012,222 + \text{controls} \\
1,237,000,000 &= 18,012,222 + 1,218,987,778
\end{aligned}
$$

$$\begin{aligned} \text{exposed pop.} &= \text{exposed cases + exposed controls} \\ \left(\frac{65,100,000}{1,237,000,000} \times 1,237,000,000\right) &= \left(\frac{38}{100} \times 18,012,222\right) + \text{exposed controls} \\ 65,100,000 &= 6,844,644 + 58,255,356 \end{aligned}$$

$$\begin{aligned} \text{not exposed population} &= 1.237 \times 10^9 - 65,100,000 \\ \text{not exposed cases} &= 18,012,222 - 6,844,644 \\ \text{not exposed controls} &= 1,218,987,778 - 58,255,356 \end{aligned}$$

$$\begin{aligned} \text{not exposed pop.} &= \text{not exposed cases + not exposed controls} \\ 1,171,900,000 &= 11,167,578 + 1,160,732,422 \end{aligned}$$

Diabetic	No	Yes
Controls	1,160,732,422	58,255,356
Cases	11,167,578	6,844,644

Table 7.16: Cases and controls in the stroke data: Diabetes.

7.4.1.3 Smoking

The necessary calculations to determine control data for smoking as an exposure are,

$$\begin{aligned} \text{population} &= \text{cases + controls} \\ 1,237,000,000 &= 18,012,222 + \text{controls} \\ 1,237,000,000 &= 18,012,222 + 1,218,987,778 \end{aligned}$$

$$\begin{aligned}
\text{exposed population} &= \text{exposed cases + exposed controls} \\
(0.14925 \times 1,237,000,000) &= \left(\frac{49}{100} \times 18,012,222\right) + \text{exposed controls} \\
184,622,250 &= 8,825,989 + 175,796,261
\end{aligned}$$

$$\begin{aligned}
\text{not exposed population} &= \text{not exposed cases + not exposed controls} \\
(1.237 \times 10^9 - 184,622,250) &= (18,012,222 - 8,825,989) + (1,218,987,778 - 175,796,261) \\
1,052,377,750 &= 9,186,233 + 1,043,191,517
\end{aligned}$$

Smoker	No	Yes
Controls	1,043,191,517	175,796,261
Cases	9,186,233	8,825,989

Table 7.17: Cases and controls in the stroke data: Smoking.

7.4.2 Results

Table 7.18 shows the results for the stroke dataset, where all the exposures of interest have increased odds ratios when using the population data compared with the initial study data. However, the confidence intervals for the hypertension population odds ratio and published odds ratio do overlap. This could suggest support from the population data method for the odds ratio for hypertension, but possible disagreement between the published and population odds ratios for the diabetes and smoking exposures; with greater disagreement when considering diabetes. One possible cause for this disagreement could be participation bias. Recall that the controls in this dataset were hospital attendees; this could have resulted in Berkson's bias,[128] since those who smoke, have hypertension, or have diabetes, may have associated conditions requiring hospital admission. This higher proportion of smoking, diabetic and hypertensive controls than in the target population could have resulted in lower odds ratios in the published results. Hence participation bias is likely to have occurred. A related discussion regarding confounders is given in §7.5.2. As with the diabetes population results (Table 7.13), Table 7.18 shows narrower confidence intervals due to the increased numbers considered when using the population data compared with the original case-control study. The population numbers include values from the large Indian

population which incorrectly lead to an illusion of very little uncertainty. A discussion for the width of the confidence intervals is provided in §7.5.3.

Exposure of interest	Published odds ratio (95% confidence interval)	Population data odds ratio (95% confidence interval)
Hypertension	3.81 (2.11, 6.86)	5.65 (5.64, 5.65)
Diabetes	3.47 (1.76, 6.87)	12.21 (12.20, 12.22)
Smoking	2.24 (1.26, 4.01)	5.70 (5.70, 5.71)

Table 7.18: Stroke data: Odds ratios and 95% confidence intervals (2dp) comparing the published odds ratios with those generated using the population data method.

7.5 Method Overview

In this chapter a method to reduce participation bias in case-control studies using population data has been introduced. The theory has been explained, simulated examples have recovered the true odds ratios, including when data are MNAR, and the method has been applied to two real datasets; one where the raw data were available and another where sufficient data were extracted from the published article. For these examples, publicly available population data have been used, but in practice researchers may have access to medical records or similar information which are likely to give more accurate odds ratios, which may be less affected by biases. The method has also been used for the identification of potential participation bias, as shown in the Indian stroke example, where Berkson's bias[128] has been suggested. The remainder of this chapter critically evaluates the method, suggests extensions and amendments, and compares this method with other approaches, such as those discussed in Chapter 4.

7.5.1 Critical Evaluation and Method Requirements

7.5.1.1 Variable Requirements

In case-control studies, the outcome variable will usually be binary. However, the remaining variables may be binary, continuous, categorical or a mixture of these. Sometimes in a case-control study, it is of interest whether the participants have been exposed or not, indicating a

binary exposure, but this is not always so. For the method introduced in this chapter, the exposure and outcome will both need to be at least categorical if not binary, and each category should be well-populated. While 'well-populated' is not defined, the number of individuals within each category relative to the overall number of individuals should be taken into account. Subsequently, the width of any resulting confidence intervals should be considered (where confidence intervals will be wider if the groups defined by the categories are more similar), to ensure the estimates can be used to draw meaningful conclusions (since wider confidence intervals may not be very informative). Within any chosen categories, there must also be both cases and controls (or the relevant control population data), or else comparisons cannot be made and there will be problems in the computation of the odds ratio. This requirement also applied to the stratification approaches described in §4.2. When cases are recruited, it must be ensured that any variables which will be included in the analysis are recorded and that effort is made to avoid other forms of bias wherever possible, such as interviewer bias and recall bias.

Table 4.9 was used to summarise whether data from non-participants, the population, or regarding the participation variable were required for the current methods to be used to reduce participation bias. If added to Table 4.9, this method would require population data only. The nature of the variable associated with participation is less important, since participation is assumed amongst the cases and generally no longer required for control data collection. This includes information such as whether the variable is binary, continuous or categorical. It is assumed that the cases (approximately) are unaffected by participation bias, since studies have shown that case participation rates are often high, (see Chapter 2). It is also assumed that the population data contain information on the entire (or close to the entire) population, hence participation and factors affecting participation are less meaningful. This is particularly useful if there are usually several (possibly unidentified) factors affecting control participation, provided they do not affect the population data or the participation of the cases. Additionally, there is not the need to obtain information relating to non-participants, assuming that approximately all of the (randomly selected and hence representative and unbiased sample of) cases asked subsequently participate in the case sample and approximately the entire population contribute to the population data, as sought.

7.5.1.2 Considerations and Limitations of Population Data

Ideal sources of population data are those which capture information from the entire population of interest and which are considered to be reliable. Examples include population-wide health databases[59,60] or appropriate census data.[418,426] If there are any resources which assist verification of the estimated population values, efforts should be made to utilise them since population data can also suffer from biases (as will be discussed in this section).

Population Data Variables

For the population data to be available, the required variables are likely to be those usually collected through the census or those recorded for research. Census data may include number of dependents, income or gender, whereas variables collected for research purposes may include the number of hospital admissions or the number of individuals with a television license. The variables need not be collected for medical research, but could be those collected for marketing or surveillance purposes, and these data may be owned by a company or research group and not necessarily accessible.

Variables which are invasive to collect, such as tissue samples or questions regarding sensitive topics, may be more difficult to obtain and hence not recorded for the population or not available due to ethical constraints. Groups which have collected these data may also be unwilling to share their findings, which have likely been expensive or time-consuming to collect. In these instances, there may be no choice other than to collect the data using a traditional case-control study.

Population Data Accuracy

The population data should be from a time period as close to the study in question as possible. If it is a new study and the population data are being collected in place of the control data, the population data would need to be as recent as possible. However, if the method is being applied for the reanalysis of a historic case-control study, then the population data would need to be from as close to the time and location of the original study as possible. Data from alternative locations or periods may be required in some instances, but this should be avoided wherever possible, since data collection or variable recording may have changed over time, and there may be unknown differences between locations.

Answers to questions such as how recent, accurate or detailed the population data should be to

be suitable, will vary from study to study. Contributing factors include the funding and urgency of the study, and the comparative costs and time associated with recruitment and data collection. For example, if up-to-date data can be collected in an afternoon using an electronic survey, it may be preferable to collect new data than to use population data which are from three years ago. Alternatively, if the data collection would take two years, be difficult to recruit for, and cost two million pounds, then population data from three years ago may be more attractive. The required closeness of the data to the study period will vary depending upon the variables of interest and how these have changed with time, with minimal changes over decades allowing more flexibility than variable values which change daily. Any limitations resulting from the structure, detail or period of the data used should be reported.

Some population values may be unknown and hence approximated using other sources, such as another similar population. These approximated values may differ from the true values, so the data source should be carefully considered. There may also be instances where the same population data are available from multiple sources. Where this occurs, it is the responsibility of the researcher to determine whether any sources are more appropriate (more suitable area or time-period), or whether one source is more reliable than another (which biases the sources may suffer from and whether either source has been collected by a more reputable organisation). If the data from separate sources agree, then this can increase confidence in the results, if they are thought to suffer from biases differently. However, if the sources are contradictory, steps should be taken to try to determine why this may be so. Some variables may be formed by combining data from different sources, if for example data on males were collected in one dataset and data on females in another. Caution should be taken when combining data to ensure no additional biases are introduced, such as when the sources have overlapping categories and assumptions need to be made.

Population Data Biases

Population data may suffer from participation bias. Individuals who cannot read or write may be unable to complete a written survey, or those with severe mental disabilities may be unable to communicate certain answers. The possibility of any participation bias should be considered and the impact of such bias discussed. The source or approach with the least bias (if known) should be taken where possible. Population data such as census data should have high response rates, whereas optional health surveys or data collected by sales companies may be less successful

when recruiting. The 2011 UK census had person responses rates in England and Wales of 94%, varying from 82% to 98% in local authorities.[427] Although these response rates are high, if the non-respondents are similar with respect to the variable(s) of interest, and if they differ from the respondents for this variable, then bias may be present. Agreement between the case-control study results and the results from reanalysis using population data, may be due to the two sources being affected by similar forms of participation bias rather than the indication of a reliable result.

Population data may also suffer from missing data. Advice is available in the missing data literature, which includes options such as complete case analysis, imputation or multiple imputation.[292] The missingness mechanism[289] should be considered and any assumptions adhered to where steps are taken to account for missing data. Where there are large amounts of missing data or where the data are known to be MNAR in the population, it may be preferable to try to collect the control data using a traditional case-control study and avoid these missing values where possible. Any missing data are unlikely to be mentioned in quoted population values, which are assumed to be accurate. Missing data could result in values higher or lower than the true values, and if imputation or similar methods are applied, the estimated value should be reported with a confidence interval allowing for the uncertainty associated with imputing.

Population Data Precision

Population data, as with many other data, could be recorded or reported with varying degrees of precision. For example, when charities or companies report the number of individuals in a given area who have a disease or who use a given product, they often do so by rounding to the nearest hundred, thousand or ten-thousand individuals, and the actual value may not be available. In fact, any variable which does not take discrete values could be considered to be rounded,[428] with the degree of precision varying between variables and studies. Some values may have been rounded by the individuals, then again by the data collectors. The amount of rounding may also depend upon the precision of the question. For example, income per day/week/month/year may be prone to different amounts of rounding, such as to the nearest unit/ten/hundred/thousand pounds respectively. Rounding may also differ for variables such as income, depending on the value reported. Individuals earning small amounts may report more accurately than those earning large amounts, possibly as they are required to be more careful with their money or because rounding may affect their value more. Limitations of using rounded data are known.[429]

Rounded data have sometimes been treated as missing data, such as in imputation of rounded values[430] but this can cause additional problems.[431] Alternative approaches have been suggested,[431] such as maximum likelihood estimation,[428] but generally there are few approaches in the literature and rounding is often ignored during analysis.[428] Recent guidance suggests that where rounded data must be used, an assessment of the impact of the rounding should be performed.[429] Sensitivity analyses could be one way in which to conduct these assessments.

Some variables may be more prone to rounding than others, such as self-reported variables where individuals believe it is preferable to round up or round down. For example, individuals may round their income up to the nearest thousand pounds per year, or may round their expenditures down to the nearest thousand pounds per year. Rounding may also be advantageous when there is a possible consequence of their answer, such as increased financial support or the promise of a bonus for achieving a given target. This intentional rounding may also occur when the variables are sensitive, such as rounding down the number of units of alcohol consumed per week, or the number of cigarettes smoked daily, or rounding up the number of calories consumed daily in those who have experienced eating disorders. Rounding may also be more likely when the individuals are asked to estimate a given number over a long period of time, where the exact value may be hard to recall. For example, the number of times they have visited a supermarket in the last year. The nature of the variable and the time frame used should be considered when the data are collected.

In survey data it is known that rounded responses result in 'heaped' data, such that there are large numbers of responses at particular expenditures.[432] Rounding in surveys can be achieved in a variety of ways, for example by selecting a rounded value for a weekly activity and scaling up to an annual value.[432] Some respondents may not wish to report an obviously rounded figure and so may deliberately provide a value which appears to be precise, yet may not be. Whatever the reason for rounding, bias may be introduced. Tests could be preformed to establish whether the values have been rounded, such as whether values are more likely to end in a zero or a five, using a test or proportions or similar. A recent medical imaging study[433] asked individuals to estimate the percentage stenosis in a vessel (the percentage by which the vessel had narrowed compared to its initial size), and it was found that between 10% and 90%, all values given were to the nearest 10. However, below 10% a value of 8% was given, and above 90% the values 92%, 93%, 95%, 98%, 99% and 100% were all given, showing how precision can change at extremes of the scale.

Another limitation of population data may be measurement error. Depending on the variable recorded and the time-point of the data collection, it may not have been measured accurately. Historical data may have been recorded using older, possibly less accurate measuring devices, or practice for recording certain variables may have changed over time, with varying degrees of accuracy. Population data may also have been recorded by different individuals, in different counties or trusts, and procedures or equipment may vary from area to area, with some more accurate than others and with different training.

Rounding to some extent will also occur when measurement tools are used, since each tool will only display values to a particular number of decimal places or increment. Digital displays will have a limited amount of display space and analogue displays will require the reader to report to the nearest increment, which itself will be marked to a certain precision. The amount of rounding will also affect different variables to different extents. For example, rounding the number of pounds an individual earns annually to the nearest ten may not affect the results greatly, whereas rounding the amount of operations an individual has had in the last year to the nearest ten is likely to lead to uninformative data. However, provided cases and controls are measured in the same way, these biases may be less important, since they can be assumed to be similar within each of the disease groups.

Population Data Structure

For the population data to be suitable for use in the method, the data must be recorded on the variables of interest and using an appropriate structure. For example, if age is a variable of interest, it may be that the required ages should be recorded in bands of five-years, whereas the population data may only be available in bands of twenty-years. In these instances, the researcher must decide whether to use the data and note the bands as a limitation of the study, or to conduct a traditional case-control study and collect the data as intended. Alternatively, approximations can be made when population data are available but not in the required format. For example, in the diabetes dataset it was assumed that the number of 15 year olds in Yorkshire was approximately a fifth of the 15–19 year old Yorkshire population.[418]

The Frequency With Which Population Data are Collected

Depending on the nature of the population data, it may be something which is collected monthly, annually, once a decade, or only through funded projects, for example through charity research

work; which may be a potential source of bias itself. In some instances, the population data may be rather out-dated and there may have been a dramatic increase or decrease in the data values since the last data collection. However, if data from the required time period are unavailable, more recent or out-dated information may be used as an approximation in older or newer studies respectively. This was true for the amniocentesis data, where only recent statistics could be found rather than the required data from the 1990s, so corresponding recent population values were used. Where this method is applied as an adaptation of the case-control method to new data, rather than as a sensitivity analysis to historical data, the case data will be recent and the results may be affected when comparing the recent case data to out-dated control data, so a new case-control study may be preferable where possible.

Population Data Summary

The requirements of the population data for this method to be advantageous over a standard case-control study, are that these population data are less biased than the control data that would otherwise have been collected. It may not be known in studies which set of data would be least biased, but the population data method offers an alternative approach for data collection, which may be cheaper and faster than conducting the control section of the study. Highly detailed, accurate and date-appropriate data would be most desirable, but there are likely to be instances where this may not be possible.

Data obtained from the population must correspond with the case data recorded in the study and the case data must include at least the number of cases exposed and unexposed in the study. If more detailed analyses are to be conducted, this additional information must be recorded from the cases during the study, or be available in records for the reanalysis of past studies.

7.5.1.3 Mechanisms of Participation Bias

For participation bias to be present in case-control studies, participation must be affected by both the exposure and outcome of interest, and be conditioned on as detailed in §2.3.4.1. Through simulation in §7.2.5 (MAR) and §7.2.6 (MNAR), this method using population data has been shown to successfully recover the true odds ratio when participation bias is present.

There may be other variables which are related to participation, or to the variables in the

association of interest. It can be difficult to distinguish between variables which affect this association and which do not. Causal diagrams such as DAGs are useful for identifying variables such as confounders which should be included in analyses. If required, additional variables can be included in this method using stratification, by collecting population data which are sufficiently detailed. Further details will be given in §7.5.2.

7.5.1.4 Strengths of the Method

The method introduced in this chapter is very simple and quick to apply, and generally far cheaper and easier than recruiting controls for a case-control study. It also had the advantage that if a population value is used and later thought to be inaccurate, the calculations can quickly and easily be rerun to generate improved estimates. In addition, if this method is used in future (rather than past) studies and only case data are collected, the ethical application for the study may be simplified. Cases can usually be obtained from one source and since the population data have already been routinely collected, the ethical considerations should be reduced. For analysis, since specialist software is not needed, there is no requirement for additional software licenses or specific high-performance technology. The method can even be easily applied by hand, allowing it to be accessible to most studies without incurring additional costs.

This approach allows the time and resources of future studies to be focused on the collection of case data, resulting in a larger sample of cases and a final dataset which has been collected more efficiently. If adopted widely, potential controls would receive fewer requests to participate in studies, and consequently this could aid recruitment in studies which still require control participants.

If carefully selected, the population data are likely to have reduced participation bias when compared with the corresponding control data, yielding more accurate results and increasing the chances of determining the true cause of a disease. Reduction of any form of bias is advantageous and if the true cause of a disease can be identified, there is the possibility for prevention techniques or potentially a cure to be developed.

All steps in the method have the potential to be conducted using case information only in the published article, without the need for the original dataset. This feature allows the method to be

applied to countless previous studies, provided sufficient data have been published or retained. This analysis could also be repeated for all the variables published or recorded in a study, to see whether any potential risk factors may have been mis-categorised. Where this method forms part of a reanalysis, it could ultimately support the findings from the original study, or identify potential bias in the results, possibly due to non-participation.

7.5.1.5 Limitations of the Method

In this method it is assumed that the outcome is known for all individuals in the population, from medical registers or similar. Those on the register are classed as cases, whereas those not on the register are considered to be controls. There is of course the possibility of misclassification bias from those who are undiagnosed or incorrectly diagnosed, however this is assumed to be minimal. The effects of misclassification are well-documented,[434] including in case-control studies[435] and this can lead to misclassification bias which is a form of information bias.[14]

The method has some assumptions, for example that the population data are available and reliable. It assumes that a large proportion of the population has been included and that the data do not suffer from participation bias in the same way that the control group in a case-control study would have. The quantity and quality of the population data do not have defined levels to attain, but they must be at least as complete and as accurately recorded as the corresponding control data would have been.

Some of the variables required for analysis may need to be measured, and hence could be affected by measurement error. The nature of the error may also differ between the cases and the controls if these measurements were collected in different ways. For example, case data may be taken from medical records during routine appointments, whereas control data may be collected retrospectively from memory and susceptible to recall bias.[14] For self-reported variables, cases or controls may (intentionally or not) report higher or lower values than the true value to hide behaviour which could be perceived negatively, or they may emphasise positive behaviour. However these limitations also apply to traditional case-control studies.

Odds ratios are calculated from this method and can be very sensitive to changes in values, particularly when the sample size is small, or when there are small proportions in some categories.

Any biases resulting from missing data or rounding (covered in §7.5.1.2) could lead to fluctuations in the study estimates and possibly lead to incorrect conclusions being drawn. However, odds ratios are also used in traditional case-control studies.

Assumptions or approximations in the population data could mean the confidence intervals for the final odds ratio should be wider than if the true values were known, as will be discussed in §7.5.3. The sample size of the cases and the robustness of the results, will also affect how influential the accuracy of the population data are. A large study of cases and an odds ratio which is not greatly affected by small changes, is preferable.

7.5.2 Extensions

This method has been presented in a relatively simple form, but can be extended to accommodate more complicated scenarios. For example, rather than having the situation where each individual is exposed or unexposed, the exposure may be continuous. As the proportions of exposed and unexposed individuals are used during the calculation in §7.2, a few options are available. The first and most simple option is to dichotomise the continuous variable into 'exposed' and 'unexposed' using a (clinically relevant) cut-off value, then the method can be utilised as described in §7.2.3. The cut-off value may be difficult to define, but published literature or expert advice may be able to advise.

Another option could be to divide the continuous exposure variable into a number of categories for a more detailed analysis. For example, let the exposure of interest be 'age' which is measured in years; age may be to split into three categories; 'children (0–17 years)', 'younger adults (18–49 years)' and 'older adults (50+ years)'. This level of detail in the analysis would require the same age categories to be used when collecting the case data, and for this level of detail to be recorded at the population level; the number of individuals with each disease status per age group. Comparisons can then be made between age groups to estimate the effect of age on the outcome of interest.

If additional variables need to be considered during the analysis, for example confounders, these can be incorporated into the method using stratification. Stratification is a recognised method used to account for confounding variables,[436] and causal diagrams such as DAGs may be useful

for identifying such variables. This step assumes the confounding variables are recorded for the population and cases in the required level of detail, which may also include stratification by other variables. If this level of detail is available, then confounding can be accounted for. For example, let the confounding variable be sex. If data regarding exposed males and females can be obtained, and data regarding male and female cases and controls, then the data can be analysed at this more detailed level, i.e. considering sex. The males and females are treated as though they are subpopulations, so the same method as detailed in §7.2 can be applied to solely males then solely females. While stratification is easily applied to this new method, the number of confounders and the number of confounder categories will be limited by the size and spread of the data across the variables, as usual with stratification and as was described in §4.2. This level of detail may not always be available in the population, and may become less likely the more variables there are and the more obscure the required data. In instances where there are many confounders, confounders with several categories, or where the confounders are continuous, it may be preferable to use an alternative approach such as regression analysis, provided the requirements of this method are satisfied.

To consider several confounders, the data would need to be recorded at each combination of the confounding factors. For example, if sex and race are confounders, then the data would need to be recorded for each combination of sex and race; white males, white females, black males, black females etc, with each combination treated as a subpopulation. If the confounder is a continuous variable such as age, this may need to be categorised, as data are unlikely to be available for each year of age at the population level. The case data will of course also need to be recorded at the same detailed level for the analysis to be conducted. Each stratification results in a smaller sample size, hence the width of the confidence intervals are likely to increase.

The basic idea of incorporating population data into a case-control study to reduce bias may be an approach which could be adapted for use in other study designs which suffer (likely to a lesser extent) from participation bias. It may also be possible to utilise this method in other areas such as survey non-response, which encounters similar problems to non-participation in case-control studies.

7.5.3 Confidence Intervals

Ideally, the estimate for each population value should have an associated level of uncertainty and corresponding confidence interval, but this is not usually reported in practice. Consequently, confidence intervals are not presented for the population data used in this chapter and subsequently not incorporated into the analysis.

The seemingly larger sample sizes used in this approach generate narrower confidence intervals, but future work could consider amending these intervals to account for the added uncertainty of any predictions at the population level. Confidence intervals narrower than appropriate could result in an increased chance of exclusion of the true odds ratio, or similarly exclusion of an odds ratio of one, suggesting an effect where there may not be.

One approach for investigating the width of the confidence interval may be to conduct a sensitivity analysis using the lowest and highest possible values for the population value prediction, which may be the lowest and highest plausible values, or the values which may have been rounded to form the estimated population value. For example, if the population value is rounded to the nearest thousand, the sensitivity analysis could investigate values from 500 below to 499 above the estimate.

Another approach may be to use the population data to form a control group for the study, but then scale the control values back to the size of the original study, giving a 1:1 or 2:1 ratio with the cases as commonly used in case-control studies. This may be the most preferable option since it would result in the population values forming a group similar to the control group, but with likely less bias and with confidence intervals of a similar width to those obtained from a comparable case-control study.

Bayesian approaches may be useful, since they allow prior information about the variable(s) to be incorporated into the analysis, plus they can represent the uncertainties related to parameter values,[437] including population data. In contrast, maximum likelihood approaches often involve fixing the values of parameters, which may impact on the final results and for which there may be some uncertainty.[437] Prior information may be particularly helpful in instances when the exact required population data are unavailable. Since Bayesian approaches combine prior beliefs with data, there is other information incorporated into the analysis besides the population value itself,

which may be affected by rounding or bias.

An approach similar in style to multiple imputation could be adopted as an alternative, whereby a population value is sampled from a range of plausible values several times, and the results combined using Rubin's rules or similar.[290] This would result in an overall estimate with a confidence interval which takes into account the uncertainty of the population data, which could subsequently be used throughout the analysis and in the final odds ratio estimate. Alternatively, the population data could be scaled down to the size of the case sample and the control part of the case-control study could be sampled from the extracted population data. This would result in wider confidence intervals which incorporate both the uncertainty from the estimated population value and the size of the case group in the study.

7.5.4 The Method as a Sensitivity Analysis

Sensitivity analyses were covered in detail in §4.1, and references to the epidemiology literature were provided. Comparisons can be made between the sensitivity analyses mentioned previously and where this approach using population data is used as a sensitivity analysis. The literature searches for sensitivity analyses in Chapter 4, and for this method in §7.1, found no evidence of this approach or similar being used as a sensitivity analysis for case-control studies.

Application as a Sensitivity Analysis

This method can be used as a sensitivity analysis in different ways. Firstly, it may be used to generate odds ratios for comparison with the original case-control study. This will essentially compare the control group in the study with a similar group generated from population data. Any differences between the odds ratio estimates will indicate differences in the two groups acting as controls and may, but not necessarily, suggest participation bias. Secondly, it can be used as a sensitivity analysis for the population data, with examples including the analysis being conducted with the highest and lowest plausible population values, or the highest and lowest population values given that the value is rounded to the nearest hundred (i.e. 49 above and 50 below the given value). It could also be used to see how robust the odds ratios are, by amending the population values by increasing amounts, to see at which point the conclusion would change; a similar approach to Rosenbaum's extension of Cornfield's inequality in §4.1.1.3.

Compared with other sensitivity analyses, such as decomposition of the odds ratio[295,301] or Rosenbaum's extension of Cornfield's inequality,[297] this is a relatively straightforward approach, since it uses simple calculations to obtain the dataset used for analysis and adopts the usual logistic regression approach to calculate the new estimate. Difficulties are more likely to be encountered in locating appropriate population data, than in the method itself.

Sensitivity Analysis Discussion

Any claim of negligible participation bias may be strengthened if the results from the two sources, population data and control data, are similar. Problems may be encountered when using this approach as a sensitivity analysis if the results generated using control data and those using population data contradict. Effort should be made to uncover why the contradiction has occurred and which biases each of the datasets may have suffered from. Ideally the dataset with the least bias (if known) should be used, but often it will not be possible to choose between the datasets on these grounds. It may be necessary to treat both sets of results with caution, especially if substantial bias is possible in both forms of data collection. Acquisition of any further sources for the population value may be advantageous for comparison. In some instances, it may be that conclusions cannot be drawn confidently as it may be unclear why the data are contradictory. Where results generated using control and population data do agree, this can be viewed as a form of triangulation, if it could reasonably be believed that the two sources would not suffer from participation bias in the same way and this may increase confidence in the results. However, agreement does not guarantee a correct estimate, since the agreement may be due to unidentified biases in the datasets. As a sensitivity analysis, this method has the advantage of being quick to apply, without requiring difficult mathematics or specialist software. It is certainly no more complex than other sensitivity analyses described in Chapter 4. It can also be easily updated and rerun if revised estimates are discovered at the population level.

Triangulation is often a qualitative rather than quantitative approach, which relies upon the compared data being collected independently.[438] Here the case group is the same for both analyses, although the population and control data are likely to be collected independently. While some use triangulation as a means to test validity, this has been questioned by others as it assumes that a weakness in one method will be corrected through another method.[439] Therefore triangulation is recommended to ensure the findings are robust and well-developed,[439] as suggested here.

This approach is also referred to as mixed method analysis and further guidance is given in the literature.[438,439]

Sensitivity analyses have been suggested as a sensible approach, whether used on their own or in conjunction with another method to reduce bias,[305,306] so using a sensitivity analysis such as the method here or one of the methods suggested in Chapter 4, would be advised. The choice of sensitivity analysis will depend upon the available data, such as the required population data, and the aim of the analysis, whether it be descriptive or quantitative. Any researchers who may be uncomfortable using data not collected specifically for their study, may wish to use this method as a form of sensitivity analysis after they have conducted their case-control study. This may be with the hope of supporting their study results, or to assist any claim of minimal participation bias they may have made.

7.5.5 Comparison with Alternative Methods

7.5.5.1 General Advantages Over Other Methods

Assumptions

The current methods have at least one of three requirements; the assumption that the variable associated with participation is recorded, that population data are available, or that data regarding the non-participants are recorded, as shown in §4.5. Each of these three assumptions are mechanisms to determine the characteristics of *who* is missing from the study data. This method which utilises population data, avoids the need for the variable associated with participation, or the non-participant data, to be recorded. The requirements for the population data may also be simple and less restrictive than other approaches that require population data, such as the bias-breaking method in §4.4.4.

This method does not adapt the biased control data as other methods do, but instead replaces this potentially biased group with population data, which should be less biased. Therefore, the missingness mechanism is less important and does not form an assumption for the method, whereas data which are MAR is an assumption in several of the methods discussed in Chapter 4, such as imputation in §4.4.2 and weighting in §4.4.1. Therefore, this method can be applied even when non-participation causes the data to be MNAR, as demonstrated using simulation in §7.2.6.

Since MAR is an untestable assumption, a method which is applied in the same way regardless of whether the data are MAR or MNAR, is beneficial. This particular aspect shows a major advantage over some existing approaches. The CEGs in Chapter 5 may also help to decide whether the data are likely to be MNAR and if so, this method can be chosen.

Other Advantages

This method has the advantage of being much quicker and cheaper than traditional case-control studies, since the control group do not need to be 'recruited'. Instead population-wide databases (possibly publicly available) can be utilised, allowing greater resources to collect case data or the ability to fund a study on a tighter budget. The simplicity of the population data method and the ability to recalculate the odds ratio should an improved population value be presented, are advantages over approaches such as the bias-breaking method, which is far more time-consuming and complicated. This may result in greater consideration of participation bias from researchers if they are not deterred by methods deemed to be difficult and lengthy.

Other methods which require population data are weighting and sensitivity analyses as shown in Figure 4.4. The method proposed in this chapter does not require the data to be MAR which is an advantage over weighting (see §4.4.1.1). Also, the method here can be used as a sensitivity analysis as described in §7.5.4 and is easier to implement than some other sensitivity analyses discussed in §4.1.

7.5.5.2 Direct Comparisons Between Methods

Considering the three most frequently used methods from the assessment in Chapter 3, which were described in detail in Chapter 4, two of these (variable adjustment and stratification) required the variable associated with participation to be identified and recorded. This may not always be possible and so this method omitted that assumption. The other method (sensitivity analysis) did not always provide an adjusted estimate for the odds ratio, whereas this method does, plus this method can be used as a sensitivity analysis as demonstrated in §7.5.4. This method therefore offers an alternative to the most frequently used methods.

This method is completed in one analysis, so offers fewer stages compared with methods such as stratification (§4.2), which often analyses subsets of the data separately then combines these

subsets. Stratification can be used in conjunction with this approach, as described in §7.5.2, but it is not necessary for the method to be used. However, where the exposure of interest is continuous and it is preferable to keep the exposure in its initial form, rather than to categorise or recategorise for use with the method proposed here, another method to reduce participation bias may be preferred.

This approach suffers from the same limitations of population data (§7.5.1.2) as inverse probability weighting (IPW), when population data provide the weights (see §4.4.1). In IPW, individuals can only be represented if there is a participant with similar characteristics, but this limitation does not apply here, since data are extracted directly from the population, regardless of the characteristics of participants. Both methods can essentially eliminate missing data by replacing unknown data with population data, hence reducing the width of the confidence intervals yet adding uncertainty in the estimate. The approach here has the advantage of not requiring the data to be MAR and is not computationally intensive as IPW can be.

Chapter 5 showed how CEGs can be used in case-control studies. The approach proposed here differs from CEGs, partly due to being numerical rather than graphical. CEGs can be created using solely the study data, although population data (or similar) can be used to specify non-uniform priors, whereas the approach here requires external population data, but avoids the need to collect control data for the study. CEGs can incorporate missing data and draw conclusions about the missingness, but the numerical approach in this chapter instead aims to replace missing data with previously recorded values. Both approaches are useful, but the choice between them will depending upon the research question, the available and accessible population data, and the structure of the data. CEGs may be more time-consuming to apply than the method utilising population data, but the population data may be time-consuming to collect from reliable sources.

7.5.5.3 Deciding Upon a Method

The current methods were summarised in Table 4.9 and Figure 4.4, with the methods split by the assumption of there being available population data, non-participant data, or data regarding the participation variable. While in some instances there may be a selection of data from reliable sources, it is more likely that the choice of method will be restricted by the available data. It is advised that the researcher eliminates unsuitable methods where particular data types are unavailable, by using Figure 4.4, then the assumptions of any remaining methods are checked. It is

likely that few methods will remain. Examples were given in §4.5.1 of how to use the flowchart for method selection. This population approach would sit alongside sensitivity analysis and weighting in the methods where relevant population data are available. As shown using the stroke study in §7.4, aside from population data, only the published data were required for basic analysis.

It is assumed that the population data are obtained using means which are easier and cheaper than the control data would have been. If the required population data are owned by a company which has restricted data sharing policies or which charges large amounts to obtain the data, then a traditional case-control study may be preferable. Where the population data are viable logistically, financially and practically, and when the data are known to be recorded accurately and as complete as possible, then population data may be more suitable for the study than control data.

The data must also be categorical for the proportions to be calculated. Where variables are continuous, there must be the ability to sensibly categorise them, using reasoning for the chosen number of groups and break-points. Experts may be able to advise on clinically-relevant divisions of the variables, but these must also be context specific, allowing both a reasonable proportion of the data to lie within each newly-formed category and for clinically-relevant outcomes to be reported. If the variables cannot be split into categories, then an alternative method may be preferable. However, with data sharing increasing it may be possible in the future for sufficient data to be available such that interval data could be analysed in this way.

7.6 Summary

Identifying the true causes or risk factors of a disease is an important step towards developing a cure or preventing others from becoming cases. An amendment to the standard case-control analysis method, such as the one proposed here which has been developed to reduce participation bias, could help to yield more accurate results and move closer towards discovering the cause of a disease. This proposed method unfortunately cannot be used in all circumstances, but has advantages over traditional case-control studies when it can. It also has dual applications; either as an adapted case-control study, or as a sensitivity analysis for previous studies, depending on the requirements of the researcher. Each study this method is applied to could benefit from increased knowledge about the possible causes of a disease, which should lead to improved healthcare.

Chapter 8

Conclusion

8.1 Overview

This chapter will summarise the findings of the entire thesis, critically evaluate them, and compare the work here with similar work in the literature. Discussions will follow for three areas of the thesis; the role of graphical models, the use of population data and an overview of the methods available to investigate participation bias. Suggestions for future work will then be provided, along with a final overview of the thesis.

8.1.1 Case-Control Studies

Despite their time and financial efficiency, the use of case-control studies has declined in recent years. This may be due to the awareness among researchers of limitations such as participation bias. However the work here has shown that case-control studies are not the only design to suffer from such bias, and that careful consideration of participation bias allows this study design to remain reliable.

The identification of biases resulting from confounding variables or through non-participation can be achieved using causal diagrams such as DAGs. However, since the true causal associations between variables are often unknown, it is recommended that changes in these causal diagrams are investigated and the robustness of the results reported. Sensitivity analyses are one way in

which the robustness of results can be assessed, and Cornfield[296] and Rosenbaum[297] have both suggested ways to quantify how large an unmeasured effect would need to be to alter conclusions.

The focus throughout this thesis has been the identification and reduction of participation bias. While this is an important bias which can be particularly problematic for case-control studies, there are other biases such as those arising from measurement error, from unmeasured confounding variables in the association of interest, or due to the misclassification of disease status. The retrospective nature of case-control studies also means they can also be susceptible to recall bias. These biases are beyond the scope of this thesis, but are important considerations of case-control studies and other study designs.

The elimination of any bias is important to yield informative results, which can ultimately lead to improved patient care. If bias remains but affects the study results only marginally and hence conclusions still stand, this is of less concern than if the bias is causing conclusions to change and subsequently the incorrect advice to be given to patients and the general public. Authors should be encouraged to provide readers with adjusted and unadjusted results, through supplementary material if need be, to allow the readers to judge for themselves (by the suitability of the adjustment and the change in results) the plausibility of the conclusions. As far as ethical approval allows, authors should also make the study data and analysis available so that the results are reproducible. This ensures not only that the findings can be critically evaluated by others, but also that no data are wasted (hence utilising patient time and research funding to its full potential) and that (as close to) the true associations are uncovered.

8.1.2 Participation Bias

Participation rates in epidemiological studies have been declining in recent years, particularly amongst the control group in case-control studies. As different study designs and topics of interest suffer from non-participation in different ways and for different reasons, it is unlikely that one strategy would increase participation rates or reduce participation bias for all studies. As a consequence there have been several methods proposed to reduce the effects of participation bias, which have not previously been thoroughly compared. Hence the aim of this research was to investigate existing solutions to minimise participation bias in case-control studies and to suggest novel solutions as appropriate.

Although factors affecting participation rates have been considered within this thesis (such as in §2.3.5), some authors correctly highlight that increased participation does not necessarily result in reduced participation bias.[172,256] Using techniques such as incentives to increase participation rates may in fact increase bias. A shift of focus from participation rates to bias may save time and resources by not chasing unwilling participants, which in turn could be used to conduct a detailed participation bias analysis.[440,441] To aid this shift, journals could insist all surveys or studies requiring participants detail a participation bias calculation, for judgment by the reader. Alternatively journals could adopt standardised formulae to calculate rates such as those proposed by The American Association for Public Opinion Research (AAPOR),[43] which would at least provide guidance to researchers and allow easier comparisons between studies.

Regardless of the requirements imposed by journals, authors should provide a participation statement so the readers can compare sample and population characteristics, to judge population representation of the sample, and hence the generalisability and validity of the results. Providing details of the population of interest where possible can also help to assess bias, for example a study may have more female than male participants but if the study is concerning breast cancer survivors, a higher number of females than males is expected.

8.1.2.1 Confusion in Participation Bias Terminology

Terms which are related to participation bias but which have different meanings are sometimes used interchangeably in the literature. This does not aid the understanding of the definitions among readers, nor their grasp of the methods to reduce these biases. Consequently, relevant literature can also be difficult to find through a database search as it can be termed as 'participation bias', 'non-response', 'selection bias', 'self-selection bias', 'co-operation bias' or other similar phrases. In addition, different fields may have their own methods (and names) for dealing with such bias, but these may not be known in other fields. There may also be similarities between these methods which might not have been identified.

In addition, in the literature there is not agreement between the definitions of participation bias, selection bias and confounding, with some authors using phrases such as selection-confounding. Here the main focus was on bias resulting from conditioning on the collider variable (participation) between the exposure (or cause of) and the outcome (or cause of).

As far back as 1981 there was awareness of the confusion caused by differences in the definitions for confounding and selection bias in case-control studies, so attempts were made to clarify them.[442] As recently as 2004, authors such as Hernan[111] still tried to rectify the confusion with a simple distinction between the two biases and this has become a key paper. However, so did Rothman[18] in his key epidemiology textbook and unfortunately these definitions subtly differ. This is therefore an ongoing problem which is still not entirely resolved. In addition, the application of the definition of participation or selection bias to case-control studies differs, since ORs are often used and are more robust. This means that they can remove bias resulting from participation or selection which is dependent upon the outcome only. These subtleties in the literature may explain the seemingly contradictory findings for when bias arises and when it can be accounted for. There is the need for more articles such as the one from Hernan[111] which try to clarify biases, but there is also the need for agreement amongst researchers in these areas and the consensus of a definition. It may be that researchers are applying definitions with different meanings without realising and hence may not be correcting for bias as intended. Before consensus is achieved, or even just as good practice, researchers should state the definition used in their work, even if just briefly (e.g. "selection bias is defined to be the conditioning on the collider variable named selection"), to avoid this confusion.

8.1.3 Methods to Investigate Participation Bias

The methods included in Chapters 4, 5, 6 and 7 to reduce participation bias are suitable for different data structures and each rely on different assumptions. However, there are similarities between these methods and often they complement each another, or can be used in conjunction with one another. There is no correct method or approach for every study, but study-specific selection should prevail. The choice of an appropriate method for reducing participation bias can be eased using a guidance tool.[3] Researchers should be aware of the possibility of participation bias and consider methods to reduce it, and readers should not immediately dismiss findings from studies which have mentioned participation bias. Sensitivity analyses are often beneficial and can usually be included as supplementary material even when there is not space in the main article. Sufficient detail should be included for the research to be reproducible, so that other research groups can amend or continue a given analysis, with a view to testing the robustness of the results and findings

under alternative assumptions. However ethical approval and copyright restrictions may prevent the full dataset or the exact method from being available in some studies.

8.2 Findings

The background information in Chapter 2 introduced case-control studies and participation bias, and drew links between them. Key findings included the increase in non-participation and the decrease of case-control studies in the literature. The review of literature from 2007–2015 showed that the characteristics of individuals who participate has generally remained unchanged compared with pre-2007, but that the way in which data are collected is shifting towards more technological means such as through social media, or using smartphones and tablets.[1]

The assessment of participation bias in three high-impact epidemiology journals reported the presence of participation bias in recent literature and the actions typically taken by authors.[2] The assessment found that many of the studies were unlikely to have been affected by participation bias due to the study design, but those that were affected used similar approaches to investigate the bias.

The methods available to investigate participation bias in case-control studies were summarised in Chapter 4 and a guidance tool was developed to aid the selection of an appropriate method. This research uncovered similarities in the requirements of the methods available, such as the need for data to be recorded on non-participants or for external data to be available.

Chapter 5 showed chain event graphs were compatible with case-control data. While these graphs are not a rigorous inferential technique which fit a model and return parameter estimates, CEGs could be used to draw conclusions from data, including those regarding the missingness mechanism, or as an explanatory tool prior to a formal analysis. Chapter 6 demonstrated how these graphs could be adapted to be more useful for case-control data and to directly investigate non-participation.

The general unsuitability of current methods when data are MNAR led to the proposal of a new approach[6] in Chapter 7, which used population data in place of the possibly biased control group within a study. This new method allowed a cost-effective way for researchers to verify past case-control study results and provided an alternative approach for future studies.

The thesis also used a diabetes dataset throughout and confirmed previous findings that childhood type I diabetes is likely to be more commonly found in children who were delivered by caesarean or whose mother had at least one amniocentesis during pregnancy. The association with diabetes was also shown to increase when both procedures were carried out.

8.2.1 Generalisability of the Findings

The review of participation in studies from 2007–2015[1] is likely to be generalisable only to similar studies to those included in the review, during a similar time-period. The review showed that the characteristics of participants are generally unchanged compared with the review pre-2007, but the way in which data are collected is changing. Further technological advances mean that these methods may continue to evolve and the findings here should not be assumed to be generalisable into the distant future. However, the discussion in the review could be helpful to researchers who are considering using new technologies to recruit participants or to collect data.

The participation bias assessment conducted in Chapter 3 is informative as a 'snapshot' of the occurrence of non-participation and how authors address the problem in high-impact epidemiology journals. However, the findings are likely to have differed had another year or month been used to conduct the assessment, if different journals were selected, or if the subjective assessment was conducted by another researcher. The assessment was intended as a general idea of the presence of non-participation and an overview of the approaches taken by researchers in a typical selection of epidemiology journals, and this was achieved. The assessment could have been improved by including more journals and assessing over a longer period of time, but this was limited by the time required to read and review each article. Therefore the assessment is informative, but should not be taken to be definitive.

The critical evaluation of the current methods in Chapter 4 and the guidance tool in §4.5 which is aimed to aid the selection of a suitable method, are generalisable to participation and selection bias. The guidance tool can be adapted to be specific to certain diseases (where particular variables may be recorded, or certain population data are available). Over time, new approaches can be added to the tool, unused approaches can be removed, and additional criteria can be added as new sources for deducing those who are missing from a study are developed. The tool could also be extended to include the outcome of interest, whether it is the odds ratio of a case-control study, how robust

the results are, or details regarding the missing values. Therefore the findings in Chapter 4 are generalisable to participation and selection bias, and may include other study designs, provided the assumptions of the methods hold.

Chapter 5 introduced chain event graphs (CEGs). The CEG methodology has previously been applied in statistics and artificial intelligence, but there was no evidence of use with case-control data. The application of CEGs with case-control data was demonstrated but was not specific to the diabetes data used in this thesis and so is generalisable to all case-control studies. Their use with other designs, such as cohort studies,[365] has previously been demonstrated and there is potential for their use with further study designs. The adaptations proposed for CEGs in Chapter 6 were intended to address problems often encountered with case-control data, and whilst applicable to all case-control data, they may also be useful for areas outside of epidemiology, such as in survey non-response.

The population method suggested in Chapter 7 was developed for use with case-control studies which may suffer from participation bias in the control group. However, this approach may be generalisable to other areas where bias is a problem, to replace potentially biased data with population data thought to be less biased. This relies on the assumption that the population data are available, in the required format, and do not suffer from biases in the same way the control data would. This method could also be applied as a sensitivity analysis, for use with previous case-control studies to verify or question results, and used in other areas provided individual level data are not required, and any available summaries are sufficient for analysis.

The findings generated from the Yorkshire diabetes data may be generalisable to other areas within the UK, for individuals who were children in the 1990s. The finding that caesarean delivery and amniocenteses are associated with an increased probability of the child having type I diabetes should be investigated further, as should the increased probability from the combination of these two variables. The structure and summaries of this dataset could be compared with more recent case-control studies of childhood type I diabetes to see if changes have occurred through time, to determine how relevant these data and findings are to current children.

8.2.2 Critical Evaluation of the Findings

The review of participation in studies from 2007–2015 confirmed that the characteristics of participants had not changed substantially in recent years and indicated the change in the way data are collected, with more studies using smartphones and social media, and these findings may assist recruitment in future studies.

The assessment of participation bias in high-impact journals is useful as an overview, but may have been improved by including more journals, over a larger period of time and if the categories of the assessment were redefined using a stricter criteria. This would enable another researcher to conduct the same assessment and reach the same conclusion. However, since each article was read and assessed by the same researcher, this ensured that each article and each journal was evaluated in the same way and consistently. The assessment was also repeated at a later date to ensure the results were reproducible. The assessment was informative for the thesis, but too subjective to declare definitive results.

The guidance tool could be viewed as too general, or too simple, since it includes a variety of methods which could be used to reduce participation bias, and it still requires the user to verify all the assumptions of the method. However, a more complex tool which incorporates all the assumptions soon becomes less user-friendly, plus it is encouraged that researchers verify assumptions from the original literature rather than just external resources. The tool was intended to be general, to be applicable to a wide range of studies and to include the methods which were presented in Chapter 4, and these points were achieved. It was introduced as a basis which could later be adapted to include more studies, or be tailored to specific study groups or research areas, and these adaptations are possible.

The CEGs introduced in Chapter 5 were an established methodology introduced to a new field, namely case-control studies. The graphs had been developed in statistics and artificial intelligence, and shown to be suitable for use with cohort data, but had not before been used with case-control data. This application ensured that a reputable method was used, and the novelty was in the research area to which it was applied. The limitation in this chapter was not in the method but in the prior information used, since a clinical expert was not available, and hence population data were utilised as a substitute. While these population data may contain more information than

could be obtained from an expert, verification from a clinician experienced in childhood type I diabetes data would have been invaluable. However, in this example the findings were shown not to be sensitive to changes in the prior information. The application of the CEG method was a new means of reporting information regarding both the missing category in variables and the likely values of these missing data. CEGs could be used as an approach which incorporates prior information and the plausibility of paths in the event tree from clinical experts, and which are suitable to use alone or in combination with odds ratios.

The methods proposed in this thesis for use with case-control studies, one newly developed using population data and CEGs taken from another field but applied to a new area and extended, have different assumptions and hence between them will be suitable for a wide range of case-control studies. They also report different findings from the data and so can be used to answer different questions or can be selected depending on the research question being answered. Since one is graphical and the other numerical, they may appeal to different researchers or to particular applications, allowing flexibility in studies. CEGs are designed to be easy to communicate with specialists through the trees which later develop into a CEG. The R code used for the CEGs is available in Appendices D.2 and D.3 hence the work using CEGs is reproducible. In comparison, the numerical approach gives the method step-by-step and hence can also be reproduced. The method using population data is designed to not include complicated models or calculations and so should be accessible to a range of researchers. This allows others to adopt the methods in their work, or extend these approaches further without needing to replicate the work here. It also allows for any limitations in the method to be highlighted. The numerical approach using population data has the advantage over many current methods of being suitable when data are MNAR. Since it is difficult to distinguish between MAR and MNAR data from the sample alone, this approach may be a wise option when MNAR data are a possibility. CEGs can also be used to investigate the missingness mechanism and hence may provide guidance as to whether a method with a MAR assumption may be plausible, offering huge additional assistance to researchers.

8.3 Contributions to the Literature

The review of participation during the last nine years has been published[1] and provides an up-to-date, relevant summary of the factors which are thought to affect participation; including

participant characteristics and aspects of the study design. This information should provide guidance for those conducting studies which require participants, as to the data collection approaches which are successful and consequently how best to allocate their research budget. It can also be used to determine which areas or person characteristics to target, whether it be those who are most likely to participate, or those which are known to be difficult to recruit and hence where more resources may be required. However, care will be needed to not introduce bias through these approaches. This information should result in more successful participant recruitment since the findings from previous studies have been collated. The review conducted through the thesis agreed with the participant characteristics reported in the previous review,[17] but differed in study design, where technology is now more dominant during data collection.

The review of how authors approach the possibility of participation bias is also published[2] as it is intended to raise awareness of this bias, particularly in studies where authors often believe it to not be a problem, such as RCTs. The occurrence of participation bias and the actions taken by researchers needed to be reported to ensure it is recognised as a bias which not only exists, but which can be reduced.

The flowchart tool has been published[3] and is not only intended to encourage researchers to consider participation bias, but is also intended to provide examples and references for further reading, plus aid the selection of a suitable method for which all assumptions hold. The review in Chapter 3 demonstrated that participation bias is present in some studies and that the application of a suitable method to reduce this bias is not always included in the literature. Raised awareness of the need to consider participation bias, and guidance towards a suitable method to control for this bias, should result in more accurate results and valid conclusions from studies. This in turn should assist finding the true associations with a disease and ultimately benefit patient care.

Contribution to the literature of the successful use of case-control data with chain event graphs is one which may lead to this approach being adopted with future case-control data. The article[4] is intended to demonstrate the use of CEGs, as well as report the findings from the diabetes case-control data, particularly with respect to non-participation. The information obtained from CEGs includes how missing categories are associated with the outcome compared with recorded categories. In addition, the likely values of the missing categories can be reported and the missingness mechanism can be investigated. The use of CEGs with missing data has already been

demonstrated, and in Chapter 5 it was shown that this idea extended to missingness resulting from non-participation. Compared with other methods, CEGs do not require the data to be MAR, nor do they require the variable associated with participation to be recorded, data from non-participants to be available, or external population data. Therefore, CEGs eliminate all three requirements shown in §4.5 for the current methods used, but are a tool to investigate non-participation rather than a direct method to reduce participation bias.

CEGs were adapted in a second article[5] which included the investigation of non-participation directly, which has not before been achieved. The diabetes dataset did not contain specific details regarding non-participants, so hypothetical examples were presented to demonstrate the methodology. These adaptations should encourage the use of CEGs to investigate non-participation in case-control data and should contribute positively to the case-control literature.

The new method which proposed using population data in place of control data in a case-control study was also published[6] and offers an alternative approach to reducing participation bias in case-control studies, provided suitable population data are available. As this method has been shown to be unbiased even when the missingness causes data to be MNAR, it offers an advantage over many existing methods. It is also easy to implement and requires no specialist software. The increased availability of data and data-sharing is making this method more possible, and therefore the use of this method could increase over time. Reproducible examples have been provided to make this approach simple to follow and use, and the possibility of application by hand ensures it is convenient for any study budget.

Two published reviews, a published guidance tool to aid the selection of a method, and the publication of a newly developed methodology specifically for case-control studies, have been achieved through this thesis. In addition, a methodology recently developed in statistics and artificial intelligence has been introduced to case-control studies, and two papers prepared for the medical literature. This interdisciplinary approach has enabled case-control studies to use strengths from statistics and artificial intelligence to investigate non-participation in case-control studies, as demonstrated with the diabetes data and CEG adaptations.

8.4 Comparisons With the Literature

There was a comprehensive review of participation bias in the literature published in 2007,[17] but there has been no such review since then. The review presented in the thesis covered articles published from 2007–2015. This is presented in Chapter 2 to fill this void and inform researchers of the changes through time, and to provide up-to-date information regarding the characteristics of participants and the aspects of the study design found to be most successful. Therefore, while there was a similar review in the literature, this review provided more up-to-date information and highlighted changes in the last nine years which centered around technological advances.

The review of participation bias in Chapter 3 had a different structure to other reviews published on participation bias, since it aimed to raise awareness of this particular bias as well as highlight the approaches being taken by researchers. Since the top three impact-factor epidemiology journals were selected for this review, these aims should be achieved.

The flowchart tool presented in §4.5 was designed to summarise the methods available to reduce participation bias, as there was no evidence of such a tool in the literature. While each of the methods are published and there are applications of these approaches in studies, it may be unclear to researchers which of the array of methods may be suitable for their work. Therefore this guidance towards suitable methods, and references for further reading, should be a useful addition to the literature.

Chain event graphs were introduced in Chapter 5 and while there are several publications in statistical and artificial intelligence journals as detailed in §5.1.2, and there are medical examples,[8,365] none are published in the epidemiology literature. It has been shown that CEGs can be used with cohort studies,[365] but CEGs have not before been used with case-control studies as shown in §5.1.2. Therefore this is a new application for CEGs and a paper has been prepared[4] demonstrating CEGs with the childhood type I diabetes data, with particular interest to non-participation. Therefore this work should add to the literature for both case-control studies and type I diabetes.

The second achievement with CEGs was their adaptation for use with scenarios found in case-control studies, and the application of these graphs specifically for investigating non-participation. These adaptations formed a second CEG paper for case-control data[5] and have highlighted new

uses for CEGs as well as additional uses for case-control data.

The new method proposed[6] which uses population data in place of control data in case-control studies, should contribute positively to the case-control literature, both as a method to reduce participation bias in new studies, and as a way in which to conduct a sensitivity analysis. Population data have been used previously to conduct sensitivity analyses, but not in this format, nor as a replacement for controls. As this method is also unbiased when data are MNAR, this contributes a method which is suitable for more scenarios than most others.

8.5 Discussion of Graphical Models

Directed acyclic graphs (DAGs) and chain event graphs (CEGs) were used in this thesis to aid analysis. DAGs were primarily used for variable selection in regression models which were aimed to reduce participation bias in §4.3, whereas CEGs formed their own methodology for investigating non-participation in case-control studies, as shown in Chapters 5 and 6.

8.5.1 Directed Acyclic Graphs

DAGs are known to be useful for causal modelling and variable selection,[112] and these uses have been applied during the thesis, primarily in Chapter 4. More widely, DAGs have been used to determine when bias can or cannot be adjusted for, and there are publications of this use[443] as well as software for their implementation.[112] This use has also been shown in §2.3.4.1 where DAGs were used to assess the likely bias due to participation or selection, which is important for this thesis, but DAGs can also be used for the identification of other biases, such as confounding bias. DAGs may also have the potential for other uses in addition to those shown in the thesis. For example if a study is prospective, it may be that DAGs could be useful to show which variables should be recorded during data collection and whether techniques such as matching should be adopted to reduce confounding bias.

DAGs have the advantage of being easy to apply and are suitable for a range of studies, plus they can act as a useful tool between analysts and experts to list plausible causal associations between variables. Their disadvantage here is that DAGs alone do not offer a means by which to reduce

participation bias. Another limitation of DAGs is that they offer a generalised result which does not take into account specific factors, such as the odds ratio which is used in the analysis of a case-control study, hence may suggest bias where there is none. DAGs are generally more cautious than may be required, but any biases identified by the DAGs can be considered by the researcher. DAGs are non-parametric and are limited in that they cannot represent the nature of the variables or the associations between them. For example, they cannot show that a variable is normally distributed or that the association between two variables is linear. Therefore they are most useful as an aid rather than a definitive form of analysis.

8.5.2 Chain Event Graphs

CEGs have been used previously for the identification of variables associated with the outcome and for highlighting combinations of categories from different variables which lead to an increased or decreased association with the outcome.[21] However they do not produce values such as odds ratios from study data. In this thesis, CEGs have been used with variables which have missing data to report how the missing category performs in relation to the recorded categories.[365] They have also been used to draw conclusions about what the values in the missing category are most likely to be.[365]

CEGs have been used in the literature to analyse cohort studies,[8,375] investigate missingness,[365] for causality,[362,444–447] model selection,[361,448] plus learning and predicting.[363,379] A benefit of CEGs is their ability to include prior information as shown in Chapter 5. While the analysis is time-consuming, information regarding the association of the missing category of variables compared with the recorded categories, plus reporting of the likely values of missing data, ensures this additional time results in a useful investigation of non-participation. The thesis has also shown that CEGs may be used to investigate non-participation directly in studies and report the characteristics of participants or the recruitment techniques associated with participation.[5]

The use of graphical models in general appears to be increasing. There are advantages in medical studies where the assistance of clinical experts is of a great value and these diagrams offer another means of communication between statisticians and clinicians, to ensure sensible analyses are conducted. The use of DAGs to aid variable selection for methods to reduce bias and the use of CEGs to investigate missingness, are encouraged.

8.6 Discussion of Population Data

Population data were used in both Chapter 5 for prior knowledge since an expert was not available, and in Chapter 7 in place of control data. The limitations of population data have already been discussed in §7.5.1.2 and the advantages of these data include that it should cover the majority of the population, and suffer from at least different biases to control data, and in many instances, fewer biases.

Population data are collected through government surveys, for marketing, and through academic research. With data collection and data sharing both increasing, plus a move towards Big Data, the presence of population data should also increase and be available for use in medical studies. If accurate data are available, this could offer huge savings of time and resources for future medical research.

The use of population data and the sharing of data are encouraged, provided ethical and copyright restrictions are adhered to. Combining knowledge and data from multiple sources, and which have been collected for different reasons, could help to reduce biases and increase the chances of finding the true associations between variables.

8.7 Discussion of Methods to Investigate Participation Bias

There are at least eight methods available to investigate participation bias in case-control studies (six shown in Table 4.9, the CEGs in Chapters 5 and 6, and the approach using population data in Chapter 7) and the tool in §4.5 has been provided to help choose a suitable approach given a particular study.

Some of the methods available to detect or correct for participation bias do not take into account the nature of the OR, and hence may state that there is bias when participation depends only upon the outcome, but the OR can account for this.[38] Unfortunately very few methods are case-control study or OR specific. Methods are also often tailored towards selection bias rather than participation bias. While these biases are similar, they are not the same and non-participants may have different reasonings and hence patterns in their missingness compared with non-random

selection. Psychological or social reasons may influence non-participation, whereas non-random selection is a feature of the study design.

There is no one 'best' method to investigate participation bias in case-control studies. The choice of method will depend upon the structure of the data obtained, which variables have been recorded (such as whether the data includes the variables associated with participation or information about the non-participants) and what, if any, population data are available. The choice will also depend upon the outcome of interest, whether it be an odds ratio estimate, a report on the robustness of the findings, or an investigation into the missing data. Each of the methods has their own strengths and limitations and these should be considered before a method is selected. The assumptions of each method must also be adhered to, and these will also depend upon the study and external data. As stated previously, many of the methods assume the missing values are MAR but this is hard to verify. To ensure this assumption does not jeopardise the study results, an approach such as the one using population data may be preferable, since this assumption is not required.

Interdisciplinary work should be encouraged, to pool resources and raise awareness of similar non-participation problems in other areas. This includes ensuring that the same method is not developed independently in more than one field. While methods should be tailored to particular applications, they should also ideally be applicable to other studies and alternative areas of interest.

8.7.1 Comparison of Odds Ratio Results Across Methods

The diabetes data in Appendix A have been used throughout the thesis, and different methods to investigate or reduce bias resulting from non-participation have been applied. Comparisons can be made between the results obtained from these different approaches, which are shown in Table 8.1. The initially published results[7] are shown at the top of the table for comparison, and the remainder of the results were calculated during this thesis. Recall that the published and thesis unadjusted odds ratios differ slightly due to the discrepancies in the published and raw dataset as described in Appendix A.

Method	Caesarean OR (CI)	Amniocentesis OR (CI)	Finding
Published results	1.84 (1.09, 3.10)	3.85 (1.34, 11.04)	Both significant risk factors
No adjustment	1.74 (1.04, 2.89)	3.52 (1.52, 8.84)	Both significant risk factors
Stratification	1.65 (0.99, 2.77)	–	Caesarean not a significant risk factor (when stratifying by amniocentesis)
Sensitivity analysis	–	–	Unobserved variable would need to cause the case or control individuals to be 1.1 times and 1.6 times more likely to be exposed for the inference to change for caesarean and amniocentesis respectively
Regression adjustment	1.66 (0.99, 2.79)	3.35 (1.43, 8.44)	Caesarean not a significant risk factor (when adjusting for amniocentesis) Amniocentesis a significant risk factor (when adjusting for caesarean)
Chain event graphs	–	–	Amniocentesis acts as a greater risk factor than caesarean. Combination of both exposures greater increases associated risk
Population data approach	2.12 (1.53, 2.95)	3.38 (2.09, 5.47)	Both significant risk factors

Table 8.1: Results obtained from the range of analysis methods used. OR = odds ratio. CI = confidence interval.

Some assumptions were required to demonstrate some of the methods. For example, it was assumed that the caesarean variable should be stratified by the amniocentesis variable to demonstrate stratification. It was also assumed that the caesarean and amniocentesis variables needed to be adjusted for by one another in regression adjustment. These assumptions were made to demonstrate the application of the methods.

In the majority of the analyses, amniocentesis and caesarean delivery were shown to be significant risk factors for type I diabetes in the child. The exceptions were when caesarean was the exposure variable and stratified by amniocentesis, and when caesarean was the exposure variable and adjusted for my amniocentesis. In both instances there was not specific reasoning for the adjustments given, and they were instead used only to demonstrate the methods. In addition, the confidence intervals were both between 0.99 and almost 2.8, indicating borderline insignificance. All methods have indicated that having at least one amniocentesis is associated with a greater risk of diabetes than delivery by caesarean, with higher odds ratios reported for no adjustment, for regression adjustment, and using population data, plus more distinction in the CEG, and a greater required effect of an unobserved variable as shown by sensitivity analyses. CEGs also identified a further increased association with diabetes when having at least one amniocentesis and delivery by caesarean.

It has already been demonstrated in §4.5 that the methods included here require different assumptions or types of data. Provided these assumptions hold and the required data are available, the preferred method will depend upon the research question of interest and the structure of the data. Some methods will provide odds ratio estimates whereas others will not, and this is shown in Table 8.1. Many of these methods require the data to be MAR, which is a key advantage of the population data approach, since this requirement is not needed. In addition, no data regarding non-participants are required, nor data regarding the variable associated with participation. The population data approach may therefore suit a wide range of situations.

The width of the odds ratio estimates will differ depending on the adjustment chosen. Stratification can lead to wider confidence intervals since there are fewer individuals in each stratum, and the population data approach can lead to narrower confidence intervals since there are effectively more individuals in the dataset. However, the width of the confidence intervals in the population data approach may need to be amended as discussed in §7.5.3.

Since the diabetes data are real, the true odds ratios are unknown, hence the methods cannot be compared to determine which approach is most appropriate for these data. Simulations would instead be a sensible option for comparing methods. However, it is likely that different methods will suit different datasets as the assumptions and data structure vary.

8.8 Future Work

There are different directions that extensions to this thesis could take, relating to the literature reviews, guidance on choosing a method to reduce bias, the application and development of chain event graphs, or further development of the method utilising population data.

The literature reviews, while useful, may not be needed in the short-term, since the review conducted here of the individual characteristics most likely to be held by a participant and the study techniques used for recruitment which proved most successful, have been summarised for the last nine years. While recruitment has become more technologically focused, the individual characteristics have largely remained unchanged. Therefore, it may be some years before a similar review would uncover any substantial changes. The same may be true for the assessment which summarised the presence of participation bias in three high-impact epidemiology journals. While the results are likely to change with each issue and between journals, the general approach to participation bias is likely to remain unchanged in the short-term. A repeat of the review in a few years could be beneficial to determine whether attitudes towards participation bias have changed over time.

The guidance tool to aid the selection of a suitable method to reduce bias could be extended. Separate tools could be developed which are specific to certain research areas or to particular outcomes of interest. Tools could also be developed which include the assumptions of each of the tests and the structures of the study data which would be appropriate. The tool developed in §4.5 was intended to be a basic aid which could be amended and added to over time, and improvement of this tool could further encourage authors to take steps towards reducing participation bias. It may even be possible to transform this flowchart tool into a software tool, where users answer a few basic questions through an online questionnaire, and are guided towards appropriate methods. The tool could also prompt the user to check the assumptions for the chosen method and if possible,

could conduct the analysis if the data are provided.

Since CEGs are new to the medical literature and this is the first time they have been applied to case-control data, there is scope for further development. The analysis of historical case-control data, particularly relating to non-participation, could be conducted to further utilise data from past studies as shown with the diabetes data. Study recruitment and data collection in studies can be very time-consuming, so the full analysis of collected data should be encouraged, and as much information obtained from them as possible. CEGs could also be applied to future case-control studies, particularly to draw conclusions regarding non-participation from the adapted graphs. From one given study it could be possible to draw conclusions about the variable associations with the outcome, the associations of variable category combinations with the outcome, the likely mechanism for missing data, the likely values of missing data, how the missing data are associated with the outcome compared with the recorded data, which types of individuals were associated with participation, the recruitment approaches which were successful, and to also draw comparisons with previous studies. The additional benefits of incorporating prior knowledge and the increased ease of communicating with the clinical expert, should also result in more reliable conclusions. The application of CEGs could therefore increase in the medical literature.

The approach suggested in Chapter 7 which uses population data in place of the control group could be further developed. It has already been discussed in §7.5.2 how the method could be extended, and in §7.5.3 how the confidence intervals could be adapted to account for the possible limitations in the population data. The extensions in §7.5.2 could be detailed and demonstrated with examples to show how they could be applied, and the adjustments to the confidence intervals should be incorporated as part of the method. These developments would further promote the method and increase its suitability for estimating ORs.

To complement the work in the thesis, R packages (or similar) could be developed to assist with the analyses proposed. For example, one package could have commands to run the new method described in Chapter 7 when given the three requirements for each estimate; the exposure in the population, the size of the population and the number of cases in the population. A second package could be formed with contains the AHC algorithm for returning stages in the development of a CEG when given the data, any priors and the equivalent sample size. It could also have a second command which forms positions from these stages, and ideally would have functions to plot the

event tree and CEG. Currently the formation of positions and the development of the CEG are done by hand, which can be time-consuming when there are a large number of variables or variable categories. In addition, the event trees and CEGs are currently drawn in LaTeX which results in neater figures than other software, but can also be time-consuming.

All examples given in the thesis have been retrospective, as the focus has been case-control data. Therefore, participation has been affected by factors such as the exposure or outcome of interest. In prospective studies, participation may *influence* other variables and hence different causal structures may be required.[38] While not covered here, many of the ideas will be the same, but further work could draw similarities and differences between the analyses of these two study types.

While briefly touched on in §7.5.2, further work could be done to accommodate matched case-control data. This may be applicable to future case-control work, but where the approaches described in this thesis are applied to historical case-control data, there may be several which have used a matched design and the implications of this could be investigated, and any changes to the methods needed could be summarised.

8.9 Summary

This thesis aimed to investigate solutions to reduce participation bias in case-control studies. The thesis summarised background literature for both case-control studies and participation bias, to introduce both topics and draw links between them. There was little information available since 2007 for participation bias, so a new review was conducted which summarised the characteristics of participants and the recruitment techniques which were found to be successful. An assessment was then conducted, which took the three highest impact-factor epidemiology journals at the time, to summarise whether participation was possible in typical publications, and what actions were taken by researchers for this. This aimed to give a snapshot of the current awareness of and attitudes towards participation bias, to estimate the scale of the problem.

The methods currently used to investigate participation bias were described next and critically evaluated, with both hypothetical examples plus a diabetes dataset applied. A flowchart was created which aimed to guide users towards methods which may be suitable for their study, given

particular data sources. The diabetes data were then used throughout the thesis as a common theme and as a means by which to compare approaches. The diabetes data were from a real case-control study conducted in Yorkshire and hence required ethical approval, and problems such as data cleaning were encountered, as common with real datasets.

Next, chain event graphs were introduced to the case-control literature, where they had not before been applied. The diabetes data were successfully used with the graphs, particularly to investigate missingness. To address non-participation and to make the graphs more applicable to case-control studies, seven adaptations to the graphs were suggested specifically for scenarios encountered with case-control data. The diabetes data did not have all the required information recorded and so hypothetical examples were used to illustrate these adaptations. This also highlighted the information which would need to be recorded to use these adaptations successfully. The chain event graphs were used to explore the missingness mechanism of any missing data, and used to estimate not only whether these values were MAR or not, but also what values these missing values were likely to have taken. The missing category of a variable was also used as an informative category, so no data were wasted. If data were shown to be MNAR, the previous research on the methods included in the earlier flowchart tool had identified that there were very few options. Therefore, the next step in the thesis was to develop a new method which used population data in place of the control group and hence proposed a method which was suitable when data were MNAR.

A researcher new to participation bias in case-control studies could use the review in §2.3.5 to understand factors associated with participation, and DAGs to decide whether participation bias is a possibility. The article in Chapter 3 could be used to see how non-participation has been approached in the literature, and the CEGs in Chapters 5 and 6 could be used to investigate non-participation and whether the data are likely to be MAR. The flowchart in §4.5 could assist with the selection of a tool to reduce bias, and if the data are found to be MNAR, the population data method proposed in Chapter 7 could be implemented.

These two participation reviews, the flowchart tool for method selection, the development of a new method, and the successful application of an existing approach for use with case-control data, plus the adaptation of this method for more specific uses with case-control data, have formed the investigation into solutions to reduce participation bias in case-control studies as intended.

Appendices

A Diabetes Dataset Details

The diabetes dataset was taken from a case-control study,[7] which had recorded cases of children under 16 years old diagnosed with insulin-dependent diabetes mellitus (IDDM), or type 1 diabetes, while resident in the area of the former Yorkshire Regional Health Authority, since 1978, with data collected 1993–1994. The dataset consisted of 196 matched cases and 325 controls (129 matched triplets and 67 matched pairs) after exclusions; 13 ineligible, 15 refusals and 35 unmatched (6 cases and 29 controls), with cases selected from the Yorkshire Childhood Diabetes Register, and age and sex matched controls recruited using The Family Health Service Authority through general practitioner contact.

A.1 Ethical Approval

The data were provided by the principal investigator to investigate whether the published results might have been affected by participation bias. To obtain this information, ethical approval was sought through the University of Leeds Research Ethics Committee; see Appendix B.

A.2 Exploratory Analysis

Exploratory analysis was conducted; for understanding, and to ensure the data were complete and not corrupted. The study outcome was type I diabetes and the study exposures included caesarean delivery and amniocenteses during pregnancy, chosen as previous analyses concluded these variables to give significantly raised odds ratios in univariable analyses.[7,449]

A.2.1 The Raw Data

Table 3 in the original study[7] showed delivery to contain three categories; normal, assisted and caesarean. Unfortunately the values in the dataset did not correspond to the published numbers for normal or assisted births. However, when delivery was dichotomised into caesarean or not, the values matched, hence two categories were used. The amniocentesis values also differed slightly, see §A.3.2.

Summaries of the data are shown in Tables A.1 and A.2, and Figure A.1, showing the exposures are rare, and that higher proportions of cases are found amongst those who were delivered by caesarean and whose mothers underwent amniocenteses. This is expected for an exposure shown to be a risk factor for type I diabetes in the original study.[7]

Exposure	With	Without
Caesarean	69	452
Amniocentesis	24	497

Table A.1: Diabetes data: Number of mothers with each exposure of interest.

Exposure	Caesarean (No)	Caesarean (Yes)
Amniocentesis (No)	434 (152 cases)	63 (28 cases)
Amniocentesis (Yes)	18 (10 cases)	6 (6 cases)

Table A.2: Diabetes data: Caesarean and amniocentesis numbers, with number of cases.

A.3 Reproducing the Original Results

The original results were replicated to ensure a fair comparison between the methods used in the thesis. The relevant variables were read into R^{303} and univariable analyses performed as in Table 3 of the original study.[7]

A.3.1 Caesarean

Table A.3 shows that $\frac{34}{196} \approx 0.17$ cases and $\frac{35}{325} \approx 0.11$ controls were delivered by caesarean, hence agreement of increased odds as concluded in the original study,[7] despite the new binary

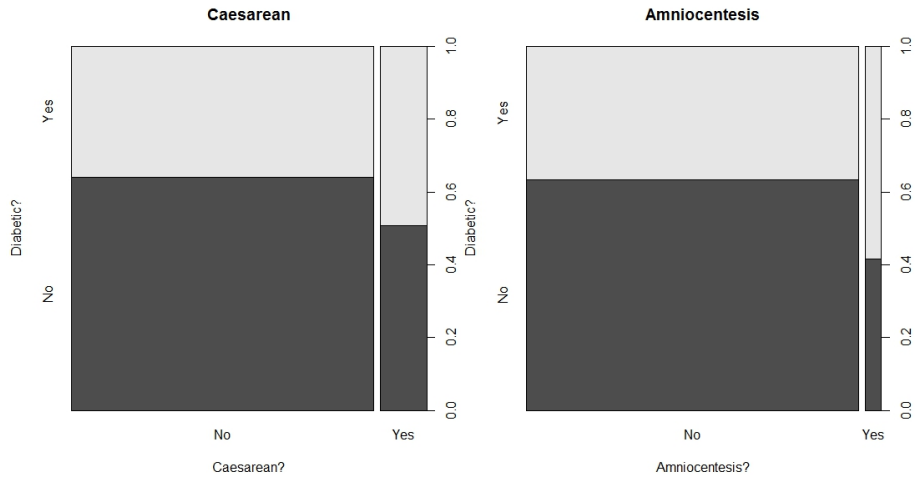

Figure A.1: Diabetes data: The cases and controls in each exposure category of interest.

category allocation. The odds ratio can be calculated as $\frac{290\times34}{162\times35} = 1.74$ (2dp), with delivery not by caesarean as the reference category. Alternatively logistic regression can be used, with the results in Table A.4 shown as 1.74 (1.04, 2.89). The original study odds ratio was 1.84 (1.09, 3.10) when the reference category was normal birth. However since assisted birth was also an option, the odds ratios are expected to differ.

Delivered by Caesarean?	No	Yes
Controls	290	35
Cases	162	34

Table A.3: Cases and controls in the diabetes data: Caesarean delivery.

Caesarean	Estimate	Lower 95% CI	Upper 95% CI
(Intercept)	0.56	0.46	0.68
Caesarean (Yes)	1.74	1.04	2.89

Table A.4: Odds ratios for the diabetes data calculated using logistic regression: Caesarean.

A.3.2 Amniocentesis

Table A.5 shows that 14 cases and 10 controls has mothers who underwent amniocentesis, whereas the published results show 13 cases and 6 controls as shown in Table 3.[7] However, Table 4 in the

original study[7] lists 13 controls and 9 cases in detail, which is much closer to the numbers found in Table A.5, possibly suggesting minor disagreement within the article. Table A.5 shows $\frac{10}{325} \approx 0.03$ controls and $\frac{14}{196} \approx 0.07$ cases had mothers who underwent amniocentesis during pregnancy, explaining the increased odds found in the original article.[7] The odds ratio can be calculated as $\frac{315 \times 14}{182 \times 10} = 2.42$ (2dp), with no amniocentesis as the reference category. Again logistic regression can be used as shown in Table A.6, with an estimate of 2.42 (1.05, 5.57), differing from the original study estimate of 3.85 (1.34,11.04), which may be due to the disagreement within the article. Whether the raw data or the published data contained the correct values, both showed an increased risk of diabetes for those born by caesarean or with mothers who underwent amniocentesis. and the differences between the datasets were relatively small.

Amniocentesis?	No	Yes
Controls	315	10
Cases	182	14

Table A.5: Cases and controls in the diabetes data: Amniocentesis.

Amniocentesis	Estimate	Lower 95% CI	Upper 95% CI
(Intercept)	0.58	0.48	0.69
Amniocentesis (Yes)	2.42	1.05	5.57

Table A.6: Odds ratios for the diabetes data calculated using logistic regression: Amniocentesis.

B Ethical Approval Paperwork: Diabetes Case-Control Study

Faculty of Medicine and Health
Research Office

Room 10.110, Level 10
Worsley Building
Clarendon Way
Leeds LS2 9NL

T (General Enquiries) +44 (0) 113 343 4361
F +44 (0) 113 343 4373

UNIVERSITY OF LEEDS

Miss Claire Keeble
PhD Student
Division of Epidemiology and Biostatistics
School of Medicine, LIGHT
8.001 Worsley Building
University of Leeds
Leeds LS2 9JT

27 September 2012

Dear Claire

Re: **HSLTLM/12/008**

Title: **Investigating Solutions to Minimise Participation Bias in Case-Control Studies**

I am pleased to inform you that the above research application has been reviewed by Darren Shickle, Acting Chair of HSLTLM and I can confirm a conditional favourable ethical opinion on the basis described in the application form as submitted at date of this letter.

This project is conditionally approved on the grounds that the protections referred to in the proposal are complied with:

1. The consent forms continue to be retained at least for the duration of the PhD and for as long as deemed appropriate by the data controller
2. A confidentiality agreement is signed, as required by the data controller
3. The data released by the data controller is stored on a secure server
4. No names of patients or parents are released

You should also delete address data as soon as practicable after analysis or at least reduce to partial postcodes to limit scope for identifying participants.

Please notify the committee if you intend to make any amendments to the original research as submitted at date of this approval. This includes recruitment methodology and all changes must be ethically approved prior to implementation. Please contact the Faculty Research Ethics and Governance Administrator for further information FMHUniEthics@leeds.ac.uk

Ethical approval does not infer you have the right of access to any member of staff or student or documents and the premises of the University of Leeds. Nor does it imply any right of access to the premises of any other organisation, including clinical areas. The committee takes no responsibility for you gaining access to staff, students and/or premises prior to, during or following your research activities.

Please note: You are expected to keep a record of all your approved documentation, as well as documents such as sample consent forms, and other documents relating to the study. This should be kept in your study file, which should be readily available for audit purposes. You will be given a two week notice period if your project is to be audited.

It is our policy to remind everyone that it is your responsibility to comply with Health and Safety, Data Protection and any other legal and/or professional guidelines there may be.

I wish you every success with the project.

Yours sincerely

Professor Darren Shickle
Acting Chair, LIHS/LIGHT/LIMM Joint REC
University of Leeds

Figure B.2: Diabetes data: Ethical approval letter.

C Breakdown of the Epidemiology Articles Used in Chapter 3

Table A.7: Breakdown of the 81 articles used in Chapter 3, with the table columns ordered. (Epi. = Epidemiology, AJE = American Journal of Epidemiology, IJE= International Journal of Epidemiology).

Study ID	Journal	Data Source	Method	Method 2	Participation Bias Term	Category
1	AJE	Cohort	-	-	No	A
2	AJE	Cohort	-	-	No	D
3	AJE	Cohort	-	Adjust for variables	No	I
4	AJE	Cohort	-	Adjust for variables	No	I
5	AJE	Cohort	-	Adjust for variables	No	I
6	AJE	-	-	Adjust for variables	No	I
7	AJE	Case-control	-	Adjust for variables	No	I
8	AJE	-	-	-	No	NA
9	AJE	-	-	-	No	NA
10	AJE	-	-	-	No	NA
11	AJE	-	-	-	No	NA
12	AJE	-	-	-	No	NA
13	AJE	-	-	-	No	NA
14	AJE	Cohort	Sensitivity Analysis	-	No	R
15	AJE	Cohort	Sensitivity Analysis	-	No	R
16	AJE	Case-control	Stratification	-	No	R
17	Epi	Database	-	-	No	D
18	Epi	Cohort	-	-	No	I
19	Epi	Cohort	-	Adjust for variables	No	I
20	Epi	Cohort	-	Adjust for variables	No	I
21	Epi	Cohort	-	Sensitivity Analysis	No	I
22	Epi	-	-	-	No	NA
23	Epi	Database	-	-	No	NA
24	Epi	-	-	-	No	NA
25	Epi	Cohort	-	-	No	NA
26	Epi	-	-	-	No	NA
27	Epi	-	-	-	No	NA
28	Epi	-	-	-	No	NA
29	Epi	-	-	-	No	NA
30	Epi	-	-	-	No	NA
31	Epi	-	-	-	No	NA
32	Epi	-	-	-	No	NA
33	Epi	-	-	-	No	NA
34	Epi	-	-	-	No	NA
35	Epi	-	-	-	No	NA
36	Epi	-	-	-	No	NA
37	Epi	-	-	-	No	NA
38	Epi	-	-	-	No	NA
39	Epi	-	-	-	No	NA
40	Epi	Case-control	Adjust for variables	-	No	R
41	Epi	-	Sensitivity Analysis	Adjust for variables	No	R
42	Epi	Cohort	Sensitivity Analysis	Adjust for variables	No	R
43	Epi	Database	Stratification	-	No	R
44	IJE	-	-	-	No	A
45	IJE	Cohort	-	-	No	A
46	IJE	Cohort	-	-	No	A
47	IJE	Cohort	-	-	No	A
48	IJE	Database	-	Adjust for variables	No	A
49	IJE	Cohort	-	Adjust for variables	No	A
50	IJE	-	-	-	No	D
51	IJE	-	-	-	No	D
52	IJE	Database	-	Adjust for variables	No	D

Breakdown of the 81 articles used in Chapter 3, with the table columns ordered.

Table A.7 – Continued.

Study ID	Journal	Data Source	Method	Method 2	Participation Bias Term	Category
53	IJE	-	-	Sensitivity Analysis	No	D
54	IJE	-	Sensitivity Analysis	-	No	D
55	IJE	Cohort	-	-	No	I
56	IJE	Database	-	-	No	I
57	IJE	Cohort	-	Adjust for variables	No	I
58	IJE	Cohort	-	Adjust for variables	No	I
59	IJE	Cohort	-	Adjust for variables	No	I
60	IJE	Cohort	-	Sensitivity Analysis	No	I
61	IJE	-	-	-	No	M
62	IJE	Case-control	-	Adjust for variables	No	M
63	IJE	-	-	-	No	NA
64	IJE	-	-	-	No	NA
65	IJE	-	-	-	No	NA
66	IJE	-	-	-	No	NA
67	IJE	Cohort	-	-	No	NA
68	IJE	Database	-	-	No	NA
69	IJE	-	-	-	No	NA
70	IJE	-	-	-	No	NA
71	IJE	-	-	-	No	NA
72	IJE	-	-	-	No	NA
73	IJE	-	-	-	No	NA
74	IJE	-	-	-	No	NA
75	IJE	-	-	-	No	NA
76	IJE	-	-	-	No	NA
77	IJE	-	-	-	No	NA
78	IJE	Cohort	Sensitivity Analysis	-	No	R
79	IJE	Cohort	Sensitivity Analysis	Adjust for variables	Yes	R
80	IJE	Cohort	Sensitivity Analysis	Adjust for variables	No	R
81	IJE	Cross-sectional	Sensitivity Analysis	Adjust for variables	No	R

The columns, from left to right, are as follows:

1. Study identifier, listing the 81 studies.
2. The journal from which the article was taken.
3. The source of the data in the article.
4. Which method was applied to account for the non-participation.
5. A second method which was applied, which may have been suitable to account for the non-participation, but which was not stated as being implemented for this purpose.
6. Whether the term “participation bias” was used in the article.
7. The category to which the article was ultimately assigned.

D Chain Event Graph Supporting Material

D.1 Chain Event Graph Literature Review

D.1.1 Web of Science

The topic search term used in Web of Science[130] was `"chain event graph*"` and the results follow. None of the nine articles included case-control studies.

1. ˆThe introduction of a subclass for CEGs and an algorithm for the selection of a CEG, with application to a cohort study.[447]
2. ˆInformed missingness and CEGs, plus the application to a cohort study.[365]
3. The application of CEGs to medical data in the form of a cohort study.[364]
4. ˆCEGs as an alternative to the causal Bayesian network.[362]
5. A dynamic programming algorithm for learning CEGs.[363]
6. ˆ Bayesian maximum a posteriori (MAP) model selection of CEGs, i.e. the value which maximises the probability mass function given the data.[361]
7. ˆCEGs for causal analysis.[389]
8. ˆScoring for model selection of CEGs.[448]
9. ˆThe seminal paper introducing CEGs.[21]

D.1.2 PubMed

Two PubMed[376] searches were conducted. The first contained the two phrases `"chain event graph*"` and `"case-control"` in `All Fields`, with no results returned. The second searched simply `"chain event graph*"` in the `Title`, with again no results returned.

D.1.3 Scopus

Two Scopus[377] searches were conducted, using `title-abs-key`. The first used the search terms `"chain event graph*" AND "case-control"` but returned no results. The second used only `"chain event graph*"` and returned 12 articles. Nine articles were

already returned using Web of Science §D.1.1 and the remaining three are given below, all of which were conference proceedings. None contained any work using case-control studies.

1. An article,[450] which is in fact the same as a previous article[363] as shown in the Web of Science output, just re-listed slightly differently.
2. A theoretical paper introducing the transported CEG which is a subgraph of the CEG.[451]
3. The evaluation of causal effects using CEGs.[444]

D.1.4 Google Scholar

A Google Scholar[378] search was conducted using the term `"chain event graph*"` and 37 results were returned as follow (the remaining seven are shown in the Web of Science results with a superscript ^). Here there were some repeats where work was uploaded as a technical report or presentation, and later as a published article. Again, none used case-control studies with CEGs.

1. A theoretical paper for the identification of the conditional independence structure of models from the topology of the graph.[452]
2. A theoretical paper for staged trees.[453]
3. APFAs as mentioned in §5.1.1.[374]
4. CEGs for decision analysis.[454]
5. Potential fellowship for CEGs to explore drop-out in weight loss studies.[455]
6. Not CEGs, but contains an abstract from a conference where CEGs were presented.[456]
7. Theoretical paper for Bayesian decision theory.[457]
8. An article[458] with the same content as another search result.[374]
9. The introduction of the dynamic CEG.[375]
10. A slideshow introducing CEGs.[459]
11. An article with the same content as another search result.[451]
12. The analysis of ecosystem services which includes graphical models.[460]
13. The introduction of the dynamic staged tree for modelling discrete-valued discrete-time multivariate processes.[461]
14. The same content as another search result,[461] but from the University repository.
15. A paper on algebraic discrete causal models.[462]

16. A PhD thesis entitled "Learning and predicting with chain event graphs"[379] where case-control studies were not used.
17. Technical report on causal analysis with CEGs.[446]
18. A chapter from a book about causal probability trees.[463]
19. Conference abstract[448] with the same content as an article.[361]
20. A summary of the basic ideas in algebraic statistics with a brief reference to CEGs.[464]
21. PhD thesis for the theory and application of CEGs,[465] but not using case-control studies.
22. Causal inference PhD thesis,[466] again without case-control data.
23. Technical report on causality.[467]
24. An introduction to CEGs for causal analysis.[445]
25. An introduction to CEGs, demonstrated with an E. coli example.[468]
26. Technical report[469] with the same content as another search result.[468]
27. An introduction to the causal manipulation of CEGs.[470]
28. Probabilistic decision graphs for inference, but not using CEGs.[471]
29. Software report,[472] published before CEGs were published.

D.2 R Code for the Bayesian Agglomerative Hierarchical Clustering Algorithm[8]

```
CEG.AHC<-function(exampledata=exampledata ,equivsize=equivsize){
exampledata<-exampledata
equivsize<-equivsize
numbvariables<-dim(exampledata)[2]
numbcat <-c()
for(k in 1:numbvariables){
numbcat <-c(numbcat ,nlevels(exampledata[,k]))
}
numb<-c(1)
for(i in 2:numbvariables){
numb<-c(numb ,prod(numbcat [1:(i-1)]))
}
prior<-c()
for(i in 1:numbvariables){
for(j in 1:numb[i]){
prior<-c(prior ,list(rbind(rep(equivsize/(numbcat[i]*numb[i]),numbcat[i]))))
}
```

```
}
data<-c(list(rbind(table(exampledata[,1]))))
for (i in 2:numbvariables){
for (j in 1:numb[i]){
data<-c(data ,list(rbind(ftable(exampledata[,1:i])[j,])))
}
}
comparisonset <-c()
for (i in 2:numbvariables){
comparisonset <-c(comparisonset ,list(c((sum(numb[1:(i-1)])+1):(sum(numb[1:i]))
)))
}
labelling <-c()
for (k in 1:(numbvariables -1)){
label <-c(1,rep("NA",sum(numb[1:k]) -1))
label<-c(label ,rep(levels(exampledata[,k]),numb[k]))
if (k<(numbvariables -1)){
for (i in (k+1):(numbvariables -1)){
label<-c(label ,rep(levels(exampledata[,k]),each=numb[i+1]/numb[k+1],numb[k+1]
/numbcat[k]))
}
}
labelling<-cbind(labelling ,label)
}
mergedlist <-c()
for (i in 1:sum(numb)){
mergedlist<-c(mergedlist ,list(labelling[i,]))
}
merged1 <-c()
lik<-0
for( i in 1: sum(numb)){
alpha<-unlist(prior[i])
N<-unlist(data[i])
lik<-lik+sum(lgamma(alpha+N)-lgamma(alpha))+sum(lgamma(sum(alpha))-lgamma(
sum(alpha+N)))
}
score<-c(lik)
diff.end<-1
```

```
while(diff.end >0){
difference<-0
for (k in 1:length(comparisonset)){
if(length(comparisonset[[k]]) >1){
for (i in 1:( length(comparisonset[[k]]) -1)){
for (j in (i+1):length(comparisonset[[k]])){
compare1 <-comparisonset[[k]][i]
compare2 <-comparisonset[[k]][j]
result<-lgamma(sum(prior[[compare1]]+prior[[compare2]]))-lgamma(sum(prior[[
compare1]]+data[[compare1]]+prior[[compare2]]+data[[compare2]]))+
sum(lgamma(prior[[compare1]]+data[[compare1]]+prior[[compare2]]+data[[
compare2]]))-sum(lgamma(prior[[compare1]]+prior[[compare2]]))-
(lgamma(sum(prior[[compare1]]))-lgamma(sum(prior[[compare1]]+data[[compare1
]]))+sum(lgamma(prior[[compare1]]+data[[compare1]]))-
sum(lgamma(prior[[compare1]]))+lgamma(sum(prior[[compare2]]))-lgamma(sum(
prior[[compare2]]+data[[compare2]]))+
sum(lgamma(prior[[compare2]]+data[[compare2]]))-sum(lgamma(prior[[compare2]])
) )
if (result > difference){
difference<-result
merged<-c(compare1 ,compare2 ,k)
}
}
}
}
}
diff.end<-difference
if(diff.end >0){
prior[[merged [1]]]<-prior[[merged [1]]]+ prior[[merged [2]]]
prior[[ merged [2]]] <-cbind(NA ,NA)
data[[merged [1]]]<-data[[merged [1]]]+data[[merged [2]]]
data[[ merged [2]]] <-cbind(NA,NA)
comparisonset[[merged [3]]]<-comparisonset[[merged[3]]][-(which(comparisonset
[[merged [3]]]== merged[2]))]
mergedlist[[merged [1]]]<-cbind(mergedlist[[merged[1]]],mergedlist[[merged
[2]]])
mergedlist [[ merged [2]]] <-cbind(NA ,NA)
lik<-lik+diff.end
```

```
score<-c(score ,lik)
merged1<-cbind(merged1 ,merged)
}
}
stages<-c(1)
for (i in 2:numbvariables){
stages<-c(stages ,comparisonset[[i-1]])
}
result<-mergedlist[stages]
newlist <-list(prior=prior ,data=data ,stages=stages ,result=result ,score=score ,
merged=merged1 ,comparisonset=comparisonset ,mergedlist=mergedlist ,lik=lik)
return(newlist)
}
```

D.3 Adapted Bayesian Agglomerative Hierarchical Clustering Code for Use With Non-Uniform Priors

```
CEG.AHC.priors<-function(exampledata=exampledata ,equivsize=equivsize,prior=prior){
exampledata<-exampledata
equivsize<-equivsize
prior<-prior
numbvariables<-dim(exampledata)[2]
numbcat <-c()
for(k in 1:numbvariables){
numbcat <-c(numbcat ,nlevels(exampledata[,k]))
}
numb<-c(1)
for(i in 2:numbvariables){
numb<-c(numb ,prod(numbcat [1:(i-1)]))
}
data<-c(list(rbind(table(exampledata[,1]))))
```

⋮

The remainder of the code is the same as in Appendix D.2.

D.4 Three Variables; Amniocentesis, Caesarean and Diabetes Status

To form the tree, the three variables need to be ordered accordingly; amniocentesis before birth, caesarean delivery during birth and the disease status detected after birth. The resulting tree is shown in Figure D.3, labelled with the number of individuals along each edge, with no individuals following the path of amniocentesis (yes) $\rightarrow$ caesarean (yes) $\rightarrow$ diabetes (no). This may be a coincidence in the study group, or there may be reasoning for this; hypotheses can be generated and tested for this combination of exposure variables, with diabetes as the outcome.

With the tree as a basis, the data were analysed using the AHC algorithm in §5.2.2.1 implemented in R.[380] The output for the algorithm is shown in Appendix D.5 for reference, and returned

$$u_0 = \{s_0\}, u_1 = \{s_1, s_2\}, u_2 = \{s_3\}, u_3 = \{s_4, s_5, s_6\},$$

where u_i are stages and s_j are situations. The situations merged at each iteration can be seen in Table A.8 along with the resulting score, where the algorithm selects the CEG with the maximum score. The corresponding staged tree is given in Figure D.4.

Iteration	Situations	Score
0	-	-655.3
1	$\{s_4, s_6\}$	-653.1
2	$\{s_5, s_6\}$	-652.0
3	$\{s_1, s_2\}$	-651.4

Table A.8: Output from the agglomerative hierarchical clustering algorithm: Three variables.

Using the stages in Figure D.4 the positions, w_k, can be listed as

$$w_0 = \{s_0\}, w_1 = \{s_1\}, w_2 = \{s_2\}, w_3 = \{s_3\}, w_4 = \{s_4, s_5, s_6\},$$

since only situations s_4, s_5, s_6 have the same coloured subtrees. Figure D.5 shows the ordinal CEG resulting from collapsing Figure D.4 over its positions. Vertices s_4, s_5 and s_6 collapse to form position w_4, with vertices s_0, s_1, s_2, s_3 forming positions w_0, w_1, w_2, w_3 respectively. Vertices s_1 and s_2 corresponding to positions w_1 and w_2 are in the same stage, hence there is a dashed line between them and their corresponding edges are assigned the same colours (as in Figure D.4). The path from w_6 to w_∞ denoting controls is given a dotted line since the path is possible but not populated.

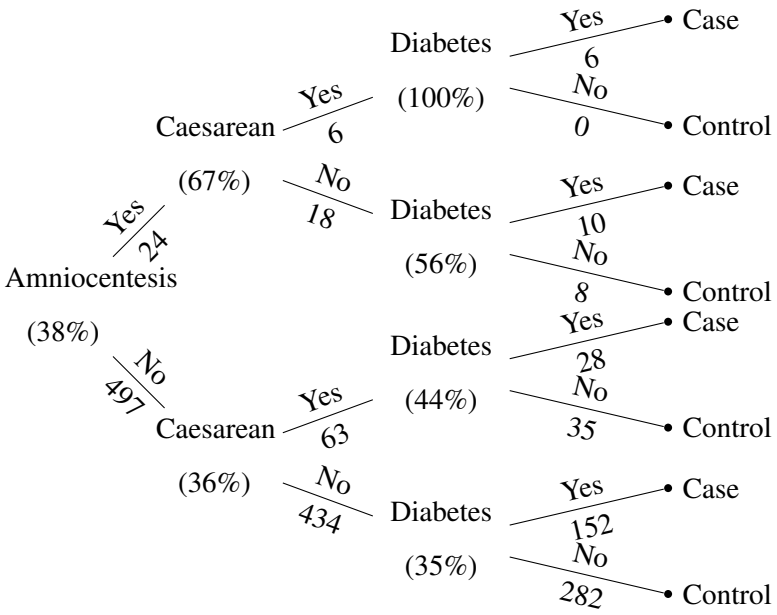

Figure D.3: Tree for the diabetes dataset, three variables, no missing data.

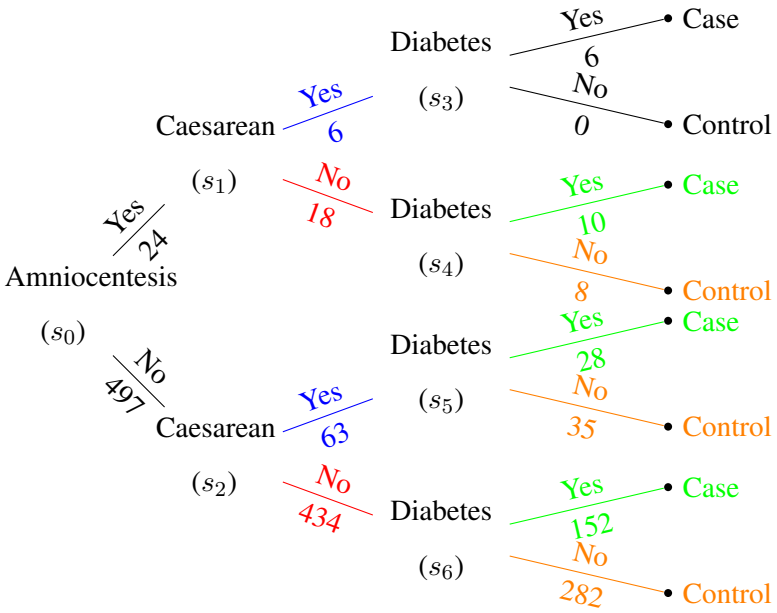

Figure D.4: Staged tree for the diabetes dataset, three variables, no missing data.

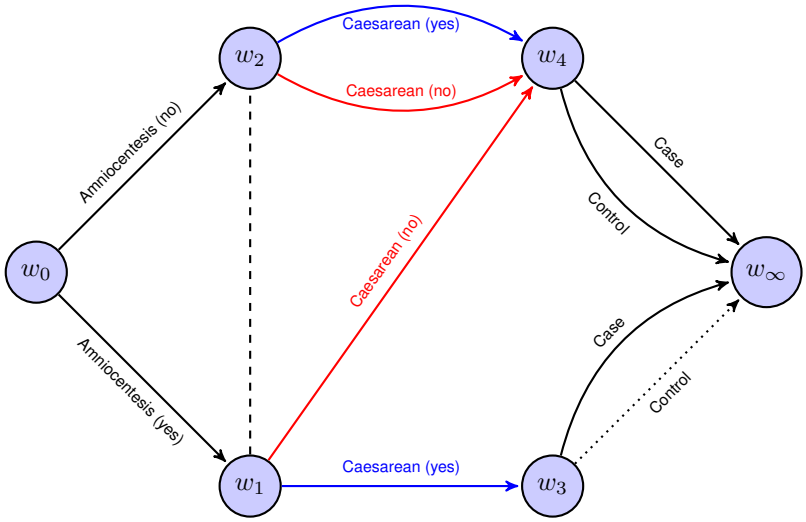

Figure D.5: Ordinal chain event graph for the diabetes dataset, three variables, no missing data.

In Figure D.5 the combination of at least one amniocentesis during pregnancy followed by delivery by caesarean is associated with the greatest proportion of cases (w_3) and if there is no amniocentesis, the diabetes status is independent of whether or not the child was delivered by caesarean (w_4). At least one amniocentesis followed by non-caesarean delivery, leads to the same vertex as when there is no amniocentesis (w_4). The population distribution of caesarean is indistinguishable for amniocentesis yes/no, since w_1 and w_2 are in the same stage. All of the children whose mother had at least one amniocentesis and delivered by caesarean, were cases. However, since this was only six children, this conclusion should be treated with caution.

D.5 Example of the AHC Algorithm Output From §D.4

```
$prior #Uniform Priors
$prior[[1]] #Amniocentesis priors
     [,1] [,2]
[1,]    1    1

$prior[[2]]        #Caesarean priors
     [,1] [,2]
[1,]    1    1

$prior[[3]]
     [,1] [,2]
[1,]   NA   NA

$prior[[4]]        #Diabetes priors
     [,1] [,2]
[1,] 0.75 0.75

$prior[[5]]
     [,1] [,2]
[1,]   NA   NA

$prior[[6]]        #Diabetes priors
     [,1] [,2]
[1,] 0.25 0.25

$prior[[7]]
     [,1] [,2]
[1,]   NA   NA


$data
$data[[1]] #Amniocentesis data
      no yes
[1,] 497  24

$data[[2]] #Caesarean data
     [,1] [,2]
[1,]   69  452

$data[[3]]
     [,1] [,2]
[1,]   NA   NA

$data[[4]] #Diabetes data
     [,1] [,2]
[1,]  190  325

$data[[5]]
     [,1] [,2]
[1,]   NA   NA

$data[[6]] #Diabetes data
     [,1] [,2]
[1,]    6    0

$data[[7]]
     [,1] [,2]
[1,]   NA   NA


$stages        #The four stages returned by
[1] 1 2 4 6        #the algorithm

$result
#The paths in each of the stages
$result[[1]] #Path for first stage
label label #(Amniocentesis stage)
  "1"   "1"

$result[[2]] #Path for second stage
      [,1] [,2]
label "no" "yes"
label "NA" "NA"
#(Two caesarean vertices are
#in the same stage)

$result[[3]
      [,1]   [,2]      [,3]
label "no"   "yes"     "no"
label "csec" "notcsec" "notcsec"
#Path for third stage -
#leading to three of the
#four diabetes vertices
```

```
$result[[4]]
 label  label
 "yes" "csec"
#Path for fourth stage
#Leading to the diabetes vertex with
#amniocentesis and caesarean

$score
[1] -655.2726 -653.0582 -652.0007 -651.3952
#The scores for the iterations of the algorithm

$merged        #Which vertices were merged
     merged merged merged
[1,]      4      4      2
[2,]      7      5      3
[3,]      2      2      1
#Read as columns not rows
#Merged vertices 4 and 7 to 2nd non-trivial stage
#Merged vertices 4 and 5 to 2nd non-trivial stage
#Merged vertices 2 and 3 to 1st non-trivial stage

$comparisonset
$comparisonset[[1]]
[1] 2

$comparisonset[[2]]
[1] 4 6


$mergedlist
$mergedlist[[1]]
label label
  "1"   "1"
#The paths associated with the output in $merged

$mergedlist[[2]]
      [,1] [,2]
label "no" "yes"
label "NA" "NA"

$mergedlist[[3]]
     [,1] [,2]
[1,]   NA   NA

$mergedlist[[4]]
      [,1]   [,2]      [,3]
label "no"   "yes"     "no"
label "csec" "notcsec" "notcsec"

$mergedlist[[5]]
     [,1] [,2]
[1,]   NA   NA

$mergedlist[[6]]
 label  label
 "yes" "csec"

$mergedlist[[7]]
     [,1] [,2]
[1,]   NA   NA


$lik #The final score for the chosen CEG
[1] -651.3952
```

Bibliography

[1] C Keeble, PD Baxter, S Barber, and GR Law. Participation rates in epidemiology studies and surveys: A review 2007–2015. *The Internet Journal of Epidemiology*, 14(1):1–14, 2015.

[2] C Keeble, S Barber, GR Law, and PD Baxter. Participation bias assessment in three high impact journals. *Sage Open*, 3(4):1–5, 2013.

[3] C Keeble, G Law, S Barber, and PD Baxter. Choosing a method to reduce selection bias: A tool for researchers. *Open Journal of Epidemiology*, 5:155–162, 2015.

[4] C Keeble, PA Thwaites, PD Baxter, S Barber, RC Parslow, and GR Law. Learning through chain event graphs: The role of maternal factors in a childhood type I diabetes. In draft, 2016.

[5] C Keeble, PA Thwaites, S Barber, GR Law, and PB Baxter. Adaptation of chain event graphs for use with case-control studies. In draft, 2016.

[6] C Keeble, S Barber, PD Baxter, RC Parslow, and GR Law. Reducing participation bias in case-control studies: Type 1 diabetes in children and stroke in adults. *Open Journal of Epidemiology*, 4(3):129–134, 2014.

[7] PA McKinney, R Parslow, K Gurney, G Law, HJ Bodansky, and DRR Williams. Antenatal risk factors for childhood diabetes mellitus; A case-control study of the medical record data in Yorkshire, UK. *Diabetologia*, 40:933–939, 1997.

[8] LM Barclay. *Modelling and reasoning with chain event graphs in health studies*. PhD thesis, University of Warwick, 2014.

[9] Thomson Reuters. ISI Web of Knowledge. Journal Citation Reports, JCR Science Edition 2010. http://webofknowledge.com, 2010.

[10] EB Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

[11] JM Last. *A Dictionary of Epidemiology, Fourth Edition*. Oxford University Press, 2001.

[12] S Greenland, J Pearl, and JM Robins. Causal diagrams for epidemiological research. *Epidemiology*, 10:37–48, 1999.

[13] Health Knowledge. Introduction to study designs - cohort studies. http://www.healthknowledge.org.uk/e-learning/epidemiology/practitioners/introduction-study-design-cs, 2016. Accessed online: 27/04/2016.

[14] JJ Schlesselman. *Case-Control Studies: Design, Conduct, Analysis*. New York: Oxford University Press, 1982.

[15] YK Tu, RW West, GDH Ellison, and MS Gilthorpe. Why evidence for the fetal origins of adult disease can be statistical artifact: The reversal paradox examined for hypertension. *American Journal of Epidemiology*, 161(1):27–32, 2004.

[16] CH Hennekens and JE Buring. *Epidemiology in Medicine. SL Mayrent [Ed.]*. Boston: Little, Brown and Co., 1987.

[17] S Galea and M Tracy. Participation rates in epidemiologic studies. *Annals of Epidemiology*, 17(9):643–653, 2007.

[18] KJ Rothman, S Greenland, and TL Lash. *Modern Epidemiology, 3rd Edition*. Philadelphia (PA): Lipincott-Raven, 2008.

[19] NICE: National Institute for Health and Care Excellence. Glossary: Letter R. https://www.nice.org.uk/glossary?letter=r, 2016. Accessed online: 27/04/2016.

[20] J Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2000.

[21] JQ Smith and PE Anderson. Conditional independence and chain event graphs. *Artificial Intelligence*, 172:42–68, 2008.

[22] BR Kirkwood. *Cohort and Case-Control Studies. In Essentials of Medical Statistics*. Oxford: Blackwell Scientific Publications, 1988.

[23] P Cole. The evolving case-control study. *Journal of Chronic Diseases*, 32:15–27, 1979.

[24] Y Jiang, AJ Scott, and CJ Wild. Adjusting for non-response in population-based case-control studies. *International Statistical Review*, 79(2):145–159, 2011.

[25] J Cornfield. A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast and cervix. *Journal of the National Cancer Institute*, 11:1269–1275, 1951.

[26] C J Mann. Observational research methods. Research design II: Cohort, cross sectional, and case-control studies. *Emergency Medicine Journal*, 20(1):54–60, 2003.

[27] MJ Sneyd and B Cox. Commentary: Decreasing response rates require investigators to quantify and report the impact of selection bias in case-control studies. *International Journal of Epidemiology*, 40:1355–1357, 2011.

[28] GS Marquis, JP Habicht, CF Lanata, RE Black, and KM Rasmussen. Association of breastfeeding and stunting in Peruvian toddlers: An example of reverse causality. *International Journal of Epidemiology*, 26(2):349–356, 1997.

[29] Government Digitial Service. HSCIC changing its name to NHS Digital. https://www.gov.uk/government/news/hscic-changing-its-name-to-nhs-digital, 2016. Accessed online: 05/05/2016.

[30] Health & Social Care Information Centre. Hospital Episode Statistics. http://www.hscic.gov.uk/hes, 2016. Accessed online: 05/05/2016.

[31] The Health Improvement Network (THIN) Research Team. THIN Database. https://www.ucl.ac.uk/pcph/research-groups-themes/thin-pub/database, 2016.

[32] R Farmer and R Lawrenson. *Epidemiology and Public Health Medicine, 5th Edition*. Oxford: Blackwell Publishing, 2004.

[33] KJ Rothman and S Greenland. *Modern Epidemiology, 2nd Edition*. Philadelphia (PA): Lipincott-Raven, 1998.

[34] M Gail, R Williams, DP Byar, and C Brown. How many controls? *Journal of Chronic Diseases*, 29:723–731, 1976.

[35] A Waldram, C McKerr, M Gobin, G Adak, JM Stuart, and P Cleary. Control selection methods in recent case-control studies conducted as part of infectious disease outbreaks. *European Journal of Epidemiology*, 30:465–471, 2015.

[36] PG Smith, LC Rodrigues, and PEM Fine. Assessment of the protective efficacy of vaccines against common diseases using case-control and cohort studies. *International Journal of Epidemiology*, 13:87–93, 1984.

[37] B Kirkwood and J Sterne. *Essential Medical Statistics*. Malden (MA): Wiley-Blackwell, 2003.

[38] RM Daniel, MG Kenward, SN Cousens, and BL De Stavola. Using causal diagrams to guide analysis in missing data problems. *Statistical Methods in Medical Research*, 2012.

[39] N Mantel and W Haenszel. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22:719–748, 1959.

[40] IS Silva. *Cancer Epidemiology: Principles and Methods*. International Agency for Research on Cancer, 2016.

[41] E Bareinboim and J Pearl. Controlling selection bias in causal inference. Technical report, Department of Computer Science, University of California, 2012.

[42] P Hartge. Raising response rates: Getting to yes. *Epidemiology*, 10(2):105–107, 1999.

[43] The American Association for Public Opinion Research. Standard definitions, final dispositions of case codes and outcome rates for surveys. http://aapor.org, 2011. Accessed online: 16/12/2011.

[44] LM Morton, J Cahill, and P Hartge. Reporting participation in epidemiologic studies: A survey of practice. *American Journal of Epidemiology*, 163:197–203, 2005.

[45] P Hartge. Participation in population studies. *Epidemiology*, 17(3):252–254, 2006.

[46] A Rogers, MA Murtaugh, S Edwards, and ML Slattery. Contacting controls: Are we working harder for similar response rates, and does it make a difference? *American Journal of Epidemiology*, 160(1):85–90, 2004.

[47] RM Groves and LE Lyberg. *Telephone Survey Methodology*. New York: John Wiley & Sons, 1988.

[48] J Jones. The effects of non-response on statistical inference. *Journal of Health & Social Policy*, 8:49–62, 1996.

[49] RM Groves. *Survey Errors and Survey Costs*. Hoboken (NJ): John Wiley & Sons, 2004.

[50] K Kypri, S Stephenson, and J Langley. Assessment of nonresponse bias in an Internet survey of alcohol use. *Alcoholism: Clinical and Experimental Research*, 28:630–634, 2004.

[51] A Stang and KH Jockel. Studies with low response proportions may be less biased than studies with high response proportions. *American Journal of Epidemiology*, 159:204–210, 2004.

[52] LF Voigt, DM Boudreau, NS Weiss, KE Malone, CI Li, and JR Daling. Re: "Studies with low response proportions may be less biased than studies with high response proportions". *American Journal of Epidemiology*, 161(4):401–402, 2005.

[53] N Pandeya, GM Williams, AC Green, PM Webb, and DC Whiteman. Do low control response rates always affect the findings? Assessments of smoking and obesity in two Australian case-control studies of cancer. *Australian and New Zealand Journal of Public Health*, 33(4):312–319, 2009.

[54] G Law, A Smith, and E Roman. The importance of full participation: Lessons from a national case-control study. *British Journal of Cancer*, 86:350–355, 2002.

[55] S Wacholder, DT Silverman, JK McLaughlin, and S Mandel. Selection of controls in case-control studies: II. Types of controls. *American Journal of Epidemiology*, 135:1029–1041, 1992.

[56] VL Holt, JR Daling, A Stergachis, LF Voigt, and NS Weiss. Results and effect of refusal recontact in a case-control study of ectopic pregnancy. *Epidemiology*, 2(5):375–379, 1991.

[57] MP Madigan, R Troisi, N Potischman, D Brogan, MD Gammon, KE Malone, and LA Brinton. Characteristics of respondents and non-respondents from a case-control study of breast cancer in younger women. *International Journal of Epidemiology*, 29:793–798, 2000.

[58] M Wrensch. Are prior head injuries or diagnostic X-rays associated with glioma in adults? The effects of control selection bias. *Neuroepidemiology*, 19:234–244, 2000.

[59] JF Ludvigsson, P Otterblad-Olausson, BU Pettersson, and A Ekbom. The Swedish personal identity number: Possibilities and pitfalls in healthcare and medical research. *European Journal of Epidemiology*, 24(11):659–667, 2009.

[60] C Sortsø, LC Thygesen, and H Brønnum-Hansen. Database on Danish population-based registers for public health and welfare research. *Scandinavian Journal of Public Health*, 39(S7):17–19, 2011.

[61] EA Nohr, M Frydenberg, TB Henriksen, and J Olsen. Does low participation in cohort studies induce bias? *Epidemiology*, 17:413–418, 2006.

[62] RM Groves and MP Couper. *Nonresponse in Household Interview Surveys*. New York: John Wiley & Sons, 1998.

[63] MJ O'Neil. Estimating the nonresponse bias due to refusals in telephone surveys. *Public Opinion Quarterly*, 43(2):218–232, 1979.

[64] L Hardell, KH Mild, and M Carlberg. Case-control study on the use of cellular and cordless phones and the risk for malignant brain tumours. *International Journal of Radiation Biology*, 78(10):931–936, 2002.

[65] C Pelucchi, S Franceschi, F Levi, D Trichopoulos, C Bosetti, E Negri, and C La Vecchia. Fried potatoes and human cancer. *International Journal of Cancer*, 105(4):558–560, 2003.

[66] J Kennet, J Groerer, KR Bowman, PC Martin, and DE Cunningham. *Evaluating and Improving Methods used in the National Survey on Drug Use and Health*. Rockville (MD): Substance Abuse and Mental Health Services, 2005.

[67] E Singer, RM Groves, DA Dillman, JL Eltinger, and RJA Little. *The Use of Incentives to Reduce Nonresponse in Household Surveys*. New York: Wiley, 2002.

[68] UK Childhood Cancer Study Investigators. The United Kingdom Childhood Cancer Study: Objectives, materials and methods. *British Journal of Cancer*, 82:1073–1102, 2000.

[69] A Goodman and R Gatward. Who are we missing? Area deprivation and survey participation. *European Journal of Epidemiology*, 23(6):379–387, 2008.

[70] JA Burg, SL Allred, and JH Sapp. The potential for bias due to attrition in the National Exposure Registry: An examination of reasons for nonresponse, nonrespondent characteristics, and the response rate. *Toxicology and Industrial Health*, 13:1–13, 1997.

[71] TC Wild, J Cunningham, and E Adlaf. Nonresponse in a follow-up to a representative telephone survey of adult drinkers. *Journal of Studies on Alcohol*, 62:257–261, 2001.

[72] TM Eagan, GE Eide, A Gulsvik, and PS Bakke. Nonresponse in a community cohort study: Predictors and consequences for exposure-disease associations. *Journal of Clinical Epidemiology*, 55:775–781, 2002.

[73] KM Dunn, K Jordan, RJ Lacey, M Shapley, and C Jinks. Patterns of consent in epidemiologic research: Evidence from over 25,000 responders. *American Journal of Epidemiology*, 159:1087–1094, 2004.

[74] ET Hille, L Elbertse, JB Gravenhorst, R Brand, and SP Verloove-Vanhorick (on behalf of the Dutch POPS-19 Collaborative Study Group). Nonresponse bias in a follow-up study of 19-year-old adolescents born as preterm infants. *Pediatrics*, 116:e662–e666, 2005.

[75] CB Cunradi, R Moore, M Killoran, and G Ames. Survey nonresponse bias among young adults: The role of alcohol, tobacco, and drugs. *Substance Use and Misuse*, 40:171–185, 2005.

[76] MR Partin, M Malone, M Winnett, J Slater, A Bar-Cohen, and L Caplan. The impact of survey nonresponse bias on conclusions drawn from a mammography intervention trial. *Journal of Clinical Epidemiology*, 56:867–873, 2003.

[77] CN Weaver, SL Holmes, and ND Glenn. Some characteristics of inaccessible respondents in a telephone survey. *Journal of Applied Psychology*, 60:260–262, 1975.

[78] E Shahar, AR Folsom, and R Jackson. The effect of nonresponse on prevalence estimates for a referent population: Insights from a population-based cohort study. Atherosclerosis Risk in Communities (ARIC) Study Investigators. *Annals of Epidemiology*, 6:498–506, 1996.

[79] L Richiardi, P Boffetta, and F Merletti. Analysis of nonresponse bias in a population-based case-control study on lung cancer. *Journal of Clinical Epidemiology*, 55:1033–1040, 2002.

[80] M Maclure and S Hankinson. Analysis of selection bias in a case-control study of renal adenocarcinoma. *Epidemiology*, 1(6):441–447, 1990.

[81] EE Hatch, MS Linet, RA Kleinerman, RE Tarone, RK Severson, CT Hartsock, C Haines, WT Kaune, D Friedman, LL Robison, and S Wacholder. Association between childhood acute lymphoblastic leukemia and use of electrical appliances during pregnancy and childhood. *Epidemiology*, 9(3):234–245, 1998.

[82] EE Hatch, RA Kleinerman, MS Linet, RE Tarone, WT Kaune, A Auvinen, D Baris, LL Robison, and S Wacholder. Do confounding or selection factors of residential wiring codes and magnetic fields distort findings of electromagnetic field studies? *Epidemiology*, 11:189–198, 2000.

[83] JG Gurney, S Davis, SM Schwartz, BA Mueller, WT Kaune, and RG Stevens. Childhood cancer occurrence in relation to power line configurations: A study of potential selection bias in case-control studies. *Epidemiology*, 6(1):31–35, 1995.

[84] JH Lubin, PE Burns, WJ Blot, RG Ziegler, AW Lees, and JF Fraumeni Jr. Dietary factors and breast cancer risk. *International Journal of Cancer*, 28(6):685–689, 1982.

[85] A McTiernan, NS Weiss, and JR Daling. Bias resulting from using the card-back system to contact patients in an epidemiologic study. *American Journal of Public Health*, 76(1):71–73, 1986.

[86] AC Mertens and LL Robison. Evaluation of parental participation in a case-control study of infant leukaemia. *Paediatric and Perinatal Epidemiology*, 11(2):240–246, 1997.

[87] JS Silberberg, J Wlodarczyk, J Fryer, CD Ray, and MJ Hensley. Correction for biases in a population-based study of family history and coronary heart disease: The Newcastle Family History Study. *American Journal of Epidemiology*, 147(12):1123–1132, 1998.

[88] J Dockerty, D Skegg, J Elwood, G Herbison, D Becroft, and M Lewis. Infections, vaccinations, and the risk of childhood leukaemia. *British Journal of Cancer*, 80:1483–1489, 1999.

[89] JM Mond, B Rodgers, PJ Hay, C Owen, and PJ Beumont. Nonresponse bias in a general population survey of eating-disordered behaviour. *International Journal of Eating Disorders*, 36(1):89–98, 2004.

[90] SJ Beglin and CG Fairburn. Women who chose not to participate in surveys on eating disorders. *International Journal of Eating Disorders*, 12:113–116, 1992.

[91] LA McNutt and R Lee. Intimate partner violence prevalence estimation using telephone surveys: Understanding the effect of nonresponse bias. *American Journal of Epidemiology*, 152:438–441, 2000.

[92] DA Mott, CA Pedersen, WR Doucette, CA Gaither, and JC Schommer. A national survey of U.S. pharmacists in 2000: Assessing nonresponse bias of a survey methodology. *American Association of Pharmaceutical Scientists*, 3(4):76–86, 2001.

[93] GR Law, PD Baxter, and MS Gilthorpe. *Selection Bias in Epidemiological Studies. In Statistical Epidemiology, Y-K Tu and D Greenwood [Eds.]*. New York: Springer, 2011.

[94] S Geneletti, N Best, MB Toledano, P Elliott, and S Richardson. Uncovering selection bias in case-control studies using Bayesian post-stratification. *Statistics in Medicine*, 32(15):2555–2570, 2013.

[95] RCGP. *Profile of UK Practices, RCGP Information Sheet No 2*. London: Royal College of General Practitioners, 1999.

[96] IM Bongers and JA van Oers. Mode effects on self-reported alcohol use and problem drinking: Mail questionnaires and personal interviewing compared. *Journal of Studies on Alcohol*, 59:280–285, 1998.

[97] PA Wingo, HW Ory, PM Layde, NC Lee, and The Cancer and Steroid Hormone Study Group. The evaluation of the data collection process for a multicenter, population-based, case-control design. *American Journal of Epidemiology*, 128(1):206–217, 1988.

[98] GW Yun and CW Trumbo. Comparative response to a survey executed by post, e-mail & web form. http://jcmc.indiana.edu, 2000. Accessed online: 22/11/2011.

[99] FJ Fowler Jr, PM Gallagher, VL Stringfellow, AM Zaslavsky, JW Thompson, and PD Cleary. Using telephone interviews to reduce nonresponse bias to mail surveys of health plan members. *Medical Care*, 40:190–200, 2002.

[100] AM Zaslavsky, LB Zaborski, and PB Cleary. Factors affecting response rates to the Consumer Assessment of Health Plans Study survey. *Medical Care*, 40:485–499, 2002.

[101] TJ Beebe, ME Davern, DD McAlpine, KT Call, and TH Rockwood. Increasing response rates in a survey of Medicaid enrollees: The effect of a prepaid monetary incentive and mixed modes (mail and telephone). *Medical Care*, 43(4):411–414, 2005.

[102] KN Carter, F Imlach-Gunasekara, SK McKenzie, and T Blakely. Differential loss of participants does not necessarily cause selection bias. *Australian and New Zealand Journal of Public Health*, 36:218–222, 2012.

[103] A de Winter, A Oldehinkel, R Veenstra, J Brunnekreef, F Verhulst, and J Ormel. Evaluation of non-response bias in mental health determinants and outcomes in a large sample of pre-adolescents. *European Journal of Epidemiology*, 20:173–181, 2005.

[104] A Alonso, M Segu-Gmez, J de Irala, A Snchez-Villegas, J Beunza, and M Martnez-Gonzalez. Predictors of follow-up and assessment of selection bias from dropouts using inverse probability weighting in a cohort of university graduates. *European Journal of Epidemiology*, 21:351–358, 2006.

[105] GD Batty and Gale CR. Impact of resurvey non-response on the associations between baseline risk factors and cardiovascular disease mortality: Prospective cohort study. *Journal of Epidemiology & Community Health*, 63:952–955, 2009.

[106] J Powers and D Loxton. The impact of attrition in an 11-year prospective longitudinal study of younger women. *Annals of Epidemiology*, 20:318–321, 2010.

[107] J Banks, A Murieal, and JP Smith. Attrition and health in ageing studies: Evidence from ELSA and HRS. *Logitudinal and Life Course Studies*, 2:101–126, 2011.

[108] LD Howe, B Galobardes, K Tilling, and Lawlor DA. Does drop-out from cohort studies bias estimates of socioeconomic inequalities in health? *Journal of Epidemiology & Community Health*, 65:A31, 2011.

[109] LD Howe, K Tilling, B Galobardes, and Lawlor DA. Loss to follow-up in cohort studies: Bias in estimates of socioeconomic inequalities. *Epidemiology*, 24:1–9, 2013.

[110] S Greenland. Quantifying biases in causal models: Classical confounding vs collider-stratification bias. *Epidemiology*, 14:300–306, 2003.

[111] M Hernan, S Hernandez-Diaz, and J Robins. A structural approach to selection bias. *Epidemiology*, 15:615–625, 2004.

[112] J Textor, J Hardt, and S Knppel. DAGitty: A graphical tool for analyzing causal diagrams. *Epidemiology*, 5(22):745, 2011. Letter to the Editor.

[113] J Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–710, 1995.

[114] J Pearl. *Probabilistic Reasoning in Intelligence Systems*. San Francisco (CA): Morgan Kaufmann, 1988.

[115] SR Cole, RW Platt, EF Schisterman, H Chu, D Westreich, D Richardson, and C Poole. Illustrating bias due to conditioning on a collider. *International Journal of Epidemiology*, 39(2):417–420, 2010.

[116] F Elwert and C Winship. Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology*, 40:31–53, 2014.

[117] S Haneuse, J Schildcrout, P Crane, J Sonnen, J Breitner, and E Larson. Adjustment for selection bias in observational studies with application to the analysis of autopsy data. *Neuroepidemiology*, 32:229–239, 2009.

[118] P Spirtes, C Glymour, and R Scheines. *Causation, Prediction and Search. Lecture Notes in Statistics 81*. New York: Springer-Verlag, 1993.

[119] E Bareinboim, J Tian, and J Pearl. Recovering from selection bias in causal and statistical inference. *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[120] C Pizzi, B De Stavola, F Merletti, R Bellocco, I dos Santos Silva, N Pearce, and L Richiardi. Sample selection and validity of exposure-disease association estimates in cohort studies. *Journal of Epidemiology & Community Health*, 65:407–411, 2011.

[121] C Pizzi, BL De Stavola, N Pearce, F Lazzarato, P Ghiotti, F Merletti, and L Richiardi. Selection bias and patterns of confounding in cohort studies: The case of the NINFEA web-based birth cohort. *Journal of Epidemiology & Community Health*, 66:976–981, 2012.

[122] SD Simon. Understanding the odds ratio and the relative risk. *Journal of Andrology*, 22:533–536, 2001.

[123] G King and L Zeng. Logistic regression in rare events data. *Political Analysis*, 9:137–163, 2001.

[124] I Langner, R Bender, R Lenz-Tonjes, H Kuchenhoff, and M Blettner. Bias of maximum-likelihood estimates in logistic and Cox regression models: A comparative simulation study. http://www.statistik.lmu.de/sfb386/papers/dsp/paper362.pdf, 2003. Accessed online: 09/04/2015.

[125] S Nemes, JM Jonasson, A Genell, and G Steineck. Bias in odds ratios by logistic regression modelling and sample size. *BMC Medical Research Methodology*, 9, 2009. doi: 10.1186/1471-2288-9-56.

[126] JS Bergtold, EA Yeager, and AM Featherstone. Sample size and robustness of inferences from logistic regression in the presence of nonlinearity and multicollinearity, 2011. The Agricultural & Applied Economics Association's 2011 AAEA & NAREA Joint Annual Meeting, Pittsburgh, Pennsylvania, 24–26 July 2011.

[127] C Winship and RD Mare. Models for sample selection bias. *Annual Review of Sociology*, 18:327–350, 1992.

[128] J Berkson. Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, 2(3):47–53, 1946.

[129] EA Clark and TW Anderson. Does screening by "pap" smears help prevent cervical cancer? A case-control study. *Lancet*, 2(8132):1–4, 1979.

[130] Thomson Reuters. Web of Science. http://webofknowledge.com, 2015.

[131] LM O'Keeffe, PM Kearney, and RA Greene. Pregnancy risk assessment monitoring system in Ireland: Methods and response rates. *Maternal and Child Health Journal*, 19(3):480–486, 2015.

[132] J Matias-Guiu, PJ Serrano-Castro, JA Mauri-Llerda, FJ Hernandez-Ramos, JC Sanchez-Alvarez, and M Sanz. Analysis of factors influencing telephone call response rate in an epidemiological study. *The Scientific World Journal*, 2014, 2014. doi: 10.1155/2014/179375.

[133] AE Hall, RW Sanson-Fisher, MC Lynagh, T Threlfall, and CA D'Este. Format and readability of an enhanced invitation letter did not affect participation rates in a cancer registry-based study: A randomized controlled trial. *Journal of Clinical Epidemiology*, 66(1):85–94, 2013.

[134] K Chen, H Lei, G Li, W Huang, and L Mu. Cash incentives improve participation rate in a face-to-face survey: An intervention study. *Journal of Clinical Epidemiology*, 68(2):228–233, 2015.

[135] M Hara, Y Higaki, T Imaizumi, N Taguchi, K Nakamura, H Nanri, T Sakamoto, M Horita, K Shinchi, and K Tanaka. Factors influencing participation rate in a baseline survey of a genetic cohort in Japan. *Journal of Epidemiology*, 20(1):40–45, 2010.

[136] DF Perez, JX Nie, CI Ardern, N Radhu, and P Ritvo. Impact of participant incentives and direct and snowball sampling on survey response rate in an ethnically diverse community: Results from a pilot study of physical activity and the built environment. *Journal of Immigrant and Minority Health*, 15(1):207–214, 2013.

[137] NA Koloski, M Jones, G Eslick, and NJ. Talley. Predictors of response rates to a long term follow-up mail out survey. *Plos One*, 8(11), 2013.

[138] SA McLean, SJ Paxton, R Massey, JM Mond, B Rodgers, and PJ Hay. Prenotification but not envelope teaser increased response rates in a bulimia nervosa mental health literacy survey: A randomized controlled trial. *Journal of Clinical Epidemiology*, 67(8):870–876, 2014.

[139] SP Nota, JA Strooker, and D Ring. Differences in response rates between mail, e-mail, and telephone follow-up in hand surgery research. *Hand*, 9(4):504–510, 2014.

[140] RA Tate, M Jones, L Hull, NT Fear, R Rona, S Wessely, and M Hotopf. How many mailouts? Could attempts to increase the response rate in the Iraq war cohort study be counterproductive? *BMC Medical Research Methodology*, 7(51), 2007. doi: 10.1186/1471-2288-7-51.

[141] Y Li, W Wang, Q Wu, MH van Velthoven, L Chen, X Du, Y Zhang, I Rudan, and J Car. Increasing the response rate of text messaging data collection: A delayed randomized controlled trial. *Journal of the American Medical Informatics Association*, 22(1):51–64, 2015.

[142] ST Liu and C Geidenberger. Comparing incentives to increase response rates among African Americans in the Ohio pregnancy risk assessment monitoring system. *Maternal and Child Health Journal*, 15(4):527–533, 2011.

[143] S Lippmann, T Frese, K Herrmann, K Scheller, and H Sandholzer. Primary care research - Trade-off between representativeness and response rate of GP teachers for undergraduates. *Swiss Medical Weekly*, 142, 2012.

[144] M Stafford, S Black, I Shah, R Hardy, M Pierce, M Richards, A Wong, and D Kuh. Using a birth cohort to study ageing: Representativeness and response rates in the National Survey of Health and Development. *European Journal of Ageing*, 10(2):145–157, 2013.

[145] JD Baron, RV Breunig, D Cobb-Clark, T Gorgens, and A Sartbayeva. Does the effect of incentive payments on survey response rates differ by income support history? *Journal of Official Statistics*, 25(4):483–507, 2009.

[146] VS Talaulikar, S Hussain, A Perera, and IT Manyonda. Low participation rates amongst Asian women: Implications for research in reproductive medicine. *European Journal of Obstetrics & Gynecology and Reproductive Biology*, 174:1–4, 2014.

[147] SY Kim, M Tucker, M Danielson, CH Johnson, P Snesrud, and H Shulman. How can PRAMS survey response rates be improved among American Indian mothers? Data from 10 states. *Maternal and Child Health Journal*, 12 Suppl 1:119–125, 2008.

[148] ER Wiebe, J Kaczorowski, and J MacKay. Why are response rates in clinician surveys declining? *Canadian Family Physician*, 58(4):E225–E228, 2012.

[149] PEL Marks, B Babcock, AHN Cillessen, and NR Crick. The effects of participation rate on the internal reliability of peer nomination measures. *Social Development*, 22(3):609–622, 2013.

[150] E Bjertness, A Sagatun, K Green, L Lien, AJ Sogaard, and R Selmer. Response rates and selection problems, with emphasis on mental health variables and DNA sampling, in large population-based, cross-sectional and longitudinal studies of adolescents in Norway. *BMC Public Health*, 10, 2010. doi: 10.1186/1471-2458-10-602.

[151] L Beghin, I Huybrechts, G Vicente-Rodriguez, S De Henauw, F Gottrand, M Gonzales-Gross, J Dallongeville, M Sjostrom, C Leclercq, S Dietrich, M Castillo, M Plada, D Molnar,

M Kersting, CC Gilbert, and LA Moreno. Main characteristics and participation rate of European adolescents included in the HELENA study. *Archives of Public Health*, 70(1):14, 2012.

[152] E Banks, N Herbert, K Rogers, T Mather, and L Jorm. Randomised trial investigating the relationship of response rate for blood sample donation to site of biospecimen collection, fasting status and reminder letter: The 45 and up study. *BMC Medical Research Methodology*, 12, 2012. doi: 10.1186/1471-2288-12-147.

[153] H Tolonen, A Aistrich, and K Borodulin. Increasing health examination survey participation rates by SMS reminders and flexible examination times. *Scandinavian Journal of Public Health*, 42(7):712–717, 2014.

[154] I Garcia, C Portugal, L-H Chu, and AA Kawatkar. Response rates of three modes of survey administration and survey preferences of rheumatoid arthritis patients. *Arthritis Care & Research*, 66(3):364–370, 2014.

[155] TJ Beebe, NJ Talley, M Camilleri, SM Jenkins, KJ Anderson, and GR Locke. The HIPAA authorization form and effects on survey response rates, nonresponse bias, and data quality - A randomized community study. *Medical Care*, 45(10):959–965, 2007.

[156] V Owen-Smith, J Burgess-Allen, K Lavelle, and E Wilding. Can lifestyle surveys survive a low response rate? *Public Health*, 122(12):1382–1383, 2008.

[157] J Meiklejohn, J Connor, and K Kypri. The effect of low survey response rates on estimates of alcohol consumption in a general population survey. *Plos One*, 7(4), 2012.

[158] JB VanGeest, TP Johnson, and VL Welch. Methodologies for improving response rates in surveys of physicians - A systematic review. *Evaluation & the Health Professions*, 30(4):303–321, 2007.

[159] NL Keating, AM Zaslavsky, J Goldstein, DW West, and JZ Ayanian. Randomized trial of $20 versus $50 incentives to increase physician survey response rates. *Medical Care*, 46(8):878–881, 2008.

[160] C Thorpe, B Ryan, SL McLean, A Burt, M Stewart, JB Brown, GJ Reid, and S Harris. How to obtain excellent response rates when surveying physicians. *Family Practice*, 26(1):65–68, 2009.

[161] M Brennan and J Charbonneau. Improving mail survey response rates using chocolate and replacement questionnaires. *Public Opinion Quarterly*, 73(2):368–378, 2009.

[162] KM Hawley, JR Cook, and A Jensen-Doss. Do noncontingent incentives increase survey response rates among mental health providers? A randomized trial comparison. *Administration and Policy in Mental Health and Mental Health Services Research*, 36(5):343–348, 2009.

[163] I Balajti, L Darago, R Adany, and K Kosa. College students' response rate to an incentivized combination of postal and web-based health survey. *Evaluation & the Health Professions*, 33(2):164–176, 2010.

[164] P Doerfling, JA Kopec, MH Liang, and JM Esdaile. The effect of cash lottery on response rates to an online health survey among members of the Canadian Association of Retired Persons: A randomized experiment. *Canadian Journal of Public Health-Revue Canadienne De Sante Publique*, 101(3):251–254, 2010.

[165] TB Crews and DF Curtis. Online course evaluations: Faculty perspective and strategies for improved response rates. *Assessment & Evaluation in Higher Education*, 36(7):865–878, 2011.

[166] RT Jacob and B Jacob. Prenotification, incentives, and survey modality: An experimental test of methods to increase survey response rates of school principals. *Journal of Research on Educational Effectiveness*, 5(4):401–418, 2012.

[167] Y Martins, RI Lederman, CL Lowenstein, S Joffe, BA Neville, BT Hastings, and GA Abel. Increasing response rates from physicians in oncology research: A structured literature review and data from a recent physician survey. *British Journal of Cancer*, 106(6):1021–1026, 2012.

[168] F Olsen, B Abelsen, and JA Olsen. Improving response rate and quality of survey data with a scratch lottery ticket incentive. *BMC Medical Research Methodology*, 12, 2012. doi: 10.1186/1471-2288-12-52.

[169] J Dykema, J Stevenson, C Kniss, K Kvale, K Gonzalez, and E Cautley. Use of monetary and nonmonetary incentives to increase response rates among African Americans in the Wisconsin

pregnancy risk assessment monitoring system. *Maternal and Child Health Journal*, 16(4):785–791, 2012.

[170] H Sauermann and M Roach. Increasing web survey response rates in innovation research: An experimental study of static and dynamic contact design features. *Research Policy*, 42(1):273–286, 2013.

[171] SW Pit, T Vo, and S Pyakurel. The effectiveness of recruitment strategies on general practitioner's survey response rates - A systematic review. *BMC Medical Research Methodology*, 14, 2014. doi: 10.1186/1471-2288-14-76.

[172] NL Parsons and MJ Manierre. Investigating the relationship among prepaid token incentives, response rates, and nonresponse bias in a web survey. *Field Methods*, 26(2):191–204, 2014.

[173] M Murdoch, AB Simon, MA Polusny, AK Bangerter, JP Grill, S Noorbaloochi, and MR Partin. Impact of different privacy conditions and incentives on survey response rate, participant representativeness, and disclosure of sensitive information: A randomized controlled trial. *BMC Medical Research Methodology*, 14, 2014. doi: 10.1186/1471-2288-14-90.

[174] K Abdulaziz, J Brehaut, M Taljaard, M Emond, M-J Sirois, JS Lee, L Wilding, and JJ Perry. National survey of physicians to determine the effect of unconditional incentives on response rates of physician postal surveys. *BMJ Open*, 5(2):e007166–e007166, 2015.

[175] G Jamtvedt, S Rosenbaum, KT Dahm, and S Flottorp. Chocolate bar as an incentive did not increase response rate among physiotherapists: A randomised controlled trial. *BMC Research Notes*, 1(34), 2008. doi: 10.1186/1756-0500-1-34.

[176] IA Harris, OK Khoo, JM Young, MJ Solomon, and H Rae. Lottery incentives did not improve response rate to a mailed survey: A randomized controlled trial. *Journal of Clinical Epidemiology*, 61(6):609–610, 2008.

[177] Y Baruch and BC Holtom. Survey response rate levels and trends in organizational research. *Human Relations*, 61(8):1139–1160, 2008.

[178] I Grava-Gubins and S Scott. Effects of various methodologic strategies survey response rates among Canadian physicians and physicians-in-training. *Canadian Family Physician*, 54(10):1424–1430, 2008.

[179] BJ Kelly, TK Fraze, and RC Hornik. Response rates to a mailed survey of a representative sample of cancer patients randomly drawn from the Pennsylvania Cancer Registry: A randomized trial of incentive and length effects. *BMC Medical Research Methodology*, 10, 2010. doi: 10.1186/1471-2288-10-65.

[180] E Bruggen, M Wetzels, K de Ruyter, and N Schillewaert. Individual differences in motivation to participate in online panels: The effect on response rate and response quality perceptions. *International Journal of Market Research*, 53(3):369–390, 2011.

[181] JP Stange and SJ Zyzanski. The effect of a college pen incentive on survey response rate among recent college graduates. *Evaluation Review*, 35(1):93–99, 2011.

[182] M Clark, M Rogers, A Foster, F Dvorchak, F Saadeh, J Weaver, and V Mor. A randomized trial of the impact of survey design characteristics on response rates among nursing home providers. *Evaluation & the Health Professions*, 34(4):464–486, 2011.

[183] AJ Viera and T Edwards. Does an offer for a free on-line continuing medical education (CME) activity increase physician survey response rate? A randomized trial. *BMC Research Notes*, 5(129), 2012. doi: 10.1186/1756-0500-5-129.

[184] RL Medway and J Fulton. When more gets you less: A meta-analysis of the effect of concurrent web options on mail survey response rates. *Public Opinion Quarterly*, 76(4):733–746, 2012.

[185] JY Ziegenfuss, K Burmeister, KM James, L Haas, JC Tilburt, and TJ Beebe. Getting physicians to open the survey: Little evidence that an envelope teaser increases response rates. *BMC Medical Research Methodology*, 12, 2012. doi: 10.1186/1471-2288-12-41.

[186] L Glidewell, R Thomas, G MacLennan, D Bonetti, M Johnston, MP Eccles, R Edlin, NB Pitts, J Clarkson, N Steen, and JM Grimshaw. Do incentives, reminders or reduced burden improve healthcare professional response rates in postal questionnaires? Two randomised controlled trials. *BMC Health Services Research*, 12, 2012. doi: 10.1186/1472-6963-12-250.

[187] LB van der Mark, KE van Wonderen, J Mohrs, PJE Bindels, MA Puhan, and G ter Riet. The effect of two lottery-style incentives on response rates to postal questionnaires in a prospective cohort study in preschool children at high risk of asthma: A randomized trial. *BMC Medical Research Methodology*, 12, 2012. doi: 10.1186/1471-2288-12-186.

[188] J Dykema, J Stevenson, L Klein, Y Kim, and B Day. Effects of e-mailed versus mailed invitations and incentives on response rates, data quality, and costs in a web survey of university faculty. *Social Science Computer Review*, 31(3):359–370, 2013.

[189] J Bakan, B Chen, C Medeiros-Nancarrow, JC Hu, PW Kantoff, and CJ Recklitis. Effects of a gift certificate incentive and specialized delivery on prostate cancer survivors' response rate to a mailed survey: A randomized-controlled trial. *Journal of Geriatric Oncology*, 5(2):127–132, 2014.

[190] SW Pit, V Hansen, and D Ewald. A small unconditional non-financial incentive suggests an increase in survey response rates amongst older general practitioners (GPs): A randomised controlled trial study. *BMC Family Practice*, 14, 2013. doi: 10.1186/1471-2296-14-108.

[191] J Dykema, J Stevenson, B Day, SL Sellers, and VL Bonham. Effects of incentives and prenotification on response rates and costs in a national web survey of physicians. *Evaluation & the Health Professions*, 34(4):434–447, 2011.

[192] RAA Kanaan, SC Wessely, and D Armstrong. Differential effects of pre and post-payment on neurologists' response rates to a postal survey. *BMC Neurology*, 10, 2010. doi: 10.1186/1471-2377-10-100.

[193] KM James, JY Ziegenfuss, JC Tilburt, AM Harris, and TJ Beebe. Getting physicians to respond: The impact of incentive type and timing on physician survey response rates. *Health Services Research*, 46(1):232–242, 2011.

[194] F Keusch. How to increase response rates in list-based web survey samples. *Social Science Computer Review*, 30(3):380–388, 2012.

[195] N Mitchell, CE Hewitt, E Lenaghan, E Platt, L Shepstone, DJ Torgerson, and Scoop Study Team. Prior notification of trial participants by newsletter increased response rates: A randomized controlled trial. *Journal of Clinical Epidemiology*, 65(12):1348–1352, 2012.

[196] G MacLennan, A McDonald, G McPherson, S Treweek, A Avenell, and RECORD Trial Group. Advance telephone calls ahead of reminder questionnaires increase response rate in non-responders compared to questionnaire reminders only: The RECORD phone trial. *Trials*, 15, 2014. doi: 10.1186/1745-6215-15-13.

[197] A Hammink, P Giesen, and M Wensing. Pre-notification did not increase response rate in addition to follow-up: A randomized trial. *Journal of Clinical Epidemiology*, 63(11):1276–1278, 2010.

[198] L Koopman, LG Donselaar, JJ Rademakers, and M Hendriks. A prenotification letter increased initial response, whereas sender did not affect response rates. *Journal of Clinical Epidemiology*, 66(3):340–348, 2013.

[199] RN Carey, A Reid, TR Driscoll, DC Glass, G Benke, and L Fritschi. An advance letter did not increase the response rates in a telephone survey: A randomized trial. *Journal of Clinical Epidemiology*, 66(12):1417–1421, 2013.

[200] Y Xie and SC Ho. Prenotification had no additional effect on the response rate and survey quality: A randomized trial. *Journal of Clinical Epidemiology*, 66(12):1422–1426, 2013.

[201] AM Hart, CW Brennan, D Sym, and E Larson. The impact of personalized prenotification on response rates to an electronic survey. *Western Journal of Nursing Research*, 31(1):17–23, 2009.

[202] TJ Beebe, E Rey, JY Ziegenfuss, S Jenkins, K Lackore, NJ Talley, and RG Locke. Shortening a survey and using alternative forms of prenotification: Impact on response rate and quality. *BMC Medical Research Methodology*, 10, 2010. doi: 10.1186/1471-2288-10-50.

[203] CM Byrne, JD Harrison, JM Young, and WS Selby. Including the questionnaire with an invitation letter did not improve a telephone survey's response rate. *Journal of Clinical Epidemiology*, 60(12):1312–1314, 2007.

[204] PJ Kroth, L McPherson, R Leverence, W Pace, E Daniels, RL Rhyne, RL Williams, and Prime Net Consortium. Combining web-based and mail surveys improves response rates: A PBRN study from PRIME Net. *Annals of Family Medicine*, 7(3):245–248, 2009.

[205] E Funkhouser, JL Fellows, VV Gordan, DB Rindal, PJ Foy, GH Gilbert, and The National Dental Practice-Based Research Network. Supplementing online surveys with a mailed option to reduce bias and improve response rate: The national dental practice-based research network. *Journal of Public Health Dentistry*, 74(4):276–282, 2014.

[206] O Rolfson, R Salomonsson, LE Dahlberg, and G Garellick. Internet-based follow-up questionnaire for measuring patient-reported outcome after total hip replacement surgery - Reliability and response rate. *Value in Health*, 14(2):316–321, 2011.

[207] JS Boschman, HF van der Molen, MHW Frings-Dresen, and JK Sluiter. Response rate of bricklayers and supervisors on an internet or a paper-and-pencil questionnaire. *International Journal of Industrial Ergonomics*, 42(1):178–182, 2012.

[208] T-H Shih and X Fan. Comparing response rates in e-mail and paper surveys: A meta-analysis. *Educational Research Review*, 4(1):26–40, 2009.

[209] S Crouch, P Robinson, and M Pitts. A comparison of general practitioner response rates to electronic and postal surveys in the setting of the National STI Prevention Program. *Australian and New Zealand Journal of Public Health*, 35(2):187–189, 2011.

[210] JB Yarger, TA James, T Ashikaga, AJ Hayanga, V Takyi, Y Lum, H Kaiser, and J Mammen. Characteristics in response rates for surveys administered to surgery residents. *Surgery*, 154(1):38–45, 2013.

[211] BD Rookey, L Le, M Littlejohn, and DA Dillman. Understanding the resilience of mail-back survey methods: An analysis of 20 years of change in response rates to national park surveys. *Social Science Research*, 41(6):1404–1414, 2012.

[212] M Barrios, A Villarroya, A Borrego, and C Olle. Response rates and data quality in web and mail surveys administered to PhD holders. *Social Science Computer Review*, 29(2):208–220, 2011.

[213] EW Wolfe, PD Converse, and FL Oswald. Item-level nonresponse rates in an attitudinal survey of teachers delivered via mail and web. *Journal of Computer-Mediated Communication*, 14(1):35–66, 2008.

[214] M Denscombe. Item non-response rates: A comparison of online and paper questionnaires. *International Journal of Social Research Methodology*, 12(4):281–291, 2009.

[215] R Haer and N Meidert. Does the first impression count? Examining the effect of the welcome screen design on the response rate. *Survey Methodology*, 39(2):419–434, 2013.

[216] MJ Parker, A Manan, and S Urbanski. Prospective evaluation of direct approach with a tablet device as a strategy to enhance survey study participant response rate. *BMC Research Notes*, 5:605, 2012. doi: 10.1186/1756-0500-5-605.

[217] F Bolanos, D Herbeck, D Christou, K Lovinger, A Pham, A Raihan, L Rodriguez, P Sheaff, and M-L Brecht. Using facebook to maximize follow-up response rates in a longitudinal study of adults who use methamphetamine. *Substance Abuse : Research and Treatment*, 6:1–11, 2012.

[218] A Scott, S-H Jeon, CM Joyce, JS Humphreys, G Kalb, J Witt, and A Leahy. A randomised trial and economic evaluation of the effect of response mode on response rate, response bias, and item non-response in a survey of doctors. *BMC Medical Research Methodology*, 11, 2011. doi: 10.1186/1471-2288-11-126.

[219] C Aitken, R Power, and R Dwyer. A very low response rate in an on-line survey of medical practitioners. *Australian and New Zealand Journal of Public Health*, 32(3):288–289, 2008.

[220] KL Manfreda, M Bosniak, J Berzelak, I Haas, and V Vehovar. Web surveys versus other survey modes - A meta-analysis comparing response rates. *International Journal of Market Research*, 50(1):79–104, 2008.

[221] M Sinclair, J O'Toole, M Malawaraarachchi, and K Leder. Comparison of response rates and cost-effectiveness for a community-based survey: Postal, Internet and telephone modes with generic or personalised recruitment approaches. *BMC Medical Research Methodology*, 12, 2012. doi: 10.1186/1471-2288-12-132.

[222] SC Westrick and JK Mount. Effects of repeated callbacks on response rate and nonresponse bias: Results from a 17-state pharmacy survey. *Research in Social & Administrative Pharmacy*, 4(1):46–58, 2008.

[223] K Kiezebrink, IK Crombie, L Irvine, V Swanson, K Power, WL Wrieden, and PW Slane. Strategies for achieving a high response rate in a home interview survey. *BMC Medical Research Methodology*, 9, 2009. doi: 10.1186/1471-2288-9-46.

[224] A Christie, H Dagfinrud, O Dale, T Schulz, and KB Hagen. Collection of patient-reported outcomes; - Text messages on mobile phones provide valid scores and high response rates. *BMC Medical Research Methodology*, 14, 2014. doi: 10.1186/1471-2288-14-52.

[225] M de Bruijne and A Wijnant. Improving response rates and questionnaire design for mobile web surveys. *Public Opinion Quarterly*, 78(4):951–962, 2014.

[226] M Bosnjak, W Neubarth, MP Couper, W Bandilla, and L Kaczmirek. Prenotification in web-based access panel surveys - The influence of mobile text messaging versus e-mail on response rates and sample composition. *Social Science Computer Review*, 26(2):213–223, 2008.

[227] PD Converse, EW Wolfe, X Huang, and FL Oswald. Response rates for mixed-mode surveys using mail and e-mail/web. *American Journal of Evaluation*, 29(1):99–107, 2008.

[228] B Bonevski, P Magin, G Horton, M Foster, and A Girgis. Response rates in GP surveys trialling two recruitment strategies. *Australian Family Physician*, 40(6):427–430, 2011.

[229] PC Hardigan, CT Succar, and JM Fleisher. An analysis of response rate and economic costs between mail and web-based surveys among practicing dentists: A randomized trial. *Journal of Community Health*, 37(2):383–394, 2012.

[230] LS Edelman, R Yang, M Guymon, and LM Olson. Survey methods and response rates among rural community dwelling older adults. *Nursing Research*, 62(4):286–291, 2013.

[231] A Pedrana, M Hellard, and M Giles. Registered post achieved a higher response rate than normal mail - A randomized controlled trial. *Journal of Clinical Epidemiology*, 61(9):896–899, 2008.

[232] EA Akl, S Gaddam, R Mustafa, MC Wilson, A Symons, A Grifasi, D McGuigan, and HJ Schuenemann. The effects of tracking responses and the day of mailing on physician survey response rate: Three randomized trials. *Plos One*, 6(2), 2011.

[233] JV Cook, HO Dickinson, and MP Eccles. Response rates in postal surveys of healthcare professionals between 1996 and 2005: An observational study. *BMC Health Services Research*, 9, 2009. doi: 10.1186/1472-6963-9-160.

[234] R Horn, S Jones, and K Warren. The cost-effectiveness of postal and telephone methodologies in increasing routine outcome measurement response rates in CAMHS. *Child and Adolescent Mental Health*, 15(1):60–63, 2010.

[235] S Sahlqvist, Y Song, F Bull, E Adams, J Preston, D Ogilvie, and Iconnect Consortium. Effect of questionnaire length, personalisation and reminder type on response rate to a complex postal survey: Randomised controlled trial. *BMC Medical Research Methodology*, 11, 2011. doi: 10.1186/1471-2288-11-62.

[236] N Khamisa, K Peltzer, D Ilic, and B Oldenburg. Evaluating research recruitment strategies to improve response rates amongst South African nurses. *SA Journal of Industrial Psychology*, 40(1):01–07, 2014.

[237] G Trapp, B Giles-Corti, K Martin, A Timperio, and K Villanueva. Conducting field research in a primary school setting: Methodological considerations for maximizing response rates, data quality and quantity. *Health Education Journal*, 71(5):590–596, 2012.

[238] T-H Shih and X Fan. Comparing response rates from web and mail surveys: A meta-analysis. *Field Methods*, 20(3):249–271, 2008.

[239] C Wakabayashi, K Hayashi, K Nagai, N Sakamoto, and Y Iwasaki. Effect of stamped reply envelopes and timing of newsletter delivery on response rates of mail survey: A randomised controlled trial in a prospective cohort study. *BMJ Open*, 2(5), 2012. doi: 10.1136/bmjopen-2012-001181.

[240] M-S Man, HE Tilbrook, S Jayakody, CE Hewitt, H Cox, B Cross, and DJ Torgerson. Electronic reminders did not improve postal questionnaire response rates or response times: A randomized controlled trial. *Journal of Clinical Epidemiology*, 64(9):1001–1004, 2011.

[241] JY Ziegenfuss, JC Tilburt, K Lackore, S Jenkins, K James, and TJ Beebe. Envelope type and response rates in a survey of health professionals. *Field Methods*, 26(4):380–389, 2014.

[242] N Mitchell, CE Hewitt, DJ Torgerson, and Scoop Trial Group. A controlled trial of envelope colour for increasing response rates in older women. *Aging Clinical and Experimental Research*, 23(3):236–240, 2011.

[243] S Kereakoglow, R Gelman, and AH Partridge. Evaluating the effect of esthetically enhanced materials compared to standard materials on clinician response rates to a mailed survey. *International Journal of Social Research Methodology*, 16(4):301–306, 2013.

[244] MD Kaplowitz, F Lupi, MP Couper, and L Thorp. The effect of invitation design on web survey response rates. *Social Science Computer Review*, 30(3):339–349, 2012.

[245] TJ Beebe, SM Stoner, KJ Anderson, and AR Williams. Selected questionnaire size and color combinations were significantly related to mailed survey response rates. *Journal of Clinical Epidemiology*, 60(11):1184–1189, 2007.

[246] KA King and RA Vidourek. Effect of respondent code location on survey response rate. *Psychological Reports*, 109(3):718–722, 2011.

[247] F Kundig, A Staines, T Kinge, and TV Perneger. Numbering questionnaires had no impact on the response rate and only a slight influence on the response content of a patient safety culture survey: A randomized trial. *Journal of Clinical Epidemiology*, 64(11):1262–1265, 2011.

[248] JT Pickett, CF Metcalfe, T Baker, M Gertz, and L Bedard. Superficial survey choice: An experimental test of a potential method for increasing response rates and response quality in correctional surveys. *Journal of Quantitative Criminology*, 30(2):265–284, 2014.

[249] R Teclaw, MC Price, and K Osatuke. Demographic question placement: Effect on item response rates and means of a veterans health administration survey. *Journal of Business and Psychology*, 27(3):281–290, 2012.

[250] T Moradi, A Sidorchuk, and J Hallqvist. Translation of questionnaire increases the response rate in immigrants: Filling the language gap or feeling of inclusion? *Scandinavian Journal of Public Health*, 38(8):889–892, 2010.

[251] JM Brick, JM Montaquila, D Han, and D Williams. Improving response rates for Spanish speakers in two-phase mail surveys. *Public Opinion Quarterly*, 76(4):721–732, 2012.

[252] E Flüß, CM Bond, GT Jones, and GJ Macfarlane. The effect of an Internet option and single-sided printing format to increase the response rate to a population-based study: A randomized controlled trial. *BMC Medical Research Methodology*, 14, 2014. doi: 10.1186/1471-2288-14-104.

[253] Y Choudhury, I Hussain, S Parsons, A Rahman, S Eldridge, and M Underwood. Methodological challenges and approaches to improving response rates in population surveys in areas of extreme deprivation. *Primary Health Care Research & Development*, 13(3):211–218, 2012.

[254] EE Bolt, A van der Heide, and BD Onwuteaka-Philipsen. Reducing questionnaire length did not improve physician response rate: A randomized trial. *Journal of Clinical Epidemiology*, 67(4):477–481, 2014.

[255] J O'Toole, M Sinclair, and K Leder. Maximising response rates in household telephone surveys. *BMC Medical Research Methodology*, 8, 2008. doi: 10.1186/1471-2288-8-71.

[256] DA Dillman, G Phelps, R Tortora, K Swift, J Kohrell, J Berck, and BL Messer. Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the Internet. *Social Science Research*, 38(1):3–20, 2009.

[257] T Nguyet and JA Dilley. Achieving a high response rate with a health care provider survey, Washington State, 2006. *Preventing Chronic Disease*, 7(5), 2010.

[258] K Olson, JD Smyth, and HM Wood. Does giving people their preferred survey mode actually increase survey participation rates? An experimental examination. *Public Opinion Quarterly*, 76(4):611–635, 2012.

[259] K Nicholls, K Chapman, T Shaw, A Perkins, MM Sullivan, S Crutchfield, and E Reed. Enhancing response rates in physician surveys: The limited utility of electronic options. *Health Services Research*, 46(5):1675–1682, 2011.

[260] NT Fear, L Van Staden, A Iversen, J Hall, and S Wessely. 50 ways to trace your veteran: Increasing response rates can be cheap and effective. *European Journal of Psychotraumatology*, 1, 2010.

[261] M Blohm, J Hox, and A Koch. The influence of interviewers' contact behavior on the contact and cooperation rate in face-to-face household surveys. *International Journal of Public Opinion Research*, 19(1):97–111, 2007.

[262] P Rao. International survey research understanding national cultures to increase survey response rate. *Cross Cultural Management - An International Journal*, 16(2):165–178, 2009.

[263] F Munoz-Leiva, J Sanchez-Fernandez, FJ Montoro-Rios, and JA Ibanez-Zapata. Improving the response rate and quality in web-based surveys through the personalization and frequency of reminder mailings. *Quality & Quantity*, 44(5):1037–1052, 2010.

[264] K Olson, JM Lepkowski, and DH Garabrant. An experimental examination of the content of persuasion letters on nonresponse rates and survey estimates in a nonresponse follow-up study. *Survey Research Methods*, 5(1):21–26, 2011.

[265] A Luiten. Personalisation in advance letters does not always increase response rates. Demographic correlates in a large scale experiment. *Survey Research Methods*, 5(1):11–20, 2011.

[266] JW Dembosky, AM Haviland, MN Elliott, P Kallaur, CA Edwards, E Sekscenski, AM Zaslavsky, and JA Brown. Does naming the focal plan in a CAHPS survey of health care quality affect response rates and beneficiary evaluations? *Public Opinion Quarterly*, 77(2):455–473, 2013.

[267] S Misra, D Stokols, and AH Marino. Using norm-based appeals to increase response rates in evaluation research: A field experiment. *American Journal of Evaluation*, 33(1):88–98, 2012.

[268] W Wang, JA Geller, A Kim, TA Morrison, JK Choi, and W Macaulay. Factors affecting response rates to mailed preoperative surveys among arthroplasty patients. *World Journal of Orthopedics*, 3(1):1–4, 2012.

[269] L Claudio and JA Stingone. Improving sampling and response rates in children's health research through participatory methods. *Journal of School Health*, 78(8):445–451, 2008.

[270] MV Pruitt. Deviant research: Deception, male Internet escorts, and response rates. *Deviant Behavior*, 29(1):70–82, 2008.

[271] A Rashidian, J van der Meulen, and I Russell. Differences in the contents of two randomized surveys of GPs' prescribing intentions affected response rates. *Journal of Clinical Epidemiology*, 61(7):718–721, 2008.

[272] TVO Hansen, MK Simonsen, FC Nielsen, and YA Hundrup. Collection of blood, saliva, and buccal cell samples in a pilot study on the Danish nurse cohort: Comparison of the response rate

and quality of genomic DNA. *Cancer Epidemiology Biomarkers & Prevention*, 16(10):2072–2076, 2007.

[273] LR Frass, XL Hopkins, LU Smith, J Kyle, J Vanderknyff, and GA Hand. A collaborative approach in workforce assessment: South Carolina's strategies for high response rates. *Health Promotion Practice*, 15:14S–22S, 2014.

[274] KE van Wonderen, J Mohrs, M Ijff, PJE Bindels, and G ter Riet. Two simple strategies (adding a logo or a senior faculty's signature) failed to improve patient participation rates in a cohort study: Randomized trial. *Journal of Clinical Epidemiology*, 61(10):971–977, 2008.

[275] JY Ziegenfuss, ND Shah, JR Deming, HK Van Houten, SA Smith, and TJ Beebe. Offering results to participants in a diabetes survey: Effects on survey response rates. *Patient-Patient Centered Outcomes Research*, 4(4):241–245, 2011.

[276] C Stenhammar, P Bokstrom, B Edlund, and A Sarkadi. Using different approaches to conducting postal questionnaires affected response rates and cost-efficiency. *Journal of Clinical Epidemiology*, 64(10):1137–1143, 2011.

[277] P Ellwood, MI Asher, AW Stewart, and ISAAC Phase III Study Group. The impact of the method of consent on response rates in the ISAAC time trends study. *International Journal of Tuberculosis and Lung Disease*, 14(8):1059–1065, 2010.

[278] L Jin. Improving response rates in web surveys with default setting. The effects of default on web survey participation and permission. *International Journal of Market Research*, 53(1):75–94, 2011.

[279] DM Berman, LL Tan, and TL Cheng. Surveys and response rates. *Pediatrics in review / American Academy of Pediatrics*, 36(8):364–366, 2015.

[280] C-E Särndal, B Swensson, and J Wretman. *Model assisted survey sampling*. New York: Springer, 1992.

[281] DG Kleinbaum, H Morgenstern, and LL Kupper. Selection bias in epidemiological studies. *American Journal of Epidemiology*, 113(4):452–463, 1981.

[282] M Austin, M Criqui, E Barret-Connor, and M Holdbrook. The effect of response bias on the odds ratio. *The American Journal of Epidemiology*, 114:137–143, 1981.

[283] S Greenland and R Neutra. An analysis of detection bias and proposed corrections in the study of estrogens and endometrial cancer. *Journal of Chronic Diseases*, 34:433–438, 1982.

[284] DG Horvitz and DJ Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

[285] A Gelman. Struggles with survey weighting and regression modelling. *Statistical Science*, 22:153–164, 2007.

[286] JJ Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):153–161, 1979.

[287] B Zadrozny. Learning and evaluating classifiers under sample selection bias. *ICML '04 Proceedings of the Twenty-First International Conference on Machine Learning*, page 114, 2004.

[288] AT Smith and C Elkan. Making generative classifiers robust to selection bias. *KDD '07 Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 657–666, 2007.

[289] RJA Little and DB Rubin. *Statistical Analysis with Missing Data*. Hoboken, New Jersey: Wiley, 2002.

[290] DB Rubin. Multiple imputations in sample surveys - A phenomenological Bayesian approach to nonresponse. In *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pages 20–34, 1978.

[291] RJA Little. Missing data adjustment in large surveys. *Journal of Business and Economic Statistics, American Statistical Association*, 6:287–301, 1988.

[292] JAC Sterne, IR White, JB Carlin, M Spratt, P Royston, MG Kenward, AM Wood, and JR Carpenter. Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ*, 338:2393–2397, 2009.

[293] BD Spycher, M Feller, M Zwahlen, M Roosli, NX von der Weid, H Hengartner, M Egger, C Kuehni, for the Swiss Paediatric Oncology Group, and the Swiss National Cohort Study

Group. Childhood cancer and nuclear power plants in Switzerland: A census-based cohort study. *International Journal of Epidemiology*, 40(5):1247–1260, 2011.

[294] The Cochrane Collaboration. 9.7 Sensitivity analyses. http://handbook.cochrane.org/chapter_9/9_7_sensitivity_analyses.htm, 2016. Accessed online: 28/04/2016.

[295] S Greenland. Basic methods for sensitivity analysis of biases. *International Journal of Epidemiology*, 25:1107–1116, 1996.

[296] J Cornfield, W Haenszel, WC Hammond, AM Lilienfeld, MB Shimkin, and EL Wynder. Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, 22:173–203, 1959.

[297] PR Rosenbaum. *Encyclopedia of Statistics in Behavioral Science*, chapter : Sensitivity Analysis in Observational Studies. Chichester: John Wiley & Sons, 2005.

[298] A Saltelli, S Tarantola, and F Campolongo. Sensitivity analysis as an ingredient of modeling. *Statistical Science*, 15:377–395, 2000.

[299] S Greenland. Multiple-bias modelling for analysis of observational data. *Journal of the Royal Statistical Society, Series A: Statistics in Society*, 168:267–306, 2005.

[300] S Greenland. The impact of prior distributions for uncontrolled confounding and response bias: A case study of the relation of wire codes and magnetic fields to childhood leukemia. *Journal of the American Statistical Association*, 98(461):47–54, 2003.

[301] KJ Rothman. *Epidemiology: An Introduction*. New York: Oxford University Press, 2012.

[302] LJ Keele. *rbounds: Perform Rosenbaum bounds sensitivity tests for matched and unmatched data*, 2014. R package version 2.1.

[303] R Development Core Team. R: A language and environment for statistical computing. http://www.r-project.org/, 2012.

[304] L Keele. *An overview of rbounds: An R package for Rosenbaum bounds sensitivity analysis with matched data*, 2010.

[305] S Geneletti, A Mason, and N Best. Adjusting for selection effects in epidemiologic studies: Why sensitivity analysis is the only "solution". *Epidemiology*, 22(1):36–39, 2011.

[306] NL Allen and PW Holland. Exposing our ignorance: The only "solution" to selection bias. *Journal of Educational Statistics*, 14(2):141–145, 1989.

[307] A Scott and C Wild. Fitting logistic regression models in stratified case-control studies. *Biometrics*, 47:497–510, 1991.

[308] A Lahkola, T Salminen, and A Auvinen. Selection bias due to differential participation in a case-control study of mobile phone use and brain tumors. *Annals of Epidemiology*, 15(5):321–325, 2005.

[309] M Stevenson, T Nunes, C Heuer, J Marshall, J Sanchez, R Thornton, J Reiczigel, J Robison-Cox, P Sebastiani, P Solymos, K Yoshida, G Jones, S Pirikahu, and S Firestone. *epiR: Tools for the Analysis of Epidemiological Data*, 2015. R package version 0.9-69.

[310] RB D'Agostino Jr. Tutorial in biostatistics. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomised control group. *Statistics in Medicine*, 17:2265–2281, 1998.

[311] A Chattopadhyay. *Oral Health Epidemiology: Principles and Practice*. Sudbury (MA): Jones and Barlett Publishers, 2011.

[312] MS Bloom, EF Schisterman, and ML Hediger. The use and misuse of matching in case-control studies: The example of PCOS. *Fertility and Sterility*, 88(3):707–710, 2007.

[313] The Cochrane Collaboration. 13.6.2.1 Controlling for confounding. http://handbook.cochrane.org/chapter_13/13_6_2_1_controlling_for_confounding.htm, 2016. Accessed online: 15/02/2016.

[314] NE Breslow and NE Day. *Chapter 3: General Considerations for the Analysis of Case-Control Studies. In Statistical Methods in Cancer Research, NE Breslow and NE Day [Eds.]*. IARC Scientific Publications, Number 32, 1980.

[315] DW Hosmer Jr, S Lemeshow, and RX Sturdivant. *Applied Logistic Regression*. Hoboken (NJ): John Wiley & Sons, 2013.

[316] EF Schisterman, SR Cole, and RW Platt. Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology*, 20(4):488–495, 2009.

[317] NHS Choices. High blood pressure (hypertension). http://www.nhs.uk/conditions/Blood-pressure-(high)/Pages/Introduction.aspx, 2016. Accessed online: 12/02/2016.

[318] J Textor and B van der Zander. *dagitty: Graphical Analysis of Structural Causal Models*, 2016. R package version 0.1-10.

[319] M Haapea, J Miettunen, J Veijola, E Lauronen, P Tanskanen, and M Isohanni. Non-participation may bias the results of a psychiatric survey: An analysis from the survey including magnetic resonance imaging within the Northern Finland 1966 Birth Cohort. *Social Psychiatry and Psychiatric Epidemiology*, 42:403–409, 2007.

[320] R Lopez, M Frydenberg, and V Baelum. Non-participation and adjustment for bias in case-control studies of periodontitis. *European Journal of Oral Sciences*, 116:405–411, 2008.

[321] S Haneuse and J Chen. A multiphase design strategy for dealing with participation bias. *Biometrics*, 67:309–318, 2011.

[322] CC Tam, CD Higgins, and LC Rodrigues. Effect of reminders on mitigating participation bias in a case-control study. *BMC Medical Research Methodology*, 11:33, 2011. doi: 10.1186/1471-2288-11-33.

[323] J Henderson and S Chatfield. Who matches? Propensity scores and bias in the causal effects of education on participation. *The Journal of Politics*, 73:646–658, 2011.

[324] E Korn and B Graubard. *Analysis of Health Surveys*. Hoboken (NJ): John Wiley & Sons, 1999.

[325] MR Elliott and RJA Little. Model-based alternatives to trimming survey weights. *Journal of Official Statistics*, 16:191–210, 2000.

[326] United Nations Economic Commission for Europe (UNECE). Glossary of Terms on Statistical Data Editing. http://www.unece.org/, 2000. Accessed online: 09/05/2016.

[327] BM Ford. *An Overview of Hot-Deck Procedures. In Incomplete Data in Sample Surveys, WG Madow, L Okin and DB Rubin [Eds.].* New York: Academic Press, 1983.

[328] G Kalton and D Kasprzyk. The treatment of missing survey data. *Survey Methodology*, 12:1–16, 1986.

[329] JT Lessler and WD Kalsbeck. *Nonsampling Error in Surveys*. New York: Wiley, 1992.

[330] C-E Särndal. Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, 18:241–252, 1992.

[331] DB Rubin. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley, 1987.

[332] MG Kenward and J Carpenter. Multiple imputation: Current perspectives. *Statistical Methods in Medical Research*, 16:199–218, 2007.

[333] MA Klebanoff and SR Cole. Use of multiple imputation in the epidemiologic literature. *American Journal of Epidemiology*, 168:355–357, 2008.

[334] HJ Cordell. Estimation and testing in case-control studies: Comparison of weighted regression and multiple imputation procedures. *Genetic Epidemiology*, 30:259–275, 2006.

[335] S van Buuren and K Groothuis-Oudshoorn. MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45:1–67, 2011.

[336] R Mansson, MM Joffe, W Sun, and S Hennessy. On the estimation and use of propensity scores in case-control and case-cohort studies. *American Journal of Epidemiology*, 166(3):332–339, 2007.

[337] PC Austin. The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. *Medical Decision Making*, 29(6):661–677, 2009.

[338] PR Rosenbaum and DB Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[339] BB Reeve, A Wilder Smith, NK Arora, and RD Hays. Reducing bias in cancer research: Application of propensity score matching. *Health Care Financing Review*, 29(4):69–80, 2008.

[340] A Exuzides, C Colby, J Goldman, and A Waaler. Reducing bias in a retrospective case-control study: An application of propensity score matching. In *ISPOR 12th Annual European Congress; Paris, France, October 2009*, 2009.

[341] MC Walsh, A Trentham-Dietz, PA Newcomb, R Gangnon, and M Palta. Using propensity scores to reduce case-control selection bias. *Epidemiology*, 23(5):772–773, 2012.

[342] AK Jaffer, WK Barsoum, V Krebs, JG Hurbanek, N Morra, and JD Brotman. Duration of anesthesia and venous thromboembolism after hip and knee arthroplasty. *Mayo Clinic Proceedings*, 80(6):732–738, 2005.

[343] BD Smith, GL Smith, and BG Haffty. Postmastectomy radiation and mortality in women with T1-2 node-positive breast cancer. *Journal of Clinical Oncology*, 23(7):1409–1419, 2005.

[344] AB Hill, D Obrand, K O'Rourke, OK Steinmetz, and N Miller. Hemispheric stroke following cardiac surgery: A case-control estimate of the risk resulting from ipsilateral asymptomatic carotid artery stenosis. *Annals of Vascular Surgery*, 14(3):200–209, 2000.

[345] P Walsh, L Shanholtzer, M Loewen, K Trinh, B McEnulty, and SJ Rothenberg. A matched case control study with propensity score balancing examining the protective effect of paracetamol against parentally reported apnoea in infants. *Resuscitation*, 83(4):440–446, 2012.

[346] MM Joffe and PR Rosenbaum. Invited commentary: Propensity scores. *American Journal of Epidemiology*, 150(4):327–333, 1999.

[347] DB Rubin and N Thomas. Matching using estimated propensity scores: Relating theory to practice. *Biometrics*, 52(1):249–264, 1996.

[348] SY Guo and MW Fraser. *Propensity Score Analysis: Statistical Methods and Applications*, chapter 4: Sample selection and related models. California (USA): SAGE Publications, 2014.

[349] NE Breslow. Presedential address: XXI International Biometric Conference, Freiburg, Germany, July 2002 - Are statistical contributions to medicine undervalued? *Biometrics*, 59(1):1–8, 2002.

[350] CS Hollenbeak, D Murphy, WC Dunagan, and VJ Fraser. Nonrandom selection and the attributable cost of surgical-site infections. *Infection control and hospital epidemiology*, 23(4):177–182, 2002.

[351] CY Wang, SJ Wang, RG Gutierrez, and RJ Carroll. Local linear regression for generalized linear models with missing data. *Annals of Statistics*, 26(3):1028–1050, 1998.

[352] S Geneletti, S Richardson, and N Best. Adjusting for selection bias in retrospective, case-control studies. *Biostatistics*, 10(1):17–31, 2009.

[353] RHM Smit, C Untiedt, and JM van Ruitenbeek. The high-bias stability of monatomic chains. *Nanotechnology*, 15(7):S472–S478, 2004.

[354] B Lindner and EM Nicola. Diffusion in different models of active Brownian motion. *European Physical Journal - Special Topics*, 157:43–52, 2008.

[355] S Kwon, ES Lee, H Seo, KJ Jeon, CC Hwang, YH Kim, and JY Park. Reversible oxidation states of single layer graphene tuned by electrostatic potential. *Surface Science*, 612:37–41, 2013.

[356] OA Arah. When selection bias in intractable to the bias breaking variable method. *American Journal of Epidemiology*, 171(Supplement 11):S107, 2010.

[357] M Dersarkissian and OA Arah. Multiply robust estimation of the effects in case control studies in the presence of selection bias. *American Journal of Epidemiology*, 177(Supplement 11):S101, 2013.

[358] NICE: National Institute for Health and Care Excellence. Glossary: Letter I. https://www.nice.org.uk/glossary?letter=i, 2016. Accessed online: 28/04/2016.

[359] Office for National Statistics. Adult drinking habits in Great Britain, 2013. http://www.ons.gov.uk/ons/rel/ghs/opinions-and-lifestyle-survey/adult-drinking-habits-in-great-britain–2013/stb-drinking-2013.html, 2015. Accessed online: 14/04/2015.

[360] NCSS Statistical Software. Data stratification. http://ncss.wpengine.netdna-cdn.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Data_Stratification.pdf, 2015. Accessed online: 14/04/15.

[361] G Freeman and JQ Smith. Bayesian MAP model selection of chain event graphs. *Journal of Multivariate Analysis*, 102:1152–1165, 2011.

[362] P Thwaites. Causal identifiability via chain event graphs. *Artificial Inteligence*, 195:291–315, 2013.

[363] T Silander and T-Y Leong. *A Dynamic Programming Algorithm for Learning Chain Event Graphs*. Berlin: Springer, 2013.

[364] LM Barclay, JL Hutton, and JQ Smith. Refining a Bayesian network using a chain event graph. *International Journal of Approximate Reasoning*, 54:1300–1309, 2013.

[365] LM Barclay, JL Hutton, and JQ Smith. Chain event graphs for informed missingness. *Bayesian Analysis*, 9:53–76, 2014.

[366] GF Cooper. The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, 42:393–405, 1990.

[367] D Heckerman. A tutorial on learning Bayesian networks. Technical report, Microsoft Research, 1995.

[368] D Ron, Y Singer, and N Tishby. On the learnability and usage of acyclic probabilistic finite automata. *Journal of Computer and System Sciences*, 56:133–152, 1998.

[369] WL Buntine. Chain graphs for learning. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI1995)*, pages 46–54, 2013.

[370] N Cruz-Ramirez, HG Acosta-Mesa, H Carrillo-Calvet, LA Nava-Fernandez, and RE Barrientos-Martinez. Diagnosis of breast cancer using Bayesian networks: A case study. *Computers in Biology and Medicine*, 37(11):1553–1564, 2007.

[371] M Lappenschaar, A Hommersom, and PJF Lucas. Qualitative chain graphs and their use in medicine. In *Sixth European Workshop on Probabilistic Graphical Models, Granada, Spain, 2012*, 2012.

[372] PW Holland. Statistical and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

[373] A Ankinakatte and D Edwards. Modelling discrete longitudinal data using acyclic probablistic finite automata. *Computational Statistics & Data Analysis*, 88:40–52, 2015.

[374] D Edwards and S Ankinakatte. Context-specific graphical models for discrete longitudinal data. *Statistical Modelling*, 15(4):301–325, 2015.

[375] LM Barclay, JQ Smith, PA Thwaites, and AE Nicholson. The dynamic chain event graph. Technical report, University of Warwick, University of Leeds & Monash University, 2013.

[376] US National Library of Medicine & National Institutes of Health. Pubmed. http://www.ncbi.nlm.nih.gov/pubmed/, 2015. Accessed online: 07/09/2015.

[377] Elsevier. Scopus. http://www.scopus.com/, 2015. Accessed online: 07/09/2015.

[378] Google. Google scholar. https://scholar.google.co.uk/, 2015. Accessed online: 07/09/2015.

[379] G Freeman. *Learning and Predicting with Chain Event Graphs*. PhD thesis, University of Warwick, 2010.

[380] R Development Core Team. R: A language and environment for statistical computing. http://www.r-project.org/, 2015.

[381] LM Barclay, RA Collazo, JQ Smith, PA Thwaites, and AE Nicholson. The dynamic chain event graph. *Electronic Journal of Statistics*, 9:2130–2169, 2015.

[382] NHS choices. Amniocentesis. http://www.nhs.uk/conditions/Amniocentesis, 2015. Accessed online: 23/10/2015.

[383] Home Health UK. The blood. http://www.homehealth-uk.com/medical/blood.htm, 2015. Accessed online: 15/12/2015.

[384] NHS. Rhesus disease - causes. http://www.nhs.uk/Conditions/Rhesus-disease/Pages/Causes.aspx, 2015. Accessed online: 15/12/2015.

[385] P Bolton. Education: Historical statistics - parliament. http://www.parliament.uk/briefing-papers/SN04252.pdf, 2012. Accessed online: 02/12/2015.

[386] Cambridge Fetal Care. Amniocentesis Test. http://www.fetalcare.co.uk, 2013. Accessed online: 13/08/2013.

[387] Birth Choice UK. Graphs of historical caesarean section rates. http://www.birthchoiceuk.com, 2011. Accessed online: 08/08/2013.

[388] T Silander, P Kontkanen, and P Myllymaki. *On Sensitivity of the MAP Bayesian Network Structure to the Equivalent Sample Size Parameter*, chapter Proceedings of The 23rd Conference on Uncertainty in Artificial Intelligence (UAI-2007), pages 360–367. 2007.

[389] P Thwaites, JQ Smith, and E Riccomagno. Causal analysis with chain event graphs. *Artificial Intelligence*, 174:889–909, 2010.

[390] Health Knowledge. Bias in epidemiological studies. http://www.healthknowledge.org.uk/public-health-textbook/research-methods/1a-epidemiology/biases, 2016. Accessed online: 06/01/2016.

[391] B Williams, L Irvine, AR McGinnis, MET McMurdo, and IK Crombie. When "no" might not quite mean "no": The importance of informed and meaningful non-consent: Results from a survey of individuals refusing participation in a health-related research project. *BMC Health Services Research*, 7(59), 2007. doi: 10.1186/1472-6963-7-59.

[392] C Snijders, U Matzat, and U-D Reips. "Big data": Big gaps of knowledge in the field of Internet science. *International Journal of Internet Science*, 7(1):1–5, 2012.

[393] C Eckmann, M Wasserman, F Latif, G Roberts, and A Beriot-Mathiot. Increased hospital length of stay attributable to Clostridium difficile infection in patients with four co-morbidities: An analysis of Hospital Episode Statistics in four European countries. *The European Journal of Health Economics*, 14(5):835–846, 2013.

[394] T Childs, A Scowcroft, and S Todd. Gender and regional differences in the treatment for hypertension: A pharmacoepidemiological analysis of the General Practice Research Database (GPRD) in the context of hypertension in Atrial Fibrillation (AF) patients. *Journal of Human Hypertension*, 27(10):648, 2013. Abstracts from the 2013 Annual Scientific Meeting of the BHS.

[395] SSR Crossfield and SE Clamp. Electronic health records research in a health sector environment with multiple provider types. In *HEALTHINF 2013: Proceedings of the International Conference on Health Informatics*, pages 104–111, 2013.

[396] V Sorganvi, MS Kulkarni, D Kadeli, and S Atharga. Risk factors for stroke: A case control study. *International Journal of Current Research and Review*, 6(3):46–52, 2014.

[397] AH Smith, NE Pearce, and PW Callas. Cancer case-control studies with other cancers as controls. *International Journal of Epidemiology*, 17(2):298–306, 1988.

[398] J Wakefield and P Elliott. Issues in the statistical analysis of small area health data. *Statistics in Medicine*, 18(17–18):2377–2399, 1999.

[399] J Jelip, GG Mathew, Y Yusin, JF Dony, N Singh, M Ashaari, N Lajanin, CS Ratnam, MY Ibrahim, and D Gopinath. Risk factors of tuberculosis among health care workers in Sabah, Malaysia. *Tuberculosis*, 84(1–2):19–23, 2004.

[400] Y Lu, H Jin, MH Chen, and CC Gluer. Reduction of sampling bias of odds ratios for vertebral fractures using propensity scores. *Osteoporosis International*, 17(4):507–520, 2006.

[401] NJ Liu, I Beerman, R Lifton, and HY Zhao. Haplotype analysis in the presence of informatively missing genotype data. *Genetic Epidemiology*, 30(4):290–300, 2006.

[402] D Conway. Oral cancer risk and smokeless tobacco products - clouded by smoke? *Evidence-based dentistry*, 9(4):114–115, 2008.

[403] B Chardon, S Host, G Pedrono, and I Gremy. Contribution of case-crossover design to the analysis of short-term health effects of air pollution: Reanalysis of air pollution and health data. *Revue d epidemiologie et de sante publique*, 56(1):31–40, 2008.

[404] E Paap, A Verbeek, DI Puliti, M Broeders, and E Paci. Minor influence of self-selection bias on the effectiveness of breast cancer screening in case-control studies in the Netherlands. *Journal of Medical Screening*, 18(3):142–146, 2011.

[405] S Heikkinen, M Koskenvuo, N Malila, T Sarkeala, E Pukkala, and J Pitkaniemi. Use of exogenous hormones and the risk of breast cancer: Results from self-reported survey data with validity assessment. *Cancer causes & control*, 27(2):249–258, 2016.

[406] R Neugebauer. Application of a capture-recapture method (the Bernoulli census) to historical epidemiology. *American Journal of Epidemiology*, 120(4):626–634, 1984.

[407] AE Czeizel, E Eros, M Rockenbauer, HT Sorensen, and J Olsen. Short-term oral diazepam treatment during pregnancy - A population-based teratological case-control study. *Clinical Drug Investigation*, 23(7):451–462, 2003.

[408] HP Chan. Detection of spatial clustering with average likelihood ratio test statistics. *Annals of Statistics*, 37(6B):3985–4010, 2009.

[409] VM Park, CA Roberts, and T Jakob. Palaeopathology in Britain: A critical analysis of publications with the aim of exploring recent trends (1997–2006). *International Journal of Osteoarchaeology*, 20(5):497–507, 2010.

[410] CD Anderson, MA Nalls, A Biffi, NS Rost, SM Greenberg, AB Singleton, JF Meschia, and J Rosand. The effect of survival bias on case-control genetic association studies of highly lethal diseases. *Circulation - Cardiovascular Genetics*, 4(2):188–196, 2011.

[411] EM Hart and L Aviles. Reconstructing local population dynamics in noisy metapopulations - The role of random catastrophes and Allee effects. *PLOS ONE*, 9(10):e110049, 2014.

[412] RM Zur, LL Pesce, and YL Jiang. Estimating screening-mammography receiver operating characteristic (ROC) curves from stratified random samples of screening mammograms: A simulation study. *Academic Radiology*, 22(5):580–590, 2015.

[413] C Sweeney, SL Edwards, KB Baumgartner, JS Herrick, LE Palmer, MA Murtaugh, A Stroup, and ML Slattery. Recruiting hispanic women for a population-based study: Validity of surname search and characteristics of nonparticipants. *American Journal of Epidemiology*, 166(10):1210–1219, 2007.

[414] FK Mensahl, EV Willett, J Simpson, AG Smith, and E Roman. Birth order and sibship size: Evaluation of the role of selection bias in a case-control study of non-Hodgkin's lymphoma. *American Journal of Epidemiology*, 166(6):717–723, 2007.

[415] U Stromberg and J Bjork. Incorporating group-level exposure information in case-control studies with missing data on dichotomous exposures. *Epidemiology*, 15(4):494–503, 2004.

[416] KL Ebi, LI Kheifets, RL Pearson, and H Wachtel. Description of a new computer wire coding method and its application to evaluate potential control selection bias in the Savitz et al. childhood cancer study. *Bioelectromagnetics*, 21(5):346–353, 2000.

[417] I Momas, JP Daures, and F Gremy. What controls in a case-control investigation - Study on bladder-cancer in Herault department. *Revue d Epidemiologie et de Sante Publique*, 39(2):197–207, 1991.

[418] Office of Population Censuses and Surveys. Subnational population projections, Series PP3, no.9, Table 5: 1993-Based population projections, 1993-2016: Sex and quinary age-groups, 1995.

[419] Nomis. Official labour market statistics. https://www.nomisweb.co.uk, 2014. Accessed online: 20/02/2014.

[420] World Health Organization. World Health Statistics 2012. http://apps.who.int, 2012. Accessed online: 20/02/2014.

[421] International Diabetes Federation. Diabetes: Facts and figures. http://www.idf.org, 2014. Accessed online: 20/02/2014.

[422] World Bank. Smoking prevalence, females (% of adults). http://data.worldbank.org, 2014. Accessed online: 20/02/2014.

[423] World Bank. Smoking prevalence, males (% of adults). http://data.worldbank.org, 2014. Accessed online: 20/02/2014.

[424] World Bank. Population (total). http://data.worldbank.org, 2014. Accessed online: 20/02/2014.

[425] Rightdiagnosis.com. Statistics by country for stroke. http://www.rightdiagnosis.com, 2014. Accessed online: 20/02/2014.

[426] Office for National Statistics. Births by mother's area of residence, Table 2a. http://www.ons.gov.uk, 2012. Accessed online: 06/08/2013.

[427] Office for National Statistics. Response rates in the 2011 census, 2012.

[428] Z Bai, S Zheng, B Zhang, and G Hu. Statistical analysis for rounded data. Technical report, NUS Risk Management Institute, 2007.

[429] PJ Borman and MJ Chatfield. Avoid the perils of using rounded data. *Journal of Pharmaceutical and Biomedical Analysis*, 115:502–508, 2015.

[430] J van der Laan and L Kuijvenhoven. Imputation of rounded data, 2011.

[431] E Triastuti Sugiyarto. Analysing rounding data using radial basis function neural networks model. Master's thesis, The University of Northampton, 2007.

[432] S Pudney. Heaping and leaping: Survey response behaviour and the dynamics of self-reported consumption expenditure. Technical report, Institute for Social & Economic Research, University of Essex, 2008.

[433] AJ Gislason-Lee, C Keeble, ON Shahim, AR Cowen, M Lupton, S Vijayan, V Auvray, M Sivananthan, and AG Davies. Is digital subtraction angiography feasible in cardiac x-ray imaging, and does it offer advantages over standard angiograms? In draft.

[434] S Greenland. The effect of misclassification in the presence of covariates. *American Journal of Epidemiology*, 12(4):564–569, 1980.

[435] S Greenland. The effect of misclassification in matched-pair case-control studies. *American Journal of Epidemiology*, 116(2):402–406, 1982.

[436] I dos Santos Silva. *Cancer Epidemiology: Principles and Methods*. Lyon, France: International Agency for Research on Cancer (IARC), 1999.

[437] Fisheries and Aquaculture Department of the Food and Agriculture Organization of the United Nations. 5. Strengths and weaknesses of the Bayesian approach. http://www.fao.org/docrep/005/y1958e/y1958e07.htm, 2016. Accessed online: 23/03/2016.

[438] A Bekhet and J Zauszniewski. Methodological triangulation: An approach to understanding data. *Nurse Researcher*, 20(2):40–43, 2012.

[439] Robert Wood Johnson Foundation. Qualitative Research Guidelines Project: Triangulation. http://www.qualres.org/HomeTria-3692.html, 2008. Accessed online: 23/03/2016.

[440] M Davern. Nonresponse rates are a problematic indicator of nonresponse bias in survey research. *Health Services Research*, 48(3):905–912, 2013.

[441] RM Groves and E Peytcheva. The impact of nonresponse rates on nonresponse bias: A meta-analysis. *Public Opinion Quarterly*, 72:167–189, 2008.

[442] RJA Little and PR Rosenbaum. Confounding and selection bias in case control studies. Technical report, United States Envirionmental Protection Agency, 1981.

[443] VW Sung. Reducing bias in pelvic floor disorders research: Using directed acyclic graphs as an aid. *Neurourology and Urodynamics*, 31:115–120, 2012.

[444] P Thwaites and J Smith. Evaluating causal effects using chain event graphs. *Proceedings of the 3rd European Workshop on Probabilistic Graphical Models, PGM 2006*, pages 293–300, 2006.

[445] E Riccomagno and JQ Smith. The causal manipulation of chain event graphs. Technical report, University of Warwick, 2007.

[446] JQ Smith, E Riccomagno, and P Thwaites. Causal analysis with chain event graphs. Technical report, University of Warwick, 2010.

[447] RG Cowell and JQ Smith. Causal discovery through MAP selection of stratified chain event graphs. *Electronic Journal of Statistics*, 8:965–997, 2014.

[448] PA Thwaites, G Freeman, and JQ Smith. Chain event graph MAP model selection. In *1st International Conference on Knowledge Engineering and Ontology Development Location: Funchal, Portugal*, pages 392–395, 2009.

[449] TJ Peters. Multifarious terminology: multivariable or multivariate? univariable or univariate? *Paediatric and Perinatal Epidemiology*, 22:506, 2008.

[450] T Silander and T-Y Leong. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Berlin: Springer, 2013.

[451] PA Thwaites, JQ Smith, and RG Cowell. Propagation using chain event graphs. *24th Conference on Uncertainty in Artificial Intelligence, UAI 2008*, pages 546–553, 2008.

[452] PA Thwaites and JQ Smith. A separation theorem for chain event graphs. Submitted to Electronic Journal of Statistics, 2015.

[453] C Görgen, M Leonelli, and JQ Smith. *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, volume 9161, chapter 9.3: A Differential Approach for Staged Trees, pages 346–355. Switzerland: Springer, 2015.

[454] PA Thwaites and JQ Smith. A new method for tackling asymmetric decision problems. In *Proceedings of the 10th Workshop on Uncertainty Processing*, 2015. WUPES'15: 10th Workshop on Uncertainty Processing, 16–19 Sep 2015, Monnec, Czech Republic.

[455] JL Hutton, I Irincheeva, and J Hager. Potential for Marie Curie post-doctoral fellowship, 2015.

[456] E Di Nardo. Symbolic methods in statistics: Elegance towards efficiency. In *Algebraic Statistics 2015, 8–12 June, University of Genoa*, 2015.

[457] M Leonelli and JQ Smith. Bayesian decision support for complex systems with many distributed experts. *Annals of Operations Research*, 2015. Online First.

[458] D Edwards and S Ankinakatte. Some context-specific graphical models for discrete longitudinal data. Technical report, Aarhus University, 2013.

[459] LM Barclay, JL Hutton, and JQ Smith. Embellishing a Bayesian network using a chain event graph. In *4th Annual Conference of the Australasian Bayesian Network Modelling Society*, 2012.

[460] RI Smith, JMcP Dick, and EM Scott. The role of statistics in the analysis of ecosystem services. *Environmetrics*, 22:608–617, 2011.

[461] G Freeman and JQ Smith. Dynamic staged trees for discrete multivariate time series : Forecasting, model selection and causal analysis. *Bayesian Analysis*, 6(2):279–305, 2011.

[462] E Riccomagno, JQ Smith, and P Thwaites. Algebraic discrete causal models. Technical report, University of Warwick, 2010.

[463] E Riccomagno and JQ Smith. *The Geometry of Causal Probability Trees that are Algebraically Constrained*, chapter 6, pages 133–154. New York: Springer, 2009.

[464] E Riccomagno. A short history of algebraic statistics. *Metrika*, 69:397–418, 2009.

[465] PA Thwaites. *Chain event graphs : Theory and application*. PhD thesis, University of Warwick, 2008.

[466] RR Ramsahai. *Causal Inference with Instruments and Other Supplementary Variables*. PhD thesis, University of Oxford, 2008.

[467] E Riccomagno and JQ Smith. Algebraic causality: Bayes nets and beyond. Technical report, University of Warwick, 2007.

[468] PE Anderson and JQ Smith. Bayesian representations using chain event graphs. Technical report, University of Warwick, 2006.

[469] PE Anderson and JQ Smith. A graphical framework for representing the semantics of asymmetric models. Technical report, University of Warwick, 2005.

[470] E Riccomagno and JQ Smith. The causal manipulation and Bayesian estimation of chain event graphs. Technical report, University of Warwick, 2005.

[471] M Jaeger. Probabilistic decision graphs - Combining verification and AI techniques for probabilistic inference. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 12:19–42, 2004.

[472] G Chiola, G Balbo, and A Jean-Marie. Requirements for a modeling and performance evaluation software environment, 1995. Computer file.