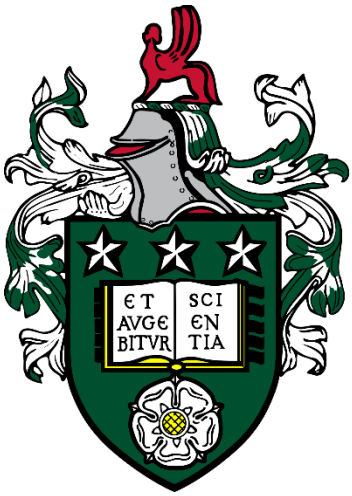
Statistical and simulation-based modelling approaches for causal inference in longitudinal data

Integrating counterfactual thinking into established methods for longitudinal data analysis

Kellyn Fair Arnold



Submitted in accordance with the requirements for the degree of
Doctor of Philosophy

The University of Leeds
School of Medicine

March 2020

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 3 contains work based on the following publications:

1. **Arnold, K.F.**, Berrie, L., Tennant, P.W.G. and Gilthorpe, M.S. A causal inference perspective on the analysis of compositional data. *International Journal of Epidemiology.* 2020, 0(0), pp.1-7. (1)
   Kellyn F. Arnold drafted the manuscript and produced all figures together with Dr Lauren Berrie. All authors conceived the idea and revised the manuscript.
2. **Arnold, K.F.**, Davies, V., de Kamps, M., Tennant, P.W.G., Mbotwa, J. and Gilthorpe, M.S. Reflections on modern methods: Generalised linear models for prognosis and intervention – theory, practice, and implications for machine learning. *International Journal of Epidemiology.* 2020, 0(0), pp.1-9. (2)
   Kellyn F. Arnold researched the literature, performed all analyses, produced all figures, and drafted the manuscript. Prof Mark S. Gilthorpe conceived the manuscript and, along with all co-authors, developed the ideas contained in the manuscript and revised the manuscript.
3. **Arnold, K.F.**, Harrison, W.J., Heppenstall, A.J. and Gilthorpe, M.S. DAG-informed regression modelling, agent-based modelling and microsimulation modelling: a critical comparison of methods for causal inference. *International Journal of Epidemiology.* 2019, 48(1), pp.243-253. (3)
   Kellyn F. Arnold researched the literature, performed all analyses, produced all tables and figures, and drafted the manuscript. All other co-authors revised the manuscript.

Chapter 4 contains work based on the following publication:

4. Tennant, P.W.G., **Arnold, K.F.**, Ellison, G.T.H. and Gilthorpe, M.S. Analyses of 'change scores' do not estimate causal effects in observational data. *ArXiv e-prints.* [Online]. 2019. (4)
   Kellyn F. Arnold researched the literature with Dr Peter W. G. Tennant, performed all path tracing analyses, and rewrote the manuscript into its current form. Prof Mark S. Gilthorpe conceived the study and, along with Dr Peter W. G. Tennant and Kellyn F. Arnold, developed the ideas contained in the manuscript. Dr Peter W. G. Tennant and Prof Mark S Gilthorpe conceived and wrote the simulations, and all authors were involved in their analysis. Prof Mark S. Gilthorpe and Dr George T. H. Ellison revised the manuscript.

Chapter 5 contains work based on the following publication:

5. **Arnold, K.F.**, Ellison, G.T.H., Gadd, S.C., Textor, J., Tennant, P.W.G., Heppenstall, A. and Gilthorpe, M.S. Adjustment for time-invariant and time-varying confounders in 'unexplained residuals' models for longitudinal data within a causal framework and associated challenges. *Statistical Methods in Medical Research.* 2019, 28(5), pp.1347-1364. (5)
   Kellyn F. Arnold researched the literature, performed all analyses and simulations, wrote all mathematical proofs, produced all tables and figures, and drafted the manuscript. Prof Mark S. Gilthorpe conceived the study and, along with all other co-authors, revised the manuscript.

Acknowledgments

Research is not conducted in a cultural or political vacuum. The years I have spent doing my PhD have been some of the most unsettled and unsettling I have known. The last four years have seen Donald Trump be elected president of the United States; a vote for Brexit and ultimately a withdrawal from the European Union; a rise in nationalism, racism, xenophobia, and polarisation throughout the world; and a global pandemic that rages on – refusing to respect international borders or academic deadlines. It is in this context that I feel particularly obliged to acknowledge and thank the people who have helped make the last four years productive, meaningful, and enjoyable.

First and foremost, I must thank my supervisors Mark Gilthorpe, Alison Heppenstall, and Wendy Harrison, for your constant support, guidance, and encouragement. To Mark, in particular, thank you for recognising my potential and for bringing me into your group of eccentric and eclectic researchers. I have always felt my ideas and skills to be valued, and could not have asked for a better PhD experience. To Alison, thank you for being a voice of reason, pragmatism, and optimism; I am constantly in awe of the way you balance so many things at once.

I am also immensely grateful to my colleagues Laurie Berrie, Peter Tennant, and George Ellison, who have provided endless amounts of creativity and support throughout my years as a PhD student. My research would not be half as good if I were not in such a stimulating and collaborative environment, nor would I enjoy it half as much. To George, thank you for always providing a new and unique perspective. To Peter, thank you for being a constant engine of creativity, and for always being available to give support and advice. To Laurie, thank you for leading the way and being an exemplar of how to get through the challenges of a PhD with poise and grace.

This research would not have been possible without the financial support of the Economic and Social Research Council, who provided me with a generous stipend that has allowed me to live comfortably and travel frequently.

Thank you to the Leeds branch of Women in Biostatistics – Mary Cronin and Sulia Celebi. You are two of the most authentic, intelligent, and unique women I know, and I am proud to call you my friends. I could not have made it through this process without your steadfast emotional support, deliciously imaginative pasta recipes, endless banter, and crazy nights in Ibiza. Thank you also to Andrea Bovo for never allowing me to take myself too seriously, and for having the courage to approach a stranger in a corridor. You have all become my family in Leeds and I love you dearly.

Thank you to my Leeds Medics and Dentists Football Club teammates for keeping me young (if not always in one piece).

Thank you also to my *viva voce* examiners – Theresa Munyombwe and Anna Pearce – for a stimulating and enjoyable examination experience.

Possibly my greatest thanks must go to my family, without whom I could not have even arrived at the point where completing a PhD was possible. Thank you for always encouraging me to follow my own path, for being the lender of last resort, and for supporting me from afar throughout this entire process. Despite the many holidays I have missed and birthdays I have forgotten, you have always been by my side.

Finally, I would like to acknowledge and thank all the badass women who carry on doing science in an increasingly hostile world. I am proud to be one of you.

Abstract

The counterfactual framework represents the dominant paradigm for testing and evaluating causal claims within epidemiology. What began as a *philosophical* framework has been formalised mathematically in the language of directed acyclic graphs (DAGs), whose underpinning theory provides a rigorous *mathematical* framework for the identification and estimation of causal effects. Moreover, DAGs provide a conceptual framework for thinking though causal processes and explicating causal assumptions.

Advances in DAG-based methods are invaluable in the era of 'big data', since we are increasingly awash with large, complex – and frequently longitudinal – datasets. However, the relative recentness of such developments means that many established methods for analysing observational data have not been considered within a robust causal framework.

This PhD thesis explores how counterfactual thinking, encoded in the language of DAGs, may be integrated into established methods for longitudinal data analysis, and illustrates several advantages of doing so. Three statistical- and simulation-based methods are considered: (1) the analysis of change, (2) regression with 'unexplained residuals', and (3) microsimulation modelling. For each method, DAGs are specifically employed to consider causal structures and to explore potential problems and/or biases that might arise when these methods are applied without sufficient consideration for such structures. In (1), DAGs are used to demonstrate that 'change scores' do not in general represent exogenous change; alternate analytical strategies for isolating change are identified. In (2), DAGs are employed to illustrate why the method works and how it may be extended to adjust for confounding. In (3), DAGs are used to explicitly consider data-generating processes, and to demonstrate some of the unique challenges faced by simulation approaches. DAGs are demonstrated to be useful tools for informing causal analyses across a wide variety of longitudinal scenarios, thereby providing a basis for integrating counterfactual thinking into other methods for longitudinal data analysis.

Table of contents

List of tables

List of figures

List of abbreviations

ABM	Agent-based model
ANCOVA	Analysis of covariance
BMI	Body mass index
CRCT	Conditionally randomised controlled trial
DAG	Directed acyclic graph
GDP	Gross domestic product
IPTW	Inverse probability of treatment weight(ing)
MSAS	Minimally sufficient adjustment set
MSM	Microsimulation model
NHANES	National Health and Nutrition Examination Survey
OLS	Ordinary least squares
RCT	Randomised controlled trial
RTM	Regression to the mean
SE	Standard error
SNM	Structural nested model
TCE	Total causal effect
UR	Unexplained residual(s)

Chapter 1
Introduction

1.1 Introduction

Estimating the causal effect of a particular factor or event (an '**exposure**'[1]) on a subsequent factor or event (an '**outcome**') is not a trivial task. Causation is a concept for which most (if not all) human beings have an intuitive understanding. Nevertheless, it is a complex phenomenon which may be difficult to even articulate. Despite thousands of years of philosophical discourse, there exists very little consensus as to what it is, how it can be defined, and – perhaps most importantly for researchers – how it can be inferred within practical research applications (7-15). Prominent theories of causation include the regularity, counterfactual, probabilistic, agency and interventionist, and mechanistic theories (16), though no single account may be considered universal because each is subject to counterexamples (17).

The counterfactual framework has risen to prominence as the dominant paradigm for testing and evaluating causal claims in many disciplines; this is likely due to both its conceptual simplicity and its recent mathematical formalisation in the form of directed acyclic graphs (DAGs).[2] However, in spite of its prominence, there exist many (purportedly causal) methods which have not been examined through this lens. This PhD thesis explores how counterfactual thinking, encoded in the language of DAGs, may be integrated into established methods for longitudinal data analysis; this thesis also seeks to demonstrate the advantages of doing so, though focus on the counterfactual framework is not intended to imply its superiority over any other causal framework. The contexts considered are primarily health- and epidemiology-focused, but the analyses performed have applicability to other domains.

Population-level health patterns emerge from a complex, dynamic, and multi-layered system, in which a multitude of different interrelationships operate (21). Estimating causal effects in this context requires somehow accounting for all potential non-causal associations and biases which may distort the association of interest (2). Historically, a 'top-down' approach has been implemented to control such complexities and minimise biases via study design (e.g. case-control studies, randomised controlled trials). However, in the era of 'big data', we are increasingly awash with large and complex datasets from the many systems and technologies that routinely record information on individual experiences. Big data offers much promise for

---

[1] The term 'treatment' is often used interchangeably with 'exposure', particularly in medical- and health-related contexts (6).

[2] This framework is substantively very similar to the 'potential outcomes' framework introduced by Jerzy Neyman in 1923 (18) and more extensively developed and popularised by Donald Rubin from 1974 (19, 20), though the two frameworks employ different terminology and possess other subtle differences.

understanding causal processes, but it does not in and of itself eliminate any of the classical challenges and data quality issues associated with observational data, such as missing or incomplete data and measurement error (22). Neither does big data eliminate the need for *a priori* subject matter knowledge, since *any* association may reach the threshold of 'statistical significance' given sufficiently large sample sizes. To fully exploit the potential of big data, robust methods for evaluating causal relationships are needed which emphasise understanding data-generating processes from the bottom up.

Longitudinal data in particular form a large proportion of the new and emerging forms of data in the digital age. For instance, smartphone apps are able to continuously track and collect data relating to location and activity levels. Hospital records constitute another example, which may additionally be linked to general practice and pharmacy records to create a more comprehensive picture of an individual's interaction with health services over time. Traditional forms of data collection like cross-sectional surveys are inherently longitudinal, since even data which are collected or measured at the same time are likely to have an implicit time ordering. This is because the time at which a variable is measured implies nothing about the time at which its value became manifest. For example, a cross-sectional survey may contain information on individuals' biological sex and their weight, but these variables nevertheless have a clear temporal ordering – sex becomes fixed at the time of conception, whereas weight represents an accumulation of infinitesimal changes throughout the life course and whose value only becomes manifest at the time it is measured. However, the term 'longitudinal' is typically applied to data for which there exist *multiple* measurements of one or more variables, and this is the meaning we adopt throughout. Such data are *explicitly* longitudinal, and are of particular interest to epidemiologists and data scientists as they allow for changes to be quantified and examined. A key focus of life course epidemiology, for instance, is to identify and quantify important periods of change or growth in an exposure, and to evaluate their effect on subsequent outcomes (23, 24).

Longitudinal data may be conceptualised both as exposures and as outcomes, but across all contexts they present analytical challenges for causal inference over and above those of cross-sectional data. This thesis explores some of those challenges in the context of three statistical- and simulation-based methods for assessing causal relationships, and demonstrates the insights that DAGs and the counterfactual framework can bring to causal analyses.

1.2 Aims and objectives

As outlined previously, the aim of this PhD thesis is to explore how DAGs can be integrated into established methods for longitudinal data analysis, and to illustrate the benefits of doing so. To this end, three specific statistical- and simulation-based methods are considered, all of which relate to distinct longitudinal scenarios but which are connected via the fact that they are purportedly used for estimating *causal* relationships.

As broad objectives, DAGs will be used to depict the longitudinal context in which each method is deployed; the principles of graphical model theory will be applied in order to identify how each method should be employed to robustly estimate causal effects; and the potential biases which result from failing to consider each method within a robust causal framework will be identified and explored.

1.3 Thesis overview

Chapter 2 provides background literature related to the counterfactual framework for causal inference, and demonstrates how this framework has been formalised mathematically in the language of DAGs. The aim of this chapter is to provide sufficient information related to the concepts and vocabulary which are necessary for understanding the contents of the remainder of the thesis.

Chapter 3 expands on Chapter 2 by introducing several methods for estimating causal effects in longitudinal data, some of which are based on DAGs but many of which are not. The utility of using DAGs to inform causal analyses is demonstrated through specific examples. Additionally, a critical comparison of statistical methods and individual-based simulation methods is provided, since both have been recognised as useful tools for evaluating counterfactual contrasts. This provides a foundation for understanding the contexts in which the methods considered in the remainder of the thesis may be used, as well as understanding some of the potential strengths and weaknesses of these methods.

Each of the next three chapters uses DAGs to examine a particular method for estimating causal effects in longitudinal data. The methods are both statistical- and simulation-based, and each method relates to a different longitudinal scenario.

Chapter 4 uses DAGs to consider the analysis of change – a topic which has historically been a matter of much disagreement but which has rarely been examined within the framework of DAGs. This context involves quantifying the relationship between a single exposure and subsequent 'change' in a longitudinal outcome. In this chapter, the concept of 'change' is considered within a formal causal framework, in order to demonstrate the analytical strategies most compatible with analysing 'change' and the problems which may arise by failing to consider underlying causal structures and data-generating processes.

Chapter 5 uses DAGs to consider regression with 'unexplained residuals' – a method which was introduced to circumvent some of the difficulties associated with estimating causal effects in longitudinal settings but which was never extended to address confounding. This context involves quantifying the relationship between *separate* measurements of a longitudinal exposure and a subsequent outcome. In this chapter, DAGs are used to illustrate why the method works in the absence of any confounding, and to provide the principles on which the method may be extended robustly to account for confounding by both time-fixed and time-varying covariates.

Chapter 6 uses DAGs to consider microsimulation modelling – a simulation method often used to estimate counterfactual quantities for policy evaluation and which shares many similarities with the statistical 'g-formula'. This context involves quantifying the relationship between *multiple* measurements of a longitudinal exposure and a subsequent outcome. In this chapter, DAGs are used to consider the parallels between the data-generating processes they represent and those which are modelled using microsimulation, and the importance of faithfully modelling data-generating processes from the bottom up in order to make causal inferences.

Chapter 7 summarises the findings and implications of all chapters, including their contributions to the literature. It additionally discusses the limitations of the research contained in the thesis, and offers suggestions for future research of this kind. Potential areas for future research are outlined.

Chapter 2
Background

2.1 Introduction

Epidemiological research relies primarily on the counterfactual theory of causation for testing and evaluating causal claims. Counterfactual reasoning underpins randomised controlled trials, long considered to be the superior and most robust method for demonstrating causal effects. However, the counterfactual framework has also been formalised in the language of DAGs, which provide a rigorous mathematical framework for causal analyses and the identification of causal effects in non-randomised contexts.

Chapter 2 provides a comprehensive introduction to the counterfactual framework for exposures which are both time-fixed and time-varying; of fundamental importance is the concept of exchangeability, which allows for the identification of causal effects in this framework. This chapter also introduces DAGs, and illustrates their utility in identifying causal effects for time-fixed and time-varying exposures. Since this thesis explores how DAGs may be integrated into established methods for longitudinal data analysis, the purpose of this chapter is to provide sufficient information related to the relevant concepts and vocabulary which are necessary for understanding the remainder of this thesis.

2.1.1 Chapter overview

A general chapter overview is provided below.

In Section 2.2, we distinguish between time-fixed and time varying variables, and consequently define what it means for an exposure to be either time-fixed (§2.2.1) or time-varying (§2.2.2).

In Section 2.3, we introduce the counterfactual framework for causal inference. We use specific examples to demonstrate how this framework conceptualises individual-level causal effects for both time-fixed (§2.3.1.1) and time-varying (§2.3.1.2) exposures. We additionally highlight a crucial concept in counterfactual causation – exchangeability (§2.3.2).

In Section 2.4, we discuss how randomisation may be used to identify average causal effects for both time-fixed (§2.4.1) and time-varying (§2.4.2) exposures. We highlight the difference between unconditional and conditional exchangeability in each context.

In Section 2.5, we introduce graphical causal models, with particular focus given to DAGs (§2.5.2). We illustrate how DAGs may be used to identify causal effects for both time-fixed (§2.5.3) and time-varying (§2.5.4) exposures by emulating randomisation.

2.2 Time-fixed versus time-varying variables

A variable is considered to be time-fixed if it occurs only once (e.g. a one-dose vaccine, birthweight), does not change over time (e.g. sex, BRCA1/BRCA2 genes (25)), or evolves over time in a deterministic way (e.g. age, time since treatment) (26). Very few time-fixed variables exist over the entire lifecourse, but over shorter periods of time certain variables may be reasonably conceptualised and/or treated as time-fixed. For example, height changes substantially over the lifecourse, though remains relatively fixed throughout middle-age.

In contrast, a variable is considered to be time-varying if it occurs multiple times (e.g. a multi-dose vaccine) or changes over time (e.g. smoking status, blood sugar) (26). Such variables form the majority of those of interest in epidemiological applications, though the complexity of dealing with variables of this type means that they are often 'reclassified' as time-fixed by defining their values at a particular point in time. For example, height at age three and height at age five could be considered as two distinct time-fixed variables.

2.2.1 Time-fixed exposures

We use the term **time-fixed exposure** to refer to an exposure whose effect on an outcome of interest is only being considered at a single point in time. For example, an epidemiologist might consider what effect obesity at age fifteen has on the risk of depression at age twenty. Although obesity is a time-varying variable, it is considered a time-fixed exposure in this context because its effect is only being considered at the specific age of fifteen.

2.2.2 Time-varying exposures

We use the term **time-varying exposure** to refer to an exposure whose effect on an outcome is being considered at multiple points in time. For example, an epidemiologist might consider what effect obesity at ages ten, fifteen, and eighteen has on the risk of depression at age twenty. A sequence of exposures such as this is often referred to an **exposure (or treatment) regime**.

2.3 The counterfactual framework for causal inference

Here, we introduce the basic concepts of, and the intuition behind, the counterfactual framework for causal inference. This framework is most easily conceptualised in the context of individual-level causal effects, and so we define such effects for both time-fixed (§2.3.1.1) and time-varying (§2.3.1.2) exposures. We additionally highlight the key concept of exchangeability (§2.3.2) and the so-called 'fundamental problem of causal inference' for the identification of individual-level causal effects (§2.3.3).

2.3.1 Individual-level causal effects

2.3.1.1 For time-fixed exposures

The counterfactual framework states that an event X (i.e. a time-fixed exposure) may be considered a cause of an event Y if, *contrary to fact*, had X not occurred then Y would not

have occurred (16, 27). As an example (adapted from (27)), we can imagine that an individual, Mary, is driving to work and comes to a fork in the road. She chooses to go left (i.e. event X) and subsequently arrives late for work (i.e. event Y). Upset, Mary declares 'I should have gone right instead!' What her statement implies is that her decision to go left at the fork in the road caused her to be late for work because, *had she gone right instead (i.e. event 'not X'), she would arrived on time (i.e. event 'not Y')*. Of course, there is no way to prove her statement, as doing so would require Mary to simultaneously go both left and right and then observe the outcome under each condition, in order to guarantee that the effect cannot be attributed to any other factor that differed between the drives. Nevertheless, this example demonstrates the intuition of (and utility behind) conceptualising causal effects as counterfactual contrasts between two scenarios that are equivalent in every way except for the putative causal factor of interest.

2.3.1.2 For time-varying exposures

The counterfactual framework – although more frequently conceptualised in the context of time-fixed exposures – can also be naturally extended to time-varying exposures. To this end, we consider a scenario involving two events X_0 and X_1 (i.e. a time-varying exposure). The events X_0 and X_1 may be considered a *joint* cause of an event of Y if, *contrary to fact*, had at least one of X_0 or X_1 not occurred then Y would not have occurred (26). As an example, we can imagine that Mary takes two doses of antibiotics to treat a chest infection (i.e. events X_0 and X_1, respectively), which clears up (i.e. event Y) after the second dose. We may conclude the two doses of antibiotics are a joint cause of clearing Mary's chest infection if there exists a counterfactual scenario in which Mary did *not* take at least one of the doses and her chest infection did not clear. For example, if Mary did not take the second dose of antibiotics (i.e. event 'not X_1') and her chest infection persisted (i.e. event 'not Y'), we could conclude that the two doses are a joint cause of her chest infection clearing. However, if Mary's chest infection cleared *regardless* of whether she took each dose of antibiotics (i.e. if all counterfactual scenarios resulted in the same outcome), the clearing cannot be attributed to the antibiotic regime.

2.3.2 Exchangeability

Exchangeability is a fundamental concept in counterfactual causation. In this framework, a causal effect is defined in terms of a comparison between two units of analysis which are in all ways equivalent except for the putative causal factor of interest – in other words, two units of analysis which are *exchangeable*.

2.3.3 The 'fundamental problem of causal inference'

The so-called 'fundamental problem of causal inference' (28) is that individual-level causal effects cannot ever be identified because it is impossible to observe an individual subjected to different values of the putative causal factor simultaneously. In other words, it is impossible to

view the unrealised counterfactual scenario(s) and therefore impossible to achieve exchangeability.

2.4 Using randomisation to identify average causal effects

Although identification of individual-level causal effects is generally agreed to be impossible within the counterfactual framework, identification of **average causal effects** *is* possible and forms the basis of a great deal of epidemiological causal inference (6). Average causal effects may be identified by creating exchangeable *groups* of individuals and comparing their *average* outcomes. This is often achieved through randomisation (29, 30).

2.4.1 Average causal effects for time-fixed exposures

To demonstrate the principle of using randomisation to identify the average causal effect for a *time-fixed* exposure, we consider a specific example involving the effect of chemotherapy versus radiotherapy on two-year survival amongst breast cancer patients. We illustrate how both unconditionally and conditionally exchangeable groups of individuals may be created by randomisation.

2.4.1.1 Exchangeability

2.4.1.1.1 Unconditional exchangeability

Epidemiologists have long considered the **randomised controlled trial (RCT)** to be the 'gold standard' for demonstrating causality because, if implemented correctly, it guarantees unconditional exchangeability (31). An RCT in our example context might involve randomly assigning each patients to receive either chemotherapy and radiotherapy, and then comparing the average outcome for each treatment group.

In this situation, the group who received chemotherapy is **unconditionally exchangeable** with the group who received radiotherapy. This is because randomisation ensures that the outcome is equally likely in both groups prior to the intervention, and so a simple comparison of the average outcome for each group after the intervention is sufficient to identify an average causal effect (32). In other words, those who received chemotherapy, *had they instead received radiotherapy*, would have experienced the same average outcomes as those who actually did receive radiotherapy (6), i.e. they are unconditionally exchangeable.[3]

2.4.1.1.2 Conditional exchangeability

We could alternately consider a ***conditionally* randomised controlled trial (CRCT)**, in which each patient is randomly assigned to receive either chemotherapy or radiotherapy *based on*

[3] If there exists differential loss to follow-up, then exchangeability may not be ensured by this process (33, 34). However, this is an additional complexity which we do not cover here, since our purpose is simply to illustrate the conceptual rationale behind such designs.

their initial cancer stage. For example, individuals in stage IV are randomised to receive chemotherapy with a higher probability than radiotherapy.

Here, a simple comparison of the average outcome for each treatment group cannot be assumed sufficient, as any difference in two-year survival might be attributable to the fact that the chemotherapy group has, on average, a worse prognosis at the beginning of the study. Nevertheless, we are still able to identify an average causal effect by comparing the average two-year survival between those who received chemotherapy and those who received radiotherapy among individuals *who had the same initial cancer stage*. Thus, *within each subgroup of cancer stage*, those who received chemotherapy, *had they instead received radiotherapy*, would have experienced the same average outcomes as those who actually did receive radiotherapy (6). The two treatment groups are **conditionally exchangeable**, i.e. they are exchangeable *conditional on* initial cancer stage.

2.4.2 Average causal effects for time-varying exposures

To demonstrate the principle of using randomisation to identify an average causal effect for a *time-varying* exposure, we return to the example from Section 2.3.1.2 involving the use of antibiotics to clear a chest infection, in which a dose of antibiotics may be prescribed at the point of initial diagnosis or at a subsequent follow-up visit.

We illustrate in this context how unconditionally and conditionally exchangeable groups of individuals may be manufactured by sequential randomisation.

2.4.2.1 Exchangeability

2.4.2.1.1 (Sequential) unconditional exchangeability

An RCT in our example context might involve randomly assigning each patient to receive each dose of antibiotics. This is referred to as ‘sequential randomisation’ (26) because patients are randomised at each time point. In this way, we create four treatment groups – those who received two doses, no doses, only the first dose, or only the second dose.

The proportion of people whose infections subsequently cleared in each of the treatment groups may then be directly compared. The process of sequential randomisation ensures that the outcome is equally likely in all groups prior to treatment both at the point of diagnosis and at the point of follow-up. Thus, a simple comparison of the average outcome for each group after the final intervention is sufficient to identify an average causal effect. For example, those who received both doses of antibiotics, *had they instead received one of the other dosing regimes*, would have experienced the same average outcomes as those who actually did receive those other dosing regimes (6), i.e. they are **unconditionally exchangeable**.

2.4.2.1.2 (Sequential) conditional exchangeability

By contrast, a CRCT in our example context might instead involve randomly assigning each patient to receive each dose of antibiotics *based on the severity of their infection at the time*.

For example individual who are initially judged to have more severe infections may be randomised to receive the first dose of antibiotics with a higher probability than those with less severe infections. Similarly, individuals with more severe infections at the follow-up visit may be randomised to receive the second dose with a higher probability.

Because each treatment group (i.e. those who received two doses, no doses, only the first dose, or only the second dose) is likely to have a different average outcome prognosis as a result of the way in which individuals were randomised, they cannot be directly compared. Moreover, we cannot even identify an average causal effect by comparing the proportion cleared chest infections among individuals who had the same infection severity at both time points, because infection severity at the second time point is itself affected by whether or not an individual received the first dose of antibiotics (i.e. *infection severity is not randomised*). However, within subgroups defined by initial infection severity, receipt of the first dose of antibiotics, and follow-up infection severity, those who received the second dose of antibiotics, *had they instead not received the second dose of antibiotics*, would have experienced the same distribution of outcomes as those who actually did not receive the second dose. The average outcome for each of the treatment groups may then be compared within levels of baseline and follow-up infection severity because they are **(sequentially) conditionally exchangeable**, i.e. they are exchangeable at each time point *conditional on* current infection severity.

We will return to this concept in Section 2.5.4, where we present a clearer graphical depiction of this issue (§2.5.4.1) and the challenges associated with identifying casual effects in sequentially randomised contexts (§2.5.4.2).

2.5 Using DAGs to identify average causal effects

For situations in which (C)RCTs are either infeasible (e.g. for extremely rare diseases), impractical (e.g. for complex and/or costly interventions), and/or unethical (e.g. for potentially deadly or otherwise harmful exposures), epidemiologists must rely on observational, *non-randomised* data. However, the average causal effect of an exposure on an outcome may still be identified by using the principles of graphical model theory to *emulate* exchangeability.

In this section, we give a brief introduction to graphical causal models and DAGs, and illustrate how they encode counterfactual statements for both time-fixed and time-varying exposures.

2.5.1 Graphical causal models

Modern causal models trace their roots to 1918, with Sewall Wright's invention of path analysis (35, 36). They also have origins in structural equation models (SEMs), which represent groups of causally related variables (both observed and latent) as systems of simultaneous linear equations (37). However, both were subsumed at the beginning of the twenty-first century under the framework of **nonparametric causal models** by Judea Pearl in his seminal book *Causality* (38).

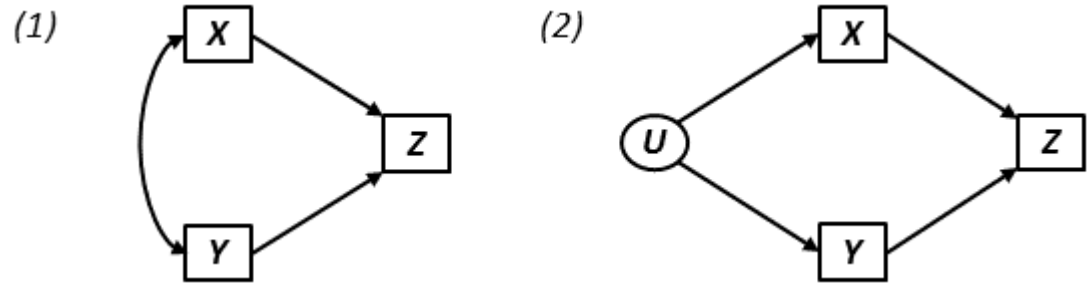
These models are typically represented graphically (hence '***graphical* causal models**'[4]) and consist of two fundamental components:

1. A set of *variables* (i.e. 'nodes'); and
2. A set of *arrows* (i.e. 'arcs' or 'edges').

Any two variables in the graph may be connected by an arrow (e.g. $A \rightarrow B$), which means that the first variable (A) exerts a direct causal effect on the second (B) for at least one member of the population (39). A variable may be either **endogenous** (i.e. having at least one direct cause represented in the graph), or **exogenous** (i.e. having no direct causes represented in the graph). However, the graph makes no assumptions about the distribution of the variables, nor does it imply or constrain either the magnitude or functional form of the causal effects (27, 39).

Two examples of graphical causal models are provided in Figure 2.1. By convention, observed and/or measured variables are denoted by rectangles, whereas unmeasured and/or unobserved (i.e. latent) variables are denoted by ovals. We also adopt the convention that time flows from left to right (6).

Figure 2.1 Graphical causal models depicting the causal relationships between three random variables X, Y, and Z



Observed and/or measured variables are depicted in rectangular boxes, and latent variables are depicted in ovals. In panel (1), X and Y cause one another. In panel (2), X and Y are caused by the unobserved variable U.

The graphical causal model in panel (1) of Figure 2.1 depicts the causal relationships between the variables X, Y, and Z; the graph implies that both X and Y are direct causes of Z, and that X and Y cause each other. The graphical causal model in panel (2) of Figure 2.1 is very similar to that in panel (1), but instead depicts X and Y as being caused by the unobserved variable U, which is the source of their mutual dependency.

The graph in panel (2) is a particular type of graphical causal model – a directed acyclic graph.

[4] Graphical causal models may alternately be referred to as 'causal diagrams' (6), 'graphical models' (27), or simply 'graphs' (27).

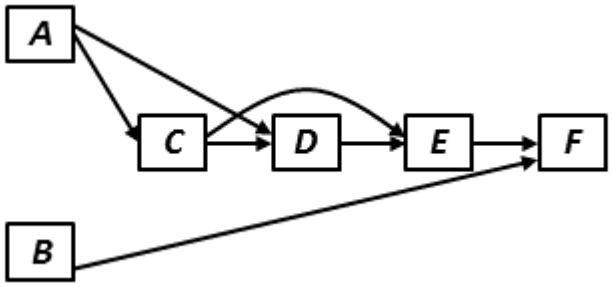
2.5.2 Directed acyclic graphs (DAGs)

Directed acyclic graphs (DAGs) represent a special subset of graphical causal models, as they form the foundation on which modern statistical causal inference methods are based.

A DAG is a graphical causal model in which all arrows are unidirectional (hence 'directed'). Moreover, no variable can indirectly cause itself (hence 'acyclic') (6, 39). As identified previously, the graphical causal model in panel 1 of Figure 2.1 is not a DAG because there exists a bidirectional arrow between X and Y, whereas the graphical causal model in panel 2 of Figure 2.1 is a DAG because the bidirectional arrow has been replaced by two unidirectional arrows emanating from the common cause U.

DAGs encode qualitative causal assumptions about the **data-generating process** in the population (39), i.e. the process by which any endogenous value in the graph obtains its value. Given information on all exogenous variables in a DAG, the values of any endogenous variable can be identified. In Figure 2.2, for example, if we know the values of A and B (the exogenous variables, which have no causes in the graph) we are able to identify the value of C, as it depends only on A for its value. Similarly, we are able to identify the values of all other endogenous variables D, E, and F, as they depend only on other variables in the graph.

Figure 2.2 DAG depicting the data-generating process for the six random variables A, B, C, D, E, and F



2.5.2.1 Key terminology

Kinship terminology is often employed to describe the relationships between variables in a DAG (39). For example, the variables which are directly caused by a given variable are called its **children** (e.g. C and D are children of A in Figure 2.2), and all variables which are directly or indirectly caused by a given variable are called its **descendants** (e.g. C, D, E, and F are descendants of A in Figure 2.2). Conversely, the variables which directly cause a given variable are called its **parents** (e.g. C and D are parents of E in Figure 2.2), and all variables which directly or indirectly cause a given variable are called its **ancestors** (e.g. A, C, and D are ancestors of E in Figure 2.2).

A **path** is a sequence of arrows connecting two variables, regardless of the orientation of those arrows; there may be multiple paths connecting any two nodes in the graph (39). For example, D and F are connected by the path $D \leftarrow C \rightarrow E \rightarrow F$ in Figure 2.2.

A **collider** on a path is a variable on a path that has two arrows pointing into it; all other variables on the path are **non-colliders** (39). For example, D is a collider on the path $C \rightarrow D \leftarrow A$ in Figure 2.2.

A **causal path** is a sequence of arrows which all flow in the same direction (39). For example, $C \rightarrow D \rightarrow E \rightarrow F$ is a causal path between C and F in Figure 2.2. Any path which connects two variables and is not a causal path is a **non-causal path**, of which backdoor paths are of particular interest.

A **backdoor path** is a path from one variable to another which begins with an arrow *into* the first variable (39). For example, the path $C \leftarrow A \rightarrow D \rightarrow E \rightarrow F$ is a backdoor path between C and F in Figure 2.2.

Conditioning refers to the act of filtering a dataset based on the values of one or more variables (27).

A path between two variables is **open** if it does not contain a collider, or if it contains a collider which has been conditioned upon (39). An open path therefore transmits a statistical association between the two variables (27). For example, the backdoor path $C \leftarrow A \rightarrow D \rightarrow E \rightarrow F$ between C and F in Figure 2.2 is open. Conversely, a path between two variables is **closed** if it contains a collider, or if it contains a non-collider which has been conditioned upon (39). A closed path transmits no statistical association between the two variables (27). The backdoor path $C \leftarrow A \rightarrow D \rightarrow E \rightarrow F$ between C and F in Figure 2.2 can be closed if we condition on any of A, D, or E.

2.5.2.2 Direct, indirect, and total causal effects

For any two variables in a DAG, there are potentially three types of **causal effects** of the first (the 'exposure') on the second (the 'outcome') which may be of interest.

The **direct causal effect** represents the change in the outcome that results directly from changing the exposure (27). This is signified by the existence of an arrow that directly connects the exposure and outcome.

An **indirect causal effect** represents the change in the outcome that results from changes to the exposure which are passed through one or more other variables (27). This is signified by the existence of a *causal pathway* between the exposure and outcome.

The **total causal effect (TCE)** comprises all of the direct and indirect causal effects (i.e. all of the direct and indirect causal pathways between the exposure and outcome) (27).[5]

To demonstrate, we consider Figure 2.2, in which C is the exposure and F is the outcome. There is no direct causal effect of C on F, since there is no arrow from C to F. However, there are two indirect causal effects of C on F, as indicated by the two causal pathways $C \rightarrow D \rightarrow$

[5] In a linear system, the total causal effect is simply a sum of the direct and indirect effects, though this does not hold in a non-linear system (27, 40).

$E \rightarrow F$ and $C \rightarrow E \rightarrow F$. The total causal effect of C on F thus comprises just the combination of two indirect causal effects.

2.5.2.3 Variable roles

The role(s) of any variable in the DAG are defined with respect to its relationship with the exposure and outcome and outcome of interest (41, 42). In the following paragraphs, we define all other potential roles that a variable in a DAG may have. We also provide an example of each role by considering the DAG in Figure 2.2, in which C is the exposure and F is the outcome.

A **confounder** is a variable which is an ancestor of both the exposure and outcome (i.e. a preceding common cause) (41). A confounder transmits a *non-causal* association between the exposure and outcome via a backdoor path. In Figure 2.2, for example, A is a confounder of the relationship between C and F. A transmits a non-causal association between C and F via the backdoor path $C \leftarrow A \rightarrow D \rightarrow E \rightarrow F$.

A **mediator** is a variable which is a descendant of the exposure and an ancestor of the outcome (i.e. a variable which lies on the causal pathway between the exposure and outcome) (41). A mediator transmits part of the causal association between the exposure and outcome. In Figure 2.2, both D and E are mediators of the causal effect of C on F. D transmits a causal association between C and F via the causal path $C \rightarrow D \rightarrow E \rightarrow F$, whereas E transmits a causal association via the causal paths $C \rightarrow D \rightarrow E \rightarrow F$ and $C \rightarrow E \rightarrow F$.

A **proxy confounder** is a variable which is a descendant of a confounder and an ancestor of either the exposure or the outcome (but not both, otherwise if would be a confounder) (41). A proxy confounder thus does not itself transmit a non-causal association between the exposure and outcome, but it may be thought of as an imperfect measure of the true confounder. In Figure 2.2, D is a proxy confounder of the relationship between C and F, since it is a descendant of the confounder A and an ancestor of the outcome F.

A **competing exposure** is a variable which is an ancestor of the outcome but is unrelated to the exposure (i.e. it is neither a confounder, mediator, or proxy confounder) (41). A competing exposure therefore represents an independent cause of the outcome. In Figure 2.2, B is a competing exposure for the causal effect of C and F, since it is a parent of F but it does not affect (nor is it affected by) C.

It is possible for a variable to have multiple roles in a single DAG. For example, when considering the exposure-outcome relationship between C and F in Figure 2.2, D acts both as a mediator and a proxy confounder (for the confounder A).

2.5.3 Average causal effects for time-fixed exposures

To illustrate how DAGs can be used to identify the causal effect of *a time-fixed* exposure on an outcome, we return to the previous context involving the effect of chemotherapy versus

radiotherapy on two-year survival amongst breast cancer patients (§2.4.1). Instead of randomised data, however, we only have data collected by hospitals.

In this dataset, it is unlikely that the group of individuals who received chemotherapy are equivalent to (i.e. exchangeable with) those who received radiotherapy. For example, chemotherapy drugs are *less* likely to be administered to individuals who are taking medications for other conditions in order to minimise the risk of adverse interactions (43), but *more* likely to be administered to individuals whose cancer is at an advanced stage and grade at the time of treatment (44).

Such imbalances between the groups receiving chemotherapy and radiotherapy mean that their average outcomes cannot be directly compared. However, this scenario is not in principle far removed from the scenario involving a CRCT. That is, if we are able to identify the causes of two-year survival which affect the treatment received, we can conceptualise the exposure as having been randomised within subgroups defined by those factors and identify the average causal effect within them. For example, we could compare average two-year survival between those who received chemotherapy and those who received radiotherapy among individuals who were taking no other medications and whose cancer was classed as stage IV and grade III. In this way, the two treatment groups are **conditionally exchangeable**, i.e. they are exchangeable *conditional on* the factors which affected the treatment received.

The power of graphical model theory is that it provides a way of determining which set(s) of variable(s) are sufficient for guaranteeing conditional exchangeability for a given DAG through a simple graphical criterion – the backdoor criterion.

2.5.3.1 The backdoor criterion

The **backdoor criterion** is actually a set of three criteria, which, if met, guarantee conditional exchangeability for a given DAG, thereby allowing for the identification of the average *total causal effect* of a time-fixed exposure on an outcome (27). Briefly, a set of variables is sufficient for guaranteeing conditional exchangeability if conditioning on those variables:

1. Closes all non-causal paths between the exposure and outcome;
2. Does not close any causal paths between the exposure and outcome; and
3. Does not open any additional non-causal paths between the exposure and outcome (6, 27).

There may be multiple sets of variables in a given DAG which satisfy these criteria, or none. A set is said to be minimally sufficient (i.e. a **minimally sufficient adjustment set**, or MSAS) if it satisfies the backdoor criterion with the smallest number of variables (39). Variable sets which satisfy the backdoor criterion may be identified algorithmically (e.g. using the 'dagitty' software (45) or R package (46, 47)), and it is for this reason that DAGs have heralded as tools which facilitate the 'algorithmisation of counterfactuals' (48).

2.5.3.2 A graphical representation of exchangeability

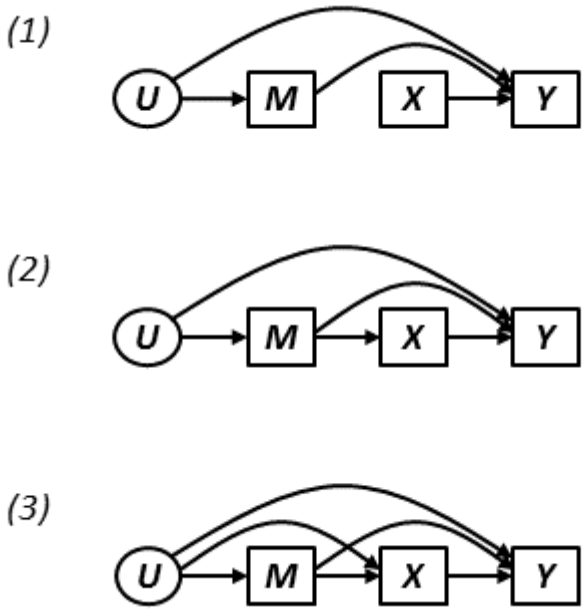
Figure 2.3 provides a graphical depiction of exchangeability, in which X represents a time-fixed exposure, Y represents an outcome, M represents a set of measured baseline causes of Y, and U represents a set of unmeasured and/or unknown baseline causes of Y (26).

In panel (1), *unconditional* exchangeability holds, since no baseline causes of the outcome (either measured or unmeasured) affect the exposure.

In panel (2), *conditional* exchangeability holds, since only measured baseline causes of the outcome affect the exposure (i.e. exchangeability can be created by conditioning on the variables in M). It is for this reason that the conditional exchangeability is sometimes referred to as the condition of '*no unmeasured confounding*' (26).

In panel (3), exchangeability does not hold, since unmeasured and/or unknown baseline causes of the outcome affect the exposure.

Figure 2.3 DAG depicting the data-generating process for a time-fixed exposure X, an outcome Y, a set M of measured baseline causes of the outcome, and a set U of measured and/or unknown baseline causes of the outcome



In panel (1), unconditional exchangeability holds. In panel (2), conditional exchangeability holds. In panel (3), exchangeability does not hold. Figure is adapted from Robins, J.M. and M.A. Hernán (26).

2.5.3.3 Other 'identifiability conditions'

Two other conditions are required to identify the average causal effect of a time-fixed exposure on a subsequent outcome; together with conditional exchangeability, these are referred to as the **identifiability conditions** (26).

Positivity is the requirement that the exposure is not deterministically allocated within any of the subgroups defined by possible combinations of the measured baseline covariates (26). Recall that, in principle, average causal effects are identified by comparing the average

outcomes within subgroups for which the distribution of confounding factors is equivalent. If there exist one or more subgroups in which every individual in the subgroup received the same value of the exposure (i.e. the exposure was deterministically allocated), then we cannot identify the causal effect of the exposure in that subgroup because there exists no 'counterfactual' scenario.

Consistency is the requirement that, for an individual who received a particular value of the exposure, their counterfactual outcome is equal to their observed outcome and is therefore known (though their other counterfactual outcome(s) remain(s) unknown) (26). Consistency may not hold for exposures comprised of a combination of various factors which affect the outcome through different mechanisms (e.g. socioeconomic position), since such an exposure would likely have multiple counterfactual outcomes associated with a single value of the exposure (49).[6] Inherent in the consistency condition is the assumption of **no interference**, i.e. the assumption that the exposure received by one individual does not affect the outcomes of any other individual (54).[7]

Taken together, the three conditions (i.e. conditional exchangeability, positivity, and consistency) imply that an observational study may be conceptualised as a CRCT, or an RCT in which the exposure was randomised conditional on the set of covariates which satisfy the backdoor criteria (26).

2.5.4 Average causal effects for time-varying exposures

DAGs can also be used to identify the causal effect of a *time-varying* exposure on an outcome in non-randomised contexts. As in the time-fixed case, this is achieved by using the principles of graphical model theory to determine set(s) of variables which are sufficient for guaranteeing *sequential* conditional exchangeability.

The criterion for determining which set(s) of variable(s) are sufficient for guaranteeing conditional exchangeability for a given DAG is known as the **sequential backdoor criterion**, which is a generalisation of the previously-introduced backdoor criterion (§2.5.3.1). The logic behind the sequential backdoor criterion is very similar to that of the original and for that reason we do not explicitly cover it here. However, a comprehensive summary of the sequential backdoor can be found in Elwert, F. (39).

---

[6] We note that there exists a substantive and contentious debate within the causal inference community as to whether consistency is an *assumption* or a *theorem* (50-52), and consequently whether causal effects can be identified in the absence of well-defined interventions (53). However, this is a philosophical debate which we believe has very little bearing on the topics covered in this thesis.

[7] The assumption of no interference may also be referred to the 'stable unit treatment value assumption' (SUTVA) (55).

2.5.4.1 A graphical representation of exchangeability

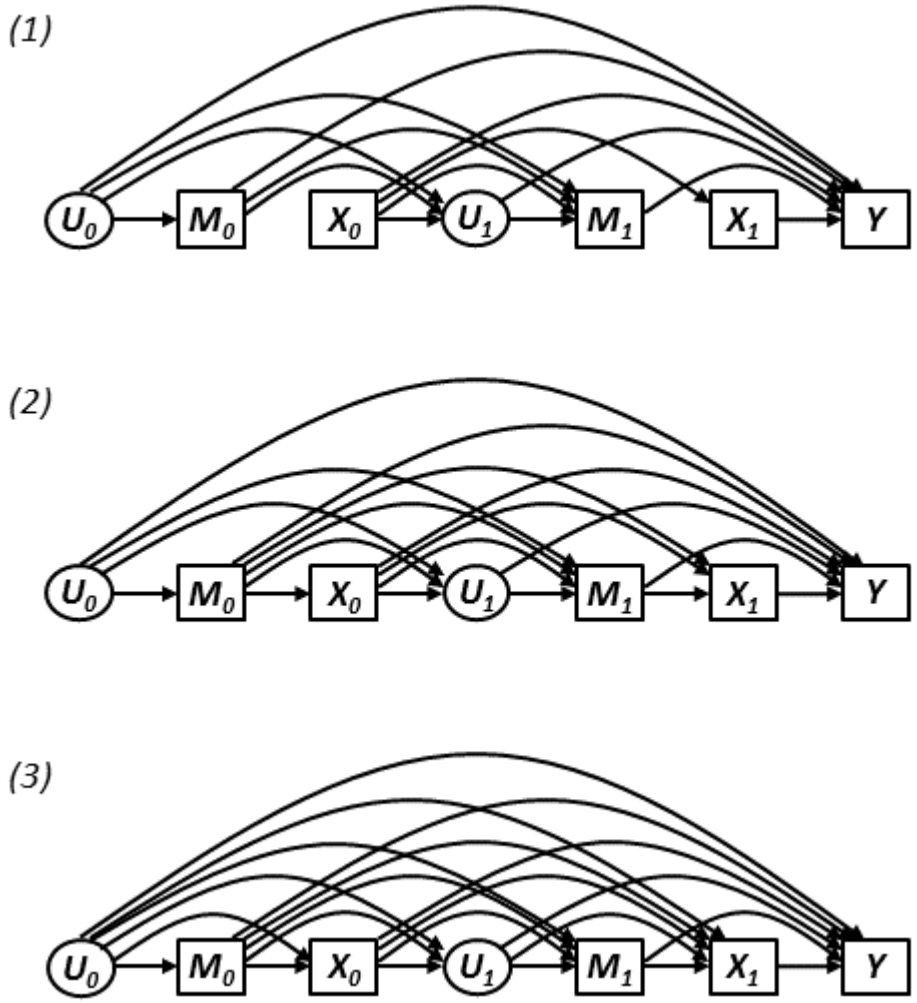
Exchangeability in a time-varying context is most easily represented graphically. Figure 2.4 provides such a representation, in which X_t represents a time-varying exposure measured at baseline (i.e. $t = 0$) and follow-up (i.e. $t = 1$), Y represents an outcome measured at or after the point of follow-up, M_t represents a set of measured causes of Y at baseline and follow-up, and U_t represents a set of unmeasured and/or unknown causes of Y at baseline and follow-up (26).

In panel (1), *unconditional* sequential exchangeability holds, since neither measured nor unmeasured/unknown causes of the outcome affect the exposure at either time point. This is signified by a lack of any arrows from M or U into X.

In panel (2), *conditional* sequential exchangeability holds, since only measured causes of the outcome affect the exposure at each time point. This is signified by a lack of any arrows from U into X.

In panel (3), sequential exchangeability does not hold, since unmeasured and/or unknown causes of the outcome affect the exposure at each time point.

Figure 2.4 DAG depicting the data-generating process for two measurements of a time-varying exposure X (i.e. X_0 and X_1), one outcome Y, two measurements of a set M of time-varying causes of the outcome (i.e. M_0 and M_1), and two measurements of a set U of time-varying unmeasured and/or unknown causes of the outcome (i.e. U_0 and U_1)



In panel (1), sequential unconditional exchangeability holds. In panel (2), sequential conditional exchangeability holds. In panel (3), sequential exchangeability does not hold.

2.5.4.2 Time-dependent confounding

Time-dependent confounding is an issue that is unique to situations involving time-varying exposures. Whilst it is conceptually challenging to understand within the counterfactual framework, the use of DAGs illustrates simply the problem posed by time-dependent confounding (39).

For example, we consider the DAG in panel (2) of Figure 2.4 where we are interested in the joint effect of X_0 and X_1 on Y. In this DAG, M_1 is a confounder for the effect of X_1 on Y, and thus the non-causal path $X_1 \leftarrow M_1 \rightarrow Y$ should be closed by conditioning on M_1. However, M_1 is also a mediator for the effect of X_0 on Y, and thus conditioning on M_1 closes the causal path $X_0 \rightarrow M_1 \rightarrow Y$. Moreover, M_1 is a collider on the path $X_0 \rightarrow M_1 \leftarrow U_1 \rightarrow Y$, such that conditioning on M_1 opens an additional non-causal path between X_1 and Y.

Thus, it is both necessary and forbidden to condition on M_1 because it simultaneously confounds and mediates the causal effect of X on Y. Estimating the joint effect of X_0 and X_1 on Y requires the use of one of the 'g-methods', which are a suite of methods which deal with time-dependent confounding in ways that do not involve direct conditioning (26, 56). The g-methods are reviewed in the next chapter.

2.5.4.3 Other 'identifiability conditions'

As in the setting of a time-fixed exposure (§2.5.3.3), two additional identifiability conditions are required to identify the average causal effect of a time-varying exposure on a subsequent outcome (26).

In a time-varying setting, **positivity** is the requirement that the exposure at each time point is not deterministically allocated within any of the subgroups defined by possible combinations of past exposure and covariate history (26). In other words, there is a non-zero chance of being exposed (or unexposed) at every time point, regardless of prior exposure and confounder. Thus, positivity is satisfied when there are both exposed and unexposed individuals within all levels of prior exposure and confounders, which can easily be evaluated empirically (for categorical variables, at least) (56).

Consistency is the requirement that, for an individual who received a particular exposure regime, their counterfactual outcome is equal to their observed outcome and is therefore known (though their other counterfactual outcomes remain unknown) (26). This condition also includes the assumption of **no interference**.

Taken together, the three conditions (i.e. conditional exchangeability, positivity, and consistency) imply that an observational study may be conceptualised as a *sequential* CRCT, or an RCT in which the exposure at each time point was randomised conditional only on prior exposure and measured covariate history (26).

2.6 Summary

The counterfactual framework for causal inference underpins much of health and social science research. Although it is impossible to identify individual-level causal effects in this framework (which is often referred to as the 'fundamental problem of causal inference'), it *is* possible to identify average causal effects. This has historically been achieved through randomisation, though recent advances in graphical model theory have provided a framework for identifying causal effects by emulating randomisation in observational data. Of particular importance are DAGs, which encode counterfactual statements in simple statistical diagrams. Time-varying exposures present additional identification challenges over and above those of time-fixed exposures due to the issue of time-dependent confounding. The utility of using DAGs to estimate causal effects and the additional methodological challenges presented by longitudinal data structures will be expanded upon in the next chapter, in which we review and critically compare statistical- and simulation-based approaches.

Chapter 3
Methods for estimating causal effects in longitudinal data

3.1 Introduction

There exist myriad methods for estimating causal effects in longitudinal data, some of which are grounded in the principles of graphical model theory but many of which are not.

Chapter 3 introduces several methods for estimating causal effects in longitudinal data. Of fundamental importance are statistical, regression-based methods which are informed by DAGs; these methods utilise the principles of graphical model theory to robustly estimate counterfactual quantities. There additionally exist individual-based simulation methods (i.e. microsimulation modelling and agent-based modelling) which are able to simulate counterfactuals; however, the conditions under which they provide meaningful causal effect estimates are not well-understood. This chapter offers a critical comparison of statistical and individual-based simulation methods for causal inference, which provides a foundation for understanding the contexts in which the methods considered in the remainder of this thesis may be used. This chapter also provides several examples of the benefits of using DAGs to consider problems and paradoxes which have historically arisen in causal analyses, thereby providing a basis for our aim to integrate DAGs and counterfactual thinking into the methods considered in the remainder of this thesis.

3.1.1 Chapter overview

A general chapter overview is provided below.

In Section 3.2, we demonstrate how DAGs can be used to inform statistical (regression) models in order to estimate causal effects in observational data, for both time-fixed (§3.2.1) and time-varying exposures (§3.2.2).

In Section 3.3, we give three specific examples which illustrate the benefits of applying DAGs and counterfactual thinking to new contexts.

In Section 3.4, we introduce three established methods for longitudinal data analysis (both statistical- and simulation-based) which will be examined in this thesis. The methods considered are the analysis of change (§3.4.1), regression with 'unexplained residuals' (§3.4.2), and microsimulation modelling (§3.4.3).

In Section 3.5, we critically compare statistical- and individual-level simulation-based approaches for causal inference.

3.1.2 Related publications

This chapter contains work based on the following publications:

Arnold, K.F., Berrie, L., Tennant, P.W.G. and Gilthorpe, M.S. A causal inference perspective on the analysis of compositional data. *International Journal of Epidemiology.* 2020, 0(0), pp.1-7. (1)

Arnold, K.F., Davies, V., de Kamps, M., Tennant, P.W.G., Mbotwa, J. and Gilthorpe, M.S. Reflections on modern methods: Generalised linear models for prognosis and intervention – theory, practice, and implications for machine learning. *International Journal of Epidemiology.* 2020, 0(0), pp.1-9. (2)

Arnold, K.F., Harrison, W.J., Heppenstall, A.J. and Gilthorpe, M.S. DAG-informed regression modelling, agent-based modelling and microsimulation modelling: a critical comparison of methods for causal inference. *International Journal of Epidemiology.* 2019, 48(1), pp.243-253. (3)

3.2 DAG-informed regression methods

As introduced previously (§2.5.2), a DAG is a qualitative (i.e. *nonparametric*) map of the data-generating process for a set of variables (39). For any given DAG, the principles of graphical model theory provide a way of determining whether a causal effect can be identified and, if so, what set(s) of variables need to be conditioned on to do so.

Where the true structure of a DAG is not known, as in almost all observational contexts, its structure must be assumed based upon subject matter knowledge and theories, and then tested and further refined according to available data (47, 57). In this way, the DAG represents the *hypothesised* data-generating process, and all inferences are made subject to the DAG being correct.

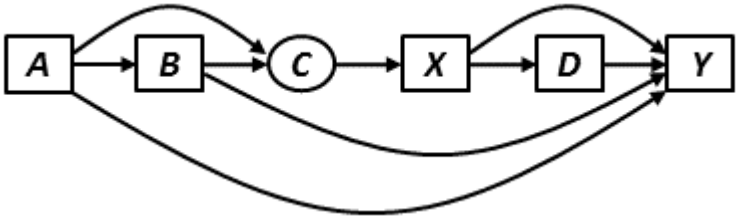
This DAG may then also be combined with *parametric* assumptions about the data-generating process in order to estimate causal effects. The primary method for achieving this is through regression modelling. In the following subsections, we outline how DAG-informed regression modelling can be implemented in order to estimate causal effects in observational data, for both time-fixed (§3.2.1) and time-varying exposures (§3.2.2).

Throughout, we use capital letters (e.g. Y) to denote random variables and small letters to denote specific values (e.g. $y = 0$ or $y = 1$), by convention (26).

3.2.1 For time-fixed exposures

To illustrate, we consider the DAG in Figure 3.1, which represents the hypothesised data-generating process for a time-fixed exposure X, outcome Y, confounders A, B, and C, and mediator D in a population of individuals (all continuous random variables).

Figure 3.1 DAG depicting the hypothesised data-generating process for a time-fixed exposure X, an outcome Y, a set of confounders A, B, and C, and a mediator D



Observed and/or measured variables are depicted in rectangular boxes, and latent variables are depicted in ovals.

By the backdoor criterion (§2.5.3.1), there exist two sets of variables which are minimally sufficient for identifying the total causal effect of X on Y:

Set 1: A and B

Set 2: C

Therefore, conditioning on either of these sets of variables will allow us to estimate the desired total causal effect. However, given that C is unmeasured, *Set 1* would be chosen as the conditioning set.

In the context of linear regression, conditioning is achieved by including the variable as a covariate in the model. Estimating the total causal effect of X on Y in our example context thus becomes a matter of estimating the parameters of the following model:

$$Y = \beta_0 + \beta_1 X + \beta_2 A + \beta_3 B + \varepsilon$$

Assuming the model has been correctly parameterised, we are able to interpret $\hat{\beta}_1$ as the estimated total causal effect of X on Y. In other words, for individuals with the same values of A and B (i.e. *conditionally exchangeable groups*), every one-unit difference in the exposure corresponds to an expected difference in the outcome of $\hat{\beta}_1$.

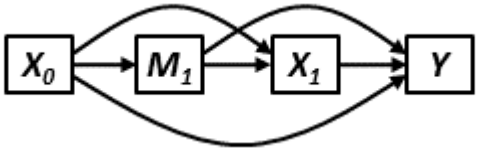
The expected counterfactual outcome associated with a particular value x of the exposure for an individual whose values of A and B were equal to a and b, respectively, can thus be computed as:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 a + \hat{\beta}_3 b$$

3.2.2 For time-varying exposures

We next consider the DAG in Figure 3.2, which represents the hypothesised data-generating process for two measurements of a time-varying exposure X (i.e. X_0 and X_1), one subsequent outcome Y, and one time-dependent confounder M_1 (all continuous random variables) in a population of individuals.

Figure 3.2 DAG depicting the hypothesised data-generating process for two measurements of a time-varying exposure X (i.e. X_0 and X_1), one outcome Y, and one time-dependent confounder M_1



The joint causal effect of X_0 and X_1 on the outcome Y is identifiable by the sequential backdoor criterion (39). However, simultaneously conditioning and not conditioning on M_1 is impossible in a conventional single-equation regression model (39). Thus, one of the three 'g-methods' may be used to estimate the average counterfactual outcomes associated with different exposure regimes. Each g-method is summarised in the following subsections; more detailed descriptions are provided by Robins, J.M. and M.A. Hernán (26), Naimi, A.I. et al. (56), Daniel, R.M. *et al.* (58), Arnold, K.F. and M.S. Gilthorpe (59), Taubman, S.L. *et al.* (60), Robins, J.M. *et al.* (61), Vansteelandt, S. and M. Joffe (62), and Picciotto, S. and A.M. Neophytou (63).

3.2.2.1 The (parametric) g-formula

Implementing the parametric g-formula requires that we first use our data to estimate the functions which govern the data-generating process, thereby creating a sequence of functions which combine to generate the values for every endogenous node in the DAG.[8] For example, if we assume a linear process, we would estimate the parameters for each of the following models:

$$M_1 = \beta_0^0 + \beta_1^0 X_0 + \varepsilon_{M_1}$$

$$X_1 = \beta_0^1 + \beta_1^1 X_0 + \beta_2^1 M_1 + \varepsilon_{X_1}$$

$$Y = \beta_0^2 + \beta_1^2 X_0 + \beta_2^2 M_1 + \beta_3^2 X_1 + \varepsilon_{Y_1}$$

Estimating the average value of Y *that would have been observed* if the exposures X_0 and X_1 had been equal to whatever values we are interested in (e.g. x_0 and x_1, respectively) therefore requires replacing X_0 with x_0 and X_1 with x_1 in our estimated models and sequentially computing the expected value of each variable, as in:

$$\widehat{M}_1 = \hat{\beta}_0^0 + \hat{\beta}_1^0 x_0$$

$$X_1 = x_0$$

$$\hat{Y} = \hat{\beta}_0^2 + \hat{\beta}_1^2 x_0 + \hat{\beta}_2^2 \widehat{M}_1 + \hat{\beta}_3^2 x_1$$

[8] In low-dimensional settings with discrete data, the conditional probability of each variable may be estimated nonparametrically; in such cases, this method is simply referred to as 'the g-formula' (64).

The g-formula thus effectively simulates the joint distribution of the variables that *would have been observed under a hypothetical intervention targeting the exposure*, based on the joint distribution that was actually observed (6).

3.2.2.2 Inverse probability of treatment weighting (IPTW) of marginal structural models

The second g-method uses *weighting* instead of conditioning to estimate the average counterfactual outcome associated with different exposure regimes.

Inverse probability of treatment weighting (IPTW) refers to the process of creating a 'pseudo-population' by estimating the expected value of each measurement of the exposure conditional on previous exposure and confounding history in the whole sample, calculating the expected value of each measurement of the exposure for each individual, and then weighting each individual by the inverse of their expected value of each measurement of the exposure.

For example, based on the DAG in Figure 3.2 and assuming linearity, we would first estimate the parameters of the following models:

$$X_0 = \alpha_0^0 + \varepsilon_{X_0}$$

$$X_1 = \alpha_0^1 + \alpha_1^1 X_0 + \alpha_2^1 M_1 + \varepsilon_{X_1}$$

For any individual, we can then calculate the expected value of X_0, and the expected value of X_1 when $X_0 = x_0$ and $M_1 = m_1$ as:

$$\hat{X}_0 = \hat{\alpha}_0^0$$

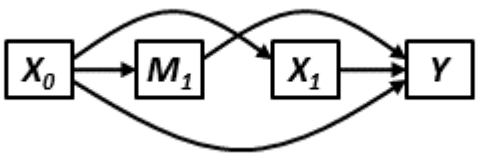
$$\hat{X}_1 = \hat{\alpha}_0^1 + \hat{\alpha}_1^1 x_0 + \hat{\alpha}_2^1 m_1$$

Each individual's weight (w) is then calculated by multiplying the inverse of their expected X_0 by the inverse of their expected X_1, i.e.:

$$w = \frac{1}{\hat{X}_0} \cdot \frac{1}{\hat{X}_1}$$

In the resulting pseudo-population, the counterfactual mean associated with each exposure regime is equal to that in the true population, but the exposure at each time point depends only on prior exposure history (i.e. *there is no time-dependent confounding*). The DAG for the pseudo-population is depicted in Figure 3.3, in which there is no arrow between M_1 and X_1.

Figure 3.3 DAG depicting the pseudo-population created by inverse probability of treatment weighting (IPTW) for the DAG in Figure 3.2



IPTW creates a pseudo-population in which there exists no time-dependent confounding (i.e. there is no arrow between M_1 and X_1).

Because there exists no time-dependent confounding in the pseudo-population, the joint effect of X_0 and X_1 on Y can be estimated by estimating the parameters of a single model:

$$Y = \beta_0 + \beta_1 X_0 + \beta_2 X_1 + \beta_3 X_0 X_1 + \varepsilon_Y$$

In the above 'marginal structural model', β_1 represents the average effect of X_0, β_2 represents the average effect of X_1, and β_3 represents the average additional joint effect of X_0 and X_1.

The average value of Y *that would have been observed* if the exposures X_0 and X_1 had been equal to whatever values we are interested in (e.g. x_0 and x_1, respectively) is therefore:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_0 + \hat{\beta}_2 x_1 + \hat{\beta}_3 x_0 x_1$$

3.2.2.3 G-estimation of structural nested models (SNMs)

The condition of sequential conditional exchangeability underlies causal inference for time-varying exposures, as outlined in Chapter 2. Moreover, the conceptualisation of longitudinal data as arising from a 'nested' sequence of trials is the foundation for g-estimation, which exploits conditional exchangeability to estimate average counterfactual outcomes (63).

Heuristically, the idea is to estimate the average effect of the exposure for the innermost (most recent) trial first (i.e. the average effect of X_1 on Y), while adjusting for past exposure and covariate history (i.e. X_0 and M_1, respectively). The estimated effect of X_1 is then removed from Y, and the process is repeated for X_0. Ultimately, the average counterfactual outcome associated with the exposure regime x_0, x_1 is computed.

For the DAG in Figure 3.2 and assuming linearity, for example, we could construct the following two structural nested models (SNMs):

$$Y = \beta_0^1 + \beta_1^1 X_1 + \beta_2^1 X_1 M_1 + \beta_3^1 X_1 X_0 + \beta_4^1 X_1 M_1 X_0 + \varepsilon_Y^1$$

$$Y = \beta_0^2 + \beta_1^2 X_0 + \varepsilon_Y^2$$

G-estimation refers to the method by which the parameters of the above models are estimated. The first model expresses the average effect of X_1 on Y, which may be modified by X_0 and M_1. The second model expresses the average effect of X_0 on Y, when the exposure at time 1 is set to some counterfactual value of interest (i.e. $X_1 = x_1$).

Sequential conditional exchangeability implies that the counterfactual outcome associated with a particular exposure regime x_0, x_1 is independent of the exposure regime that was actually observed. G-estimation directly leverages this assumption by determining the parameters for which the counterfactual outcomes are statistically independent of the observed exposures. In practice, this often involves a grid search or optimisation algorithm (63).

3.3 Examples of the benefits of DAG-based counterfactual thinking

For the identification and estimation of quantifiable causal effects, DAGs hold a privileged position due to the rigorous mathematical framework that underpins them (65).[9] However, the utility of DAGs extends far beyond the narrow framework of identifying and estimating causal effects. DAG-based thinking has also been instrumental in providing clarity to previously unresolved 'paradoxes' (66-69), for example, and in clarifying the assumptions that must hold for an association to be interpreted as a *causal* association (27, 70).

3.3.1 Example 1: Understanding the implications of conditioning on a collider

The dependency induced between two independent events when conditioning on a common outcome (i.e. a *collider*) has the potential to cause serious interpretational problems for causal analyses. Termed '**collider bias**', it often produces seemingly paradoxical results which are contrary to intuition and/or scientific feasibility (e.g. the Monty Hall problem (27)). However, there are circumstances in which conditioning on a collider may provide useful and informative causal effect estimates.

The implications of conditioning on a collider can be easily understood and illuminated using the framework of DAGs. We first discuss the birthweight paradox, which is considered to be one of the most well-known examples of collider bias. We then discuss compositional data, which present a unique context in which conditioning on a collider may be desirable (i.e. *not bias*).

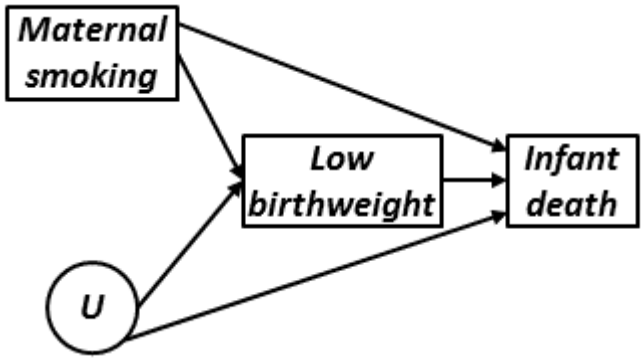
3.3.1.1 The birthweight paradox

Low birthweight has long been recognised as a strong factor which increases the risk of neonatal and infant mortality (71). Moreover, maternal smoking during pregnancy increases the risk of an infant being born with low birthweight (72). However, paradoxically, among low birthweight babies, infant mortality is substantially lower for mothers who smoke compared to those who do not smoke. Thus, among low birthweight babies, maternal smoking appears to have a *protective effect* on the risk of infant mortality. This paradox was first noted by Yerushalmy, J. (73) in 1971 and has been consistently replicated in other datasets (74).

Many hypotheses have been put forward attempting to explain the seemingly paradoxical results, e.g. (75), but perhaps the most compelling is grounded in graphical causal models (74, 76-79). To demonstrate, we represent the situation by the DAG in Figure 3.4, in which all unknown causes of low birthweight are represented by U.

[9] Nevertheless, it is worth mentioning that DAGs are currently underutilised in health research; a recent review by Tennant, P.W. et al. (57) of studies published between 1999 and 2017 identified only 234 which used a DAG to guide their analysis.

Figure 3.4 Directed acyclic graph (DAG) depicting the 'birthweight paradox'



U represents all unknown common causes of low birthweight and infant death.

In this DAG, it is apparent that low birthweight is a *collider* on the (closed) path $Maternal \rightarrow Low\ birthweight \leftarrow U$. That is, maternal smoking and U are competing causes of low birthweight. However, conditioning on low birthweight (as Yerushalmy did) opens the path between maternal smoking and U and thus creates a non-causal dependency between the competing causes of low birthweight.

Heuristically, what this means is that if a low birthweight baby has a mother who smokes, his/her low birthweight is likely a consequence of that smoking rather than a consequence of another serious unobserved cause of infant death. It is for this reason that maternal smoking then appears to be protective against infant death despite there being no obvious causal explanation.

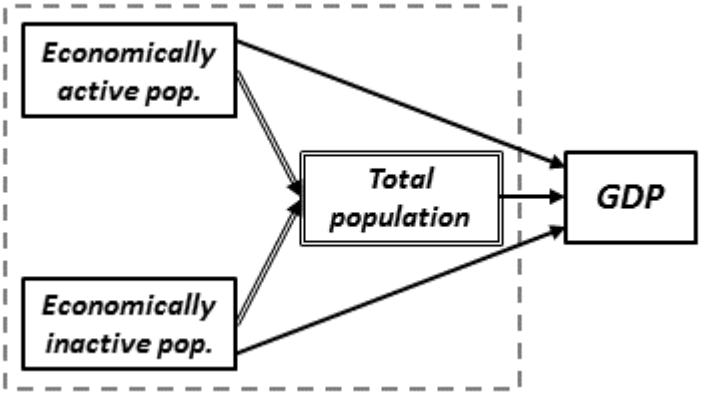
3.3.1.2 Relative versus total effects in compositional data

Although the consequences of conditioning on a collider are now well-recognised in specific probabilistic instances, such as the birthweight paradox, the consequences of doing so in deterministic instances are only beginning to be recognised. To illustrate, we consider the case of compositional data, where conditioning on a collider may in fact provide useful causal effect estimates. This example represents additional work carried out by the author of this thesis, and has been recently published in its entirety (1).

Compositional data comprise the parts (or 'components') of some whole (or 'total'), for which all parts sum to that whole (80). For example, suppose we are interested in the causal effect of the total number of economically active individuals within a geographical area on the area-level gross domestic product (GDP). This can be represented by the DAG in Figure 3.5, which explicitly depicts the compositional nature of the exposure (i.e. economically active population + economically inactive population = total population).[10]

[10] Although the components together determine the total, no time flow is indicated by the double arcs from the components to the total. Compositional data are unique in that the component parts and the total – which denote the same variable at different levels of aggregation – occur *simultaneously*. This is an additional complexity that we do not discuss here, but has been addressed in Arnold et al. (81).

Figure 3.5 DAG depicting total population in relation to gross domestic product (GDP), in which total population is subdivided into economic activity and inactivity



Deterministic relationships (i.e. total population = economically active population + economically inactive population) are indicated by double-lined arrows, and fully determined nodes are indicated by double-outlined rectangles; this notation has been adapted from Shachter, R.D. (82). A dashed box around variables indicates that those variables occur at an instantaneous point in time.

The benefit of depicting compositional data as in Figure 3.5 is that total population immediately becomes recognisable as a *collider* on the (closed) path $Economically\ active\ pop \rightarrow Total\ population \leftarrow Economically\ inactive\ pop$. Thus, conditioning on the total population when estimating the effect of the economically active population on GDP will create a dependency between the economically active and inactive populations, thereby generating a *relative effect*.

The **relative effect** of the economically active population represents the average change in GDP achieved by *swapping* economically inactive individuals for economically active individuals. This effect is therefore a *joint* effect, representing the *combined* effect of simultaneously increasing the economically active population while decreasing the economically inactive population by equal numbers (thereby retaining the same total population).

By contrast, the **total effect** of the economically active population (i.e. without conditioning on the total population) represents the average change in GDP that results from *adding* economically active individuals to the area, thereby increasing both the number of economically active individuals and the total number of individuals, whilst doing nothing to the population of economically inactive individuals.

In this scenario, both the relative and total effects reflect the population-level average effects of changing the relative numbers (i.e. the proportions) of economically active individuals to alter GDP, but by different means. It is therefore possible derive two distinct causal quantities, each of which may be of interest depending on the context or hypothetical intervention. For example, an estimate of the relative effect may be of interest if the government were considering job-training programmes for currently unemployed individuals, whereas an

estimate of the total effect may be of interest if the government were considering policies aimed at increasing economic immigration.

In general, whether the relative or total effect represents a useful estimand dependents upon the context and, in particular, the number of components which make up the total.

3.3.2 Example 2: Understanding the distinction between prediction and causal inference

Causal inference is fundamentally distinct from prediction (83, 84). However, the two are often conflated since many of the same statistical techniques (e.g. linear models) can be applied to both predictive and causal queries (70). Moreover, the relative newness of a formal framework for causal inference from observational data meant that the routine application of such techniques that became embedded was predicated on *prediction*, rather than *causal inference*.

Consequently, it remains common practice to endow certain 'predictive' variables with causal significance, either explicitly or implicitly (e.g. (85)). The framework of DAGs offers insight into the dangers of doing so by explicating the assumptions required to interpret *associations* between individual predictors and the outcome as *causal effects*.

To illustrate, we consider the following linear model, which represents the expected value a single variate Y (the 'dependent' or 'outcome' variable) from a linear combination of a set of observed covariates $X_1, \dots, X_n$ (the 'independent' or 'explanatory' variables, or simply 'predictors'):

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_n X_n \qquad \textbf{Equation 3.1}$$

Any coefficient in Equation 3.1 could potentially represent a true causal effect (either direct, total, or a subset of the total), an association due to uncontrolled confounding or collider bias, or any combination thereof. Interpreting a particular coefficient as an estimate of the total causal effect of that covariate on the outcome requires making the assumption that all other covariates in the model 'control for' all spurious associations, do not 'control for' any of the causal association, and do not create any additional spurious associations.[11] Causal modelling processes have these assumptions explicitly built into their foundations, but prediction modelling processes do not.

A model for prediction is concerned with optimally deriving the likely *value* (or *risk*) of an outcome (i.e. Y in Equation 3.1) given information from one or more 'predictors'. The goal of prediction modelling is to develop a useful tool to forecast an outcome that has yet to occur, and so the model-building process is ultimately driven by convenience and other practical considerations. It is well-suited to automated methods for covariate selection and

[11] This is closely related to the fallacy of 'mutual adjustment' – often referred to as the 'Table 2 fallacy' (86) – since interpreting *every* coefficient in the model as a total causal effect requires making these assumptions about all other covariates in the model simultaneously, which cannot be valid unless all covariates are causally unrelated (i.e. 'competing exposures') and orthogonal (i.e. uncorrelated in the sample).

parameterisation, because the specific subset of covariates that is ultimately used to predict the outcome (and the way in which they are parameterised) is relatively unimportant so long as the model has a sufficient degree of internal and external validity.

In contrast, a model for causal inference is concerned with optimally deriving the likely *change* in an outcome (i.e. $\hat{\beta}_i$ for $1 \leq i \leq n$ in Equation 3.1) due to (potentially hypothetical) change in a *particular covariate* (i.e. X_i). The causal model-building process is necessarily driven by external and *a priori* theory, and thus benefits little from algorithmic modelling methodologies. To estimate the causal effect of one variable on another, one must specify both the possible causal pathways through which those effects are realised and the possible non-causal pathways that transmit spurious associations *before* any modelling is undertaken. Although the process of identifying a suitable subset of covariates which remove all spurious associations between the exposure and outcome may be automated once all causal assumptions are made explicit (often in the form of a DAG), identifying the initial set of variables and specifying the *manner* in which they are likely to transmit spurious associations cannot be automated.

Consequently, models for prediction and causal inference are fundamentally distinct in terms of their purpose and utility, and methods optimised for one cannot be assumed optimal for the other. This has important implications for more advanced modelling methods (e.g. 'machine learning' methods (87-89)), which have been developed with the goal of *prediction* in mind.

3.4 Other established methods for longitudinal data analysis

The previous examples illustrate the benefits of bringing DAGs and counterfactual thinking to new contexts and to methods which may not have been developed in an explicit causal framework. Although several authors have been critical of the increasingly widespread use of DAGs (8, 14, 15, 90, 91), the insights into many important causal questions and apparent paradoxes (like the birthweight paradox) that have been facilitated by the use of DAGs are substantial (27, 66).

This PhD explores how counterfactual thinking, encoded in the language of DAGs, can be integrated into established methods for causal longitudinal data analysis. Although DAGs provide a formal mathematical framework for the identification and estimation of causal effects, the relative recentness of such developments has meant that many established methods for causal analysis have not been considered within a robust causal framework. The methods to be considered are both regression-based and simulation-based. The regression-based methods considered are the analysis of 'change' and the method of 'unexplained residuals' (UR) models; the simulation-based method considered is microsimulation modelling.

Each is summarised in the following subsections.

3.4.1 The analysis of change

Studies of change are an important element of health research. Understanding how people change, and the factors that may cause them to change more or less, are important for prognosis and treatment decisions.

There are two primary methods for analysing the effect of a baseline exposure X_0 on 'change' in Y, which is measured once at baseline (i.e. Y_0) and once at follow-up (i.e. Y_1).

The first method involves constructing a 'change score' by subtracting the baseline outcome from the follow-up outcome (i.e. $\Delta Y = Y_1 - Y_0$). The effect of the baseline exposure X_0 on the change score is then estimated using a regression model of the form:

$$\Delta Y = \alpha_0 + \alpha_1 X_0 + \varepsilon$$

The second method involves regressing the follow-up outcome Y_1 on the baseline exposure X_0 and adjusting for the baseline outcome Y_0, as in:

$$Y_1 = \beta_0 + \beta_1 X_0 + \beta_1 Y_0 + \varepsilon$$

This method is commonly referred to as an analysis of covariance (ANCOVA).

It has been recognised that these two methods produce discordant results, and therefore lead to *differing causal conclusions*, in situations in where the exposure is not randomised (92).

In Chapter 4, we use DAGs to consider the analysis of change and to resolve the historical disagreement between the two methods.

3.4.2 Regression with 'unexplained residuals'

Another common research question involving change relates to how changes in a time-varying exposure (e.g. X_0 and X_1) affect a subsequent outcome (Y).

Regression with 'unexplained residuals' (93), or UR models, are regression models which simultaneously estimate the effect on Y of the initial measurement of the exposure X_0 *and* subsequent change in X. A UR model is constructed by regressing Y on X_0 and all 'unexplained' changes in X (i.e. e_{X1}), as in:

$$\hat{Y} = \hat{\lambda}_0 + \hat{\lambda}_1 X_0 + \hat{\lambda}_2 e_{X1}$$

The term e_{X1} is itself derived from the regression of X_1 on X_0 (i.e. $X_1 = \hat{\gamma}_0 + \hat{\gamma}_1 X_0 + e_{X1}$), and it has been claimed that the model therefore provides insight (via the coefficient $\hat{\lambda}_2$) into the effect on Y of X increasing *more than expected* (93). However, this claim has been previously challenged (94).

UR models have also been extended *ad-hoc* in order to accommodate more than two measurements of a time-varying exposure, and to accommodate time-fixed and time-dependent confounding.

In Chapter 5, we use DAGs to evaluate the properties of UR models and their suitability for causal analyses. We also demonstrate how to extend the method robustly in order to accommodate confounding variables and more complex longitudinal scenarios.

3.4.3 Microsimulation modelling

Several authors have identified microsimulation models (MSMs) as being promising tools for causal inference, especially due to their ability to evaluate exposure regimes. In its most basic form, microsimulation is a method for generating micro-level data in order to provide an estimated cross-sectional snapshot of a population (95). However, the resulting synthetic population is then often used as the foundation for a dynamic simulation model, which simulates the evolution of individuals in the population through time and potentially space.

MSMs are able to evaluate counterfactual (or 'what if' (96, 97)) scenarios by, for example, altering the model parameters at one or more time points throughout the simulation, according to evidence derived from real-world interventions. The effects of these 'interventions' on some outcome of interest can then be compared and evaluated. Moreover, since the initial population remains unchanged, each simulation run may be thought of as being exchangeable with any other (98). However, the conditions under which simulation-based approaches like MSMs provide meaningful estimates of causal effects are not well understood (98). Similarities between MSMs and the g-formula have been noted previously (99), but there are several key differences which arise from their distinct historical evolutions, which are outlined in the next section.

In Chapter 6, we use DAGs to represent the microsimulation modelling process, and conduct a simulation to demonstrate the relative importance of faithfully modelling data-generating processes using microsimulation compared to the g-formula.

3.5 A critical comparison of statistical versus individual-based simulation methods for causal inference

Statistical, regression-based approaches to causal inference (or 'DAG-informed regression modelling') currently dominate epidemiological research. However, there have been growing calls for a more pluralistic approach in the field (8, 90, 91), many of them premised on the argument that there are numerous causal scenarios which do not lend themselves to statistical analysis. Many authors have proposed more widespread adoption of 'systems approaches' (21, 100-104), a somewhat nebulous term for a group of methods that may be used to study the nature of systems. In particular, several authors have identified individual-based simulation methods as promising tools for causal inference in complex systems, as they provide a framework for the simulation of counterfactuals (96, 98, 104). Microsimulation modelling (discussed previously in Section 3.4.3) is one such method; agent-based modelling is another, which is methodologically and conceptually similar to microsimulation but notably features *interactions* amongst individuals (3).

Microsimulation and agent-based modelling are historically distinct, but both have roots in cellular automata (105), which first emerged in the 1940s and involve simulating the evolution of a collection of coloured cells within a grid at discrete time steps, in accordance with a set of rules based on the states of neighbouring cells. From this, MSMs and agent-based models (ABMs) evolved separately as more complex simulation methods. While both methods have been in use for approximately the last half century – with Orcutt, G.H. (106) frequently credited as one of the founding fathers of the field of microsimulation and Schelling, T.C. (107) for agent-based modelling – the vast increases in computing power ushered in by the age of technology has rendered early implementations virtually unrecognisable in comparison to their modern counterparts (e.g. (108-111)).

Here, we offer a brief comparison of statistical and individual-based simulation approaches for causal inference. Though this thesis does not explicitly consider ABMs, their methodological and conceptual similarities with MSMs, and their ability to accommodate and model time-dependent confounding, nevertheless make them potentially important tools for causal inference.

Hernán, M.A. (112) provides a useful commentary on DAG-based regression modelling and agent-based modelling, in which he frames their differences in terms of their relative reliance on data versus theory and thus reflecting the relative value placed on data and theory within the disciplines in which they are typically used. This distinction is elaborated on in the following subsection (§3.5.1), in which we also consider the place of microsimulation modelling. We also discuss how the separate evolutions of DAG-based regression modelling, microsimulation modelling, and agent-based modelling have shaped the types of causal questions for which they are well-suited to evaluating (§3.5.2), their focus on fixed versus random effects (§3.5.3), and the timescales and timeframes upon which they generally operate (§3.5.4).

3.5.1 The relative importance of theory versus data

Epidemiology – though arguably a social science – has historically been associated with the field of medicine. Consequently, it has tended to direct greater focus towards causal questions that lend themselves to experimentation (112). Even when such experimentation is infeasible, large quantities of observational, individual-level data are collected and statistical methods (e.g. regression models) are employed with the aim of mathematically controlling for those factors which would typically be controlled via experimental manipulation. The recent advances in graphical model theory have provided the theoretical foundations for causal data analysis that had historically been lacking, but it nevertheless remains that epidemiology is a data-loving science.

In contrast, disciplines such as sociology and psychology, for example, tend to be interested in answering broader, more theory-driven questions. These often relate to phenomena for which data do not exist or may be difficult to measure or quantify (e.g. social norms); the theory-

driven, hypothesis- and data-generative nature of ABMs make them more suitable for modelling in such contexts.

Economics – the primary realm of MSMs – falls somewhere in between; indeed, the discipline has shown a degree of willingness to embrace graphical model-based methods (e.g. instrumental variable analysis (113)).

3.5.2 Research questions considered

The minimisation (albeit not elimination) of theory in the field of epidemiology has necessitated addressing narrower causal questions (112), and this is the context in which DAGs have been employed and in which the majority of methodological work is ongoing (e.g. VanderWeele, T.J. (40) and Burgess, S. *et al.* (114)). In contrast, the theory-driven, hypothesis- and data-generative nature of ABMs make them more suitable for modelling more abstract phenomena.

To illustrate how use of the methods differs, we can consider obesity as a case study.[12] Table 3.1 provides a sample of the stated research objectives for published studies which have examined obesity using DAG-informed regression modelling, microsimulation modelling, and agent-based modelling.

[12] This context has been chosen because the obesity epidemic has previously been characterised as containing many features of a complex system (21, 100, 115) and many elements from a wide variety of disciplines (e.g. biology, social policy, economics, psychology, geography, etc.); it therefore offers an ideal context for comparing the methods of interest. However, the analysis that follows is applicable to many other contexts.

Table 3.1 A sample of the stated research objectives for published studies which have examined obesity using DAG-informed regression modelling, microsimulation modelling, and agent-based modelling

DAG-informed regression modelling	Microsimulation modelling	Agent-based modelling
'… to estimate the joint effects of obesity and smoking on all-cause mortality and investigate whether there were additive or multiplicative interactions.' (116)*	'… to establish whether 52-week referral to an open-group weight-management programme would achieve greater weight loss and improvements in a range of health outcomes and be more cost-effective than the current practice of 12-week referrals.' (123)	'To explore the role that economic segregation can have in creating income differences in healthy eating and to explore policy levers that may be appropriate for countering income disparities in diet.' (130)
'… to estimate the independent causal effects of body mass index […] and physical activity on current asthma…' (117)*	'…to estimate the expected impact of the [1-peso-per-litre] tax [on sugar sweetened beverages] on body weight and on the prevalence of overweight, obesity and diabetes in Mexico.' (124)	'… [to compare] the effects of targeting antiobesity interventions at the most connected individuals in a network with those targeting individuals at random.' (131)
'… to study whether weight-related anthropometrics, changes in BMI SDS [standard deviation score] and physical activity at different ages in childhood are associated with atopic disease by late childhood.' (118)	'…to estimate changes in calorie intake and physical activity necessary to achieve the Healthy People 2020 objective of reducing adult obesity prevalence from 33.9% to 30.5%.' (125)	'… [to] simulate how a mass media and nutrition education campaign strengthening positive social norms about food consumption may potentially increase the proportion of the population who consume two or more servings of fruits and vegetables per day in NYC.' (132)
'… to estimate the 26-year risk of CHD [coronary heart disease] under several hypothetical weight loss strategies.' (119)*	'To assess the cost-utility of gastric bypass versus usual care for patients with severe obesity in Spain.' (126)	'… [to explore] the efficacy of a policy that improved the quality of neighborhood schools in reducing racial disparities in obesity-related behaviour and the dependence of this effect on social network influence and norms.' (133)
'… [to evaluate] the associations between early-life POP [persistent organic pollutant] exposures and body mass index…' (120)	'To analyse the cost-effectiveness of bariatric surgery in severely obese (BMI ≥ 35 kg/m^2) adults who have diabetes.' (127)	'… to examine: a) the effects of social norms on school children's BMI growth and fruit and vegetable (FV) consumption, and b) the effects of misperceptions of social norms on US children's BMI growth.' (134)
'… to assess the mediating role of anthropometric parameters in the relation of education and inflammation in the elderly.' (121)	'To estimate the impact of three federal policies on childhood obesity prevalence in 2032, after 20 years of implementation.' (128)	'…to examine the effects of different policies on unhealthy eating behaviors.' (135)
'… to examine differences in the contribution of obesity measures to adenoma risk by race.' (122)	'To determine the cost-effectiveness of gastric band surgery in overweight but not obese people who receive standard diabetes care.' (129)	

In the first column (DAG-informed regression modelling), '' denotes use of a g-method.*

Examination of Table 3.1 demonstrates several important distinctions, and also provides evidence to support Hernan's (112) observation that DAG-informed regression modelling and agent-based modelling exist along a spectrum according to the relative weights given to data and theory, with microsimulation modelling providing a bridge between the two.

The research questions addressed by DAG-informed regression modelling tend to be framed in terms of estimating the effect of a *specific factor* on a subsequent outcome. The concept of intervention is often implicit in these analyses (e.g. 'If we were to intervene to alter exposure to early-life persistent organic pollutions, how would this affect BMI?', as in Karlsen, M. et al. (120)), but may also be explicit, as in Danaei, G. et al. (119). In fact, the example of Karlsen, M. et al. (120) is illuminating due to its use of the g-formula, which shares certain similarities with microsimulation (99).

Researchers using microsimulation modelling tend to exclusively focus on estimating the effect of a *specific policy or intervention* on a subsequent outcome and, often, determining its cost-effectiveness (136, 137). Inherent in and integral to these analyses are specific comparisons between alternative intervention programmes, as in Kristensen, A.H. et al. (128).

This explicit evaluation of interventions crosses over to agent-based modelling, with several stated research objectives in the third column of Table 3.1 referring to specific hypothetical policy interventions. However, unique to agent-based modelling analyses is their exploration of social phenomena (e.g. economic segregation, as in Auchincloss, A.H. et al. (130), and social norms, as in Li, Y. et al. (132)) in the simulation framework. Agent-to-agent interactions often give rise to greater complexity, producing highly nonlinear and 'emergent' properties (138); consequently, ABMs are less-suited than MSMs to producing the detailed predictions often required by economists and policymakers, but arguably more-suited to modelling naturally complex social phenomena. Thus, although they share considerable overlap methodologically, microsimulation and agent-based modelling are distinct in their underlying purposes and practical utility.

3.5.3 Focus on fixed versus random effects

A natural consequence of using DAG-informed regression models is that greater focus is directed towards modelling mean structures and estimating mean (or 'fixed') effects instead of evaluating distributional properties and patterns of variation. Although DAGs describe causal processes that could potentially manifest in infinitely many parametric ways, the use of regression models to interrogate causal questions and identify (average) causal effects makes focus on the distributional properties of the variables of interest effectively redundant. However, there are myriad determinants of health and disease – particularly social ones (104) – which operate on many levels and in complex ways, about which the 'random' structures (possibly arising from individual interactions) are of equal if not greater importance than the 'fixed' ones. Such determinants may be of great interest to epidemiologists, yet statistical modelling is limited in the insights it can provide into the potential complexity of random

structures that contain spillover effects and interference. Consequently, causal questions involving such complexities have tended to be relegated to the social sciences, in which greater emphasis is placed on theory as opposed to data.

ABMs are theoretically very different from their statistical counterparts. As recognised by Oakes, J.M. (139), the outcome of interest is primarily the process by which group phenomena emerge. In other words, the (micro-)simulated processes of ABMs give rise to patterns and properties of a system; mean effects may be eventually derived, but the primary focus is on conceptualising and modelling the system as a whole, and how individual *agency* and *heterogeneity* interact to give rise to aggregate patterns. Although ABMs have seen some use in epidemiology, this has largely been confined to the study of infectious diseases (140-143); in such situations, there exist clear transmission mechanisms via individual interactions (144) and it is widely recognised that the effects of those interactions are a fundamental part of the causal mechanism and thus cannot be overlooked (55). Although the random effects arising from agent-to-agent interactions in ABMs are absent in MSMs, individuals remain the central focus of MSMs rather than average patterns. Indeed, in introducing the method of microsimulation, Orcutt, G.H. (106) lamented that 'current models of our socio-economic system only predict aggregates and fail to predict distributions of individuals […]'. Individual-level focus allows for the analysis of heterogeneity and distributional properties that might be masked by approaches considering only mean effects (136, 145).

3.5.4 Timescales and timeframes modelled

MSMs and ABMs tend to model much smaller timescales (e.g. days, weeks, months) than do statistical models because these are closer to the timescales upon which human behaviour and interactions generally operate, and upon which the effects of policy interventions might be realised. For ABMs in particular, in which agent-to-agent interactions are integral to the causal processes operating (e.g. for infectious diseases), modelling geolocation with high frequency is essential. Abstraction to larger scales has the potential to miss out on the complexity that these models seek to explore and/or explain; moreover, because they are not as limited by data availability, they are able to explore phenomena in finer granularity when the context requires it. Although DAG-based regression models are theoretically able to model such small timescales, their reliance on data (which has historically tended to be collected infrequently, as in cohort studies) limits this in practice.

Additionally, because DAG-based regression models are reliant upon a single dataset, they exclusively model past events; the counterfactuals represent thought experiments about what *would have happened* had some condition been different. Although public health and epidemiological are generally interested in intervening to alter future health states, DAG-based regression methods do not explicitly model this – their results must be extrapolated to infer what *would* happen in the future. In contrast, MSMs and ABMs may be used to model both past and future events by utilising and synthesising historical data and estimates to make

decisions about hypothetical future interventions; indeed, estimating the future impact of potential policy interventions has historically been fundamental to the utility of these methods (136, 145, 146).

3.6 Summary

DAGs may be combined with parametric assumptions about data-generating processes in order to estimate counterfactual quantities in non-randomised contexts. Typically, this is achieved through regression modelling, which is a methodological cornerstone of epidemiological causal inference. DAGs also have utility as conceptual tools, and have provided clarity in understanding phenomena such as collider bias. However, there are many established methods for causal analysis – both statistical and simulation-based – which have not been considered in the framework of DAGs. Individual-based simulation methods are fundamentally distinct from statistical regression-based methods, despite both being able to evaluate counterfactual quantities; these methodological differences arise from their evolution across separate research domains. In the remainder of this thesis, the integration of DAGs with both statistical- and simulation-based methods will be explored, and the utility of using DAGs to consider causal questions across different longitudinal scenarios will be demonstrated.

Chapter 4
The analysis of change

4.1 Introduction

Studies of change are a foundation for much of health research. A common target of enquiry involves quantifying the effect of a single exposure on *change* in a time-varying outcome (e.g. *'How do beta-blockers affect change in blood pressure?'*). Previous methodological research involving questions of this kind has exclusively focused on experimental contexts. However, analyses of 'change' are deceptively complex, particularly in observational data. One of the most common – yet poorly understood – challenges stems from the use and interpretation of 'change scores'.

Chapter 4 considers the analysis of change within a formal causal framework . Although studies of change are extremely common, the concept of change – and, indeed, the use of a change score as a measure of change – has received relatively limited consideration in this framework. To this end, we use DAGs to consider the concept of 'exogenous change', which is the definition of change that is of greatest utility for causal analysis and which we demonstrate is *not* isolated by the construction of a change score. We also demonstrate the utility of using DAGs to consider the different causal structures that might arise in the analysis of change. This allows us to draw conclusions about the analytical strategies most appropriate for analysing change, and highlights the importance of defining the most meaningful estimand according to the causal structure under consideration.

4.1.1 Chapter overview

A general chapter overview is provided below.

In Section 4.2, we introduce the example context that will be considered throughout the remainder of the chapter, and define the change-score analysis (§4.2.1) and the follow-up adjusted for baseline analysis (§4.2.2) in this context. Additionally, we briefly review historical perspectives on the contradictory results that these methods may provide (§4.2.3).

In Section 4.3, we argue that the estimand targeted by a change-score analysis is not useful for causal analyses, nor meaningful in most circumstances.

In Section 4.4, we use DAGs to consider the analysis of change. We consider three distinct causal structures that may arise, and we use path tracing to demonstrate the degree of discordance that may arise between a change-score analysis and a follow-up adjusted for baseline analysis in each context.

In Section 4.5, we argue that adjusting for the baseline exposure (as in a follow-up adjusted for baseline analysis) is not always appropriate, and that an unadjusted estimate may be preferred in certain scenarios.

In Section 4.6, we emphasise the importance of defining the most useful estimand on a case-by-case basis. We also present a simple simulation to illustrate the degree of inferential bias that might be introduced by a change-score analysis, and to emphasise the utility of using DAGs to determine correct adjustment.

In Section 4.7, we examine Lord's Paradox through the lens of our previous analysis.

In Section 4.8, we compare our results with those from two previous attempts to examine the analysis of change using DAGs, and explain our differences and disagreements.

In Section 4.9, we discuss the implications of our analyses.

4.1.2 Related publications

This chapter contains work conducted jointly and based on the following pre-print:

Tennant, P.W.G., **Arnold, K.F.**, Ellison, G.T.H. and Gilthorpe, M.S. Analyses of 'change scores' do not estimate causal effects in observational data. *ArXiv e-prints: 1907.02764*. 2019. (4) [13]

4.2 Methods for estimating the effect of a baseline exposure on 'change' in an outcome

Throughout, we consider a baseline exposure X_0 in relation to an outcome Y, which is measured once at baseline (i.e. Y_0) and once at follow-up (i.e. Y_1).

Historically, the issue of estimating the effect of X_0 on 'change' in Y has been framed in the context of adjusting for 'pre-existing differences' between groups (147), where the exposure X_0 represents a manipulable treatment. For example, in determining the effect of Ramipril (i.e. X_0) on blood pressure (i.e. Y_1), we would want to be certain that any effect we find is not simply the result of differences in *baseline* blood pressure (i.e. Y_0) between those who received the drug and those who did not. In other words, we would want to be certain that the effect is not *confounded*.

In the following subsections, we describe two primary methods for analysing the effect of X_0 on 'change' in Y, and briefly review historical perspectives on the apparently contradictory answers that the methods provide.

[13] This manuscript was originally submitted to the *International Journal of Epidemiology* on 5 July 2019, where it received a decision of 'revise and resubmit' on 16 March 2020. It is currently being revised in line with reviewer comments.

4.2.1 Change-score analysis

A **change-score analysis** attempts to adjust for the confounding effects of Y_0 via *offsetting* (148), i.e. by subtracting the baseline outcome from the follow-up, as in $\Delta Y = Y_1 - Y_0$.[14] The effect of X_0 on 'change' in Y is thus estimated by constructing a regression model of the form:

$$\widehat{\Delta Y} = \hat{\alpha}_0 + \hat{\alpha}_1 X_0 \qquad \textbf{Equation 4.1}$$

In this formulation, $\hat{\alpha}_1$ represents the effect of interest – the estimated effect of X_0 on ΔY.

4.2.2 Follow-up adjusted for baseline analysis

A **follow-up adjusted for baseline analysis** attempts to adjust for the confounding effects of Y_0 via *blocking* or *conditioning* (148), i.e. by examining the association between X_0 and Y_1 within levels of Y_0. The effect of X_0 on 'change' in Y is thus estimated by constructing a regression model of the form:

$$\hat{Y}_1 = \hat{\beta}_0 + \hat{\beta}_1 X_0 + \hat{\beta}_2 Y_0 \qquad \textbf{Equation 4.2}$$

In this formulation, $\hat{\beta}_1$ represents the effect of interest – the estimated effect of X_0 on Y_1, controlling for the effect of Y_0 on Y_1.

We note that a follow-up adjusted for baseline analysis is also commonly referred to as an analysis of covariance (ANCOVA) (92, 152).

4.2.3 Discordance between methods and summary of previous literature

It is widely recognised that in *randomised* contexts (i.e. where the correlation between X_0 and Y_0 is zero), both methods of analysis lead to the same, unbiased conclusions (i.e. that $\hat{\alpha}_1 = \hat{\beta}_1$) (92, 150, 153). However, in *non-randomised* contexts, the two methods provide 'contradictory results' (92), and there exists little consensus as to which is correct. Indeed, Lord's eponymous paradox centres around this very issue and the lack of an obviously correct answer (147).

Maris, E. (154), Wainer, H. (155), and Wainer, H. and L.M. Brown (156) examine both methods in the context of Lord's paradox and the potential outcomes framework. All conclude that each method makes untestable assumptions about unobserved counterfactuals and thus it is often impossible to determine which method is best. Wainer, H. (155) additionally highlights that both methods may be used to make correct *descriptive* statements, but that the 'validity of the causal inferences that naturally follow from each of these descriptive statements will all depend on different untestable assumptions.'

Allison, P.D. (157) explains that regression to the mean implies that Y_0 will usually be negatively correlated with the change score $\Delta Y = Y_1 - Y_0$, and consequently any variable (e.g. X_0) that is correlated with Y_0 will have a spurious negative relationship with ΔY; however, he goes on to argue that a change-score analysis is superior to a follow-up adjusted for baseline

[14] ΔY is typically referred to as a 'change score', and this is the language we have adopted throughout this chapter. However, change scores have alternately be known in the literature as 'difference scores' (149), 'gain scores' (148, 150), and 'change variables' (151).

analysis when X_0 is temporally subsequent to Y_0 and uncorrelated with the 'transient' component of Y_0. Van Breukelen, G.J.P. (92) acknowledges that in randomised contexts both methods are unbiased but that a follow-up adjusted for baseline analysis has more power; however, in non-randomised contexts, a change-score analysis 'seems less biased'. In contrast, Senn, S. (152) shows that baseline randomisation is not a necessary condition to estimate an unbiased effect of X_0 using a follow-up adjusted for baseline analysis; he also argues that although there may be situations in which a follow-up adjusted for baseline analysis is biased, a change-score analysis is also likely to be biased in such situations. Cronbach, L.J. and L. Furby (150) conclude even more bluntly that '[change] scores are rarely useful, no matter now they may be adjusted or refined.'

Notably, a few authors have sought to examine the analysis of change using DAGs, but unfortunately their conclusions are equally divergent as those of their predecessors. Glymour, M.M. *et al.* (158) focus on the role of measurement error and argue that follow-up adjusted for baseline analyses are likely to result in bias where measurements of Y are unreliable or unstable.[15] Kim, Y. and P.M. Steiner (148) argue that change-score analyses are immune to the potentially adverse effects of measurement error in Y_0, bias amplification, and collider bias because they do not account for Y_0 by conditioning (as in follow-up adjusted for baseline analyses). However, Shahar, E. and D.J. Shahar (151) focus exclusively on change scores, and conclude that change scores are simply 'derived variables' which have no direct causes and do not cause anything; consequently, the authors conclude that change scores are 'not of causal interest.' We will return to these authors in Section 4.8, where we re-examine their conclusions in light of our own analyses.

4.3 Considering change in a formal causal framework

Although studies of change are extremely common in health research, the use of change scores has received relatively limited consideration within a formal causal framework. Here, we use DAGs to argue that the estimand targeted by a change-score analysis is not useful for causal analyses, nor meaningful in most circumstances. This is because 'change' is fundamentally defined by the follow-up outcome only, and because change scores do not represent exogenous change – the concept of change most useful from a causal perspective. In the following subsections, we elaborate on these two arguments.

4.3.1 Change is fundamentally defined by the follow-up outcome

We first argue that 'change' is an undefined, *latent* concept that occurs post-baseline and only becomes manifest when the point of follow-up is fixed. In other words, the 'change' that occurs in a variable after the baseline cannot be defined until the point of follow-up is chosen.

---

[15] We note that Glymour, M.M. et al. (158) actually compare a change-score analysis with a change-score *adjusted for baseline* analysis, i.e. Equation 4.1 in which the baseline outcome Y_0 is additionally included as a covariate. However, this is formally equivalent to a follow-up adjusted for baseline analysis (153).

Thus, the concept of 'change' is fully encapsulated *within the follow-up variable only*. This indicates that the follow-up outcome is what should be the true target of any analysis of change. Indeed, this is most evident in the context of RCTs, in which baseline randomisation allows for fruitful comparison of average follow-up outcomes between groups.

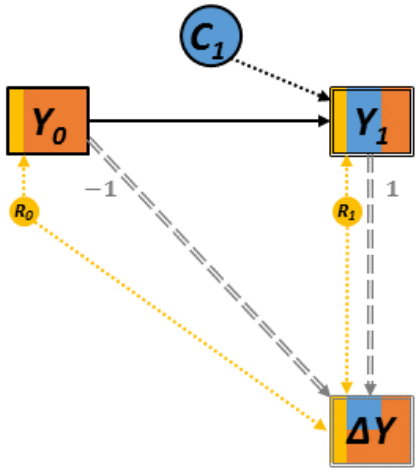
4.3.2 Change scores do not represent exogenous change

The concept of *exogenous change* is of greatest utility for causal analysis and, as we will demonstrate, this is the quantity targeted by follow-up adjusted for baseline analyses.

To illustrate the concept and utility of exogenous change, we could consider a time-varying variable like weight, which is measured at the beginning and the end of a given year in a sample of adults. While there may be limited utility in *summarising* the change that has occurred over the course of the year, from a causal perspective we are most likely interested in is the part of that change which can hypothetically be *modified* by a targeted intervention. However, each individual's weight at the end of the year is likely to be at least partly *determined* by his/her weight at the beginning of the year; that part of weight at the end of the year is therefore not modifiable and not of interest from a causal perspective. Instead, we should be interested in the part of weight at the end of the year which has *not* been determined, i.e. the part which is exogenous.

Exogenous change is the structural (i.e. non-random) part of the follow-up outcome which has not been determined at baseline, and thus may potentially be modified by intervention. This is depicted in Figure 4.1, in which the follow-up values Y_1 are partly determined by baseline values Y_0 (in orange), with the remainder determined by random features R_1 (in yellow) plus exogenous change C_1 (in blue). Here, C_1 represents all non-random change in Y that is not determined by Y_0, and thus the concept of exogenous change can be considered an average of all the processes in $C_1 \rightarrow Y$.

Figure 4.1 DAG depicting the relationship between two measurements of a longitudinal variable Y (i.e. Y_0 and Y_1) and their difference (i.e. $\Delta Y = Y_1 - Y_0$), where exogenous change (i.e. C_1) exists after baseline



The follow-up outcome Y_1 is partly determined by the baseline Y_0 (in orange), with the remainder determined by random features R_1 (in yellow) plus exogenous change C_1 (in blue). Construction of the change score ΔY does not isolate exogenous change because it conflates information from Y_0 with Y_1, whereas C_1 is fundamentally defined by Y_1. Note that the use of colours in DAGs is not commonplace, but we have introduced them here to aid understanding.

From Figure 4.1, it is clear that construction of the change score ΔY does not isolate exogenous change (i.e. C_1, in blue), since a change score represents an arbitrary linear combination of Y_0 and Y_1. Instead, C_1 is isolated by *conditioning away* the part of Y_1 that is determined by Y_0 (in orange). This is the quantity targeted by a follow-up adjusted for baseline analysis (i.e. the effect of X_0 on Y_1, *conditional on* Y_0).

4.4 Understanding analyses of change using DAGs

We now use DAGs to consider why (and the degree to which) a change-score analysis might differ from a follow-up adjusted for baseline analysis. Because the majority of past research focussing on the analysis of change has done so in an *experimental* context, the issue of *temporality* has been masked. In experimental contexts, the baseline outcome Y_0 may reasonably be assumed to occur before or at the time of the exposure X_0, but this cannot be assumed to hold in observational contexts. To this end, we use DAGs to consider three distinct potential causal structures that might arise in analyses of change:

Scenario 1: X_0 and Y_0 are causally unrelated (§4.4.1).
Scenario 2: X_0 is caused by Y_0 (§4.4.2).
Scenario 3: X_0 causes Y_0 (§4.4.3).

Because both the change-score analysis and follow-up adjusted for baseline analysis are based on linear regression, we depict each scenario as a path diagram (35, 36). A path diagram is a (linearly) parametric DAG, in which a single coefficient is assigned to every arc and all variables are constrained to have a variance of one. Throughout, we use the notation p_{ba} to represent the coefficient of the arrow $a \rightarrow b$.

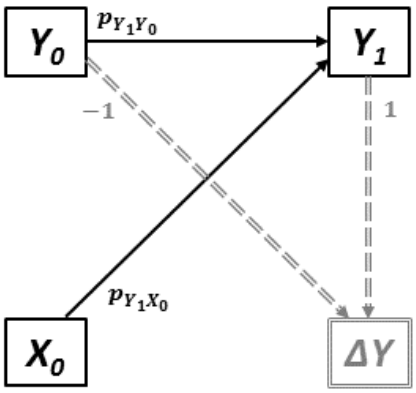
For each scenario, we are then able to use simple path tracing rules (35, 36) to identify the effect estimand for each method. We illustrate that the degree of discordance between a change-score analysis (i.e. Equation 4.1) and a follow-up adjusted for baseline analysis (i.e. Equation 4.2) depends on both the strength and nature of the relationship between the baseline exposure X_0 and the baseline outcome Y_0. Specifically, we show that although both methods purportedly estimate the effect of the baseline exposure X_0 on 'change' in Y, where there exists a causal relationship between X_0 and the baseline outcome Y_0 (i.e. in non-randomised settings), the two methods of analysis target different estimands.

Note that in all scenarios, we do not depict any direct causal relationship between X_0 and ΔY. This is because ΔY is a mathematically *determined* variable, and thus X_0 cannot have any effect on ΔY independent of its effects on the separate components Y_0 and/or Y_1. There is historical precedence this depiction (67) and it is theoretically justified by the results of Shahar, E. and D.J. Shahar (151). Throughout, deterministic relationships are indicated by double-lined arrows, and fully determined nodes are indicated by double-outlined rectangles.

4.4.1 Scenario 1: X_0 and Y_0 are causally unrelated

In Scenario 1, the baseline outcome Y_0 is a 'competing exposure' for the effect of X_0 on Y_1 (Figure 4.2). This is equivalent to a large, well-conducted randomised controlled trial, in which randomisation ensures that there exists no association between the exposure X_0 and the baseline outcome Y_0. In this scenario, both methods of analysis target the same causal association (i.e. the *total causal effect* of X_0 on Y_1).

Figure 4.2 Path diagram representing the hypothesised data-generating process for an exposure X measured once at baseline (i.e. X_0) and two measurements of a longitudinal outcome Y (i.e. Y_0 and Y_1), where X_0 and Y_0 are causally unrelated



The exposure X_0 is measured once at baseline, and the outcome Y is measured once at baseline (Y_0) and once at follow-up (Y_1). The change-score ΔY is constructed from $Y_1 - Y_0$.

The change-score analysis targets the total association between X_0 and ΔY. In this scenario, the total association between X_0 and ΔY is equal to the *causal* association $p_{Y_1X_0}$ (Table 4.1), and thus a change-score analysis will provide an unbiased estimate of the total causal effect of

X_0 on ΔY. Moreover, baseline randomisation ensures that the effect of X_0 on ΔY is in fact equal to the total causal effect of X_0 on Y_1 only.

The follow-up adjusted for baseline analysis targets the total association between X_0 and Y_1. As in the change-score analysis, this is equal to the *causal* association between X_0 and Y_1 (Table 4.1).

Thus, both the change-score analysis and the follow-up adjusted for baseline analysis estimate the *total causal effect* of X_0 on Y_1.

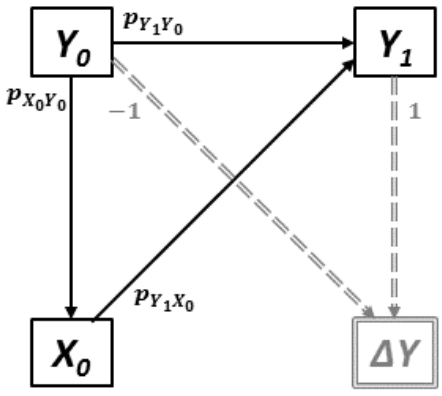
Table 4.1 Total association between X_0 and each of ΔY and Y_1, subdivided into causal and confounding associations, for the path diagram depicted in Figure 4.2

Outcome	Path		Association size	Total association
ΔY	Causal:	$X_0 \rightarrow Y_1 \rightarrow \Delta Y$	$p_{Y_1X_0} \cdot 1$	$p_{Y_1X_0}$
	Confounding:	n/a	n/a	
Y_1	Causal:	$X_0 \rightarrow Y_1$	$p_{Y_1X_0}$	$p_{Y_1X_0}$
	Confounding:	n/a	n/a	

4.4.2 Scenario 2: X_0 is caused by Y_0

In Scenario 2, the baseline outcome is a confounder of the effect of X_0 on Y_1 (Figure 4.3). Here, the change-score analysis and follow-up adjusted for baseline analysis target different estimands, and thus their results will diverge; the magnitude of this divergence is dependent upon the strength of the relationship between X_0 on Y_0.

Figure 4.3 Path diagram representing the hypothesised data-generating process for an exposure X measured once at baseline (i.e. X_0) and two measurements of a longitudinal outcome Y (i.e. Y_0 and Y_1), where X_0 is caused by Y_0



The exposure X_0 is measured once at baseline, and the outcome Y is measured once at baseline (Y_0) and once at follow-up (Y_1). The change-score ΔY is constructed from $Y_1 - Y_0$.

The change-score analysis targets the total association between X_0 and ΔY, which is $p_{Y_1X_0} + p_{X_0Y_0} \cdot p_{Y_1Y_0} - p_{X_0Y_0}$ (Table 4.2). However, in this scenario, the total association comprises both

causal and *confounding* associations, and thus the change-score analysis targets an estimand which is difficult – if not impossible – to interpret causally.

In contrast, the follow-up adjusted for baseline analysis targets *only* the causal association between X_0 and Y_1, which is $p_{Y_1X_0}$ (Table 4.2), since adjusting for Y_0 closes the confounding path $X_0 \leftarrow Y_0 \rightarrow Y_1$.

Thus, where the baseline outcome Y_0 is a confounder for the effect of X_0 on Y_1, the effects targeted by the change-score analysis and the follow-up adjusted for baseline analysis are expected to differ by $p_{X_0Y_0} \cdot (p_{Y_1Y_0} - 1)$. As the causal effect of Y_0 on X_0 strengthens (i.e. as $p_{X_0Y_0}$ increases), the magnitude of this difference will increase. Only under two specific circumstances are the two methods of analyses expected to agree:

1. Where X_0 and Y_0 are uncorrelated, such that $p_{X_0Y_0} = 0$; or
2. Where Y_0 and Y_1 are perfectly correlated, such that $p_{Y_1Y_0} - 1 = 0$.

We note that the first circumstance corresponds to the randomised experimental setting (i.e. Scenario 1), whereas the second corresponds to a setting in which Y_0 and Y_1 are related deterministically.

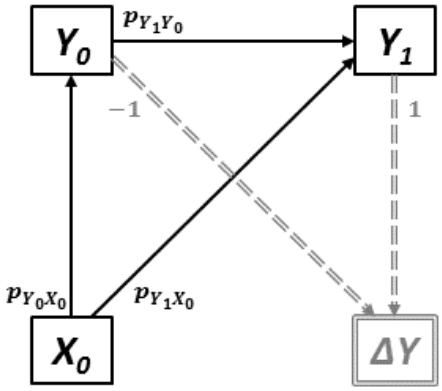
Table 4.2 Total association between X_0 and each of ΔY and Y_1, subdivided into causal and confounding associations, for the path diagram depicted in Figure 4.3

Outcome	Path		Association size	Total association
ΔY	Causal:	$X_0 \rightarrow Y_1 \rightarrow \Delta Y$	$p_{Y_1X_0} \cdot 1$	$p_{Y_1X_0} + p_{X_0Y_0} \cdot p_{Y_1Y_0} - p_{X_0Y_0}$
	Confounding:	$X_0 \leftarrow Y_0 \rightarrow Y_1 \rightarrow \Delta Y$	$p_{X_0Y_0} \cdot p_{Y_1Y_0} \cdot 1$	
		$X_0 \leftarrow Y_0 \rightarrow \Delta Y$	$p_{X_0Y_0} \cdot -1$	
Y_1	Causal:	$X_0 \rightarrow Y_1$	$p_{Y_1X_0}$	$p_{Y_1X_0} + p_{X_0Y_0} \cdot p_{Y_1Y_0}$
	Confounding:	$X_0 \leftarrow Y_0 \rightarrow Y_1$	$p_{X_0Y_0} \cdot p_{Y_1Y_0}$	

4.4.3 Scenario 3: X_0 causes Y_0

In Scenario 3, the baseline outcome is a mediator of the effect of X_0 on Y_1 (Figure 4.4). As in the previous scenario, the change-score analysis and follow-up adjusted for baseline analysis target different effects, and consequently their results will diverge according to the strength of the relationship between X_0 on Y_0.

Figure 4.4 Path diagram representing the hypothesised data-generating process for an exposure X measured once at baseline (i.e. X_0) and two measurements of a longitudinal outcome Y (i.e. Y_0 and Y_1), where X_0 causes Y_0



The exposure X_0 is measured once at baseline, and the outcome Y is measured once at baseline (Y_0) and once at follow-up (Y_1). The change-score ΔY is constructed from $Y_1 - Y_0$.

The change-score analysis targets the total association between X_0 and ΔY. This association comprises the total causal effect of X_0 on Y_1 (i.e. $p_{Y_1X_0} + p_{Y_0X_0} \cdot p_{Y_1Y_0}$) in addition the biasing component $-p_{Y_0X_0}$ that is introduced by the construction of the change-score (Table 4.3).

The follow-up adjusted for baseline analysis targets only the *direct* causal association between X_0 and Y_1, which is $p_{Y_1X_0}$ (Table 4.3), since adjusting for the baseline outcome closes the indirect causal path $X_0 \rightarrow Y_0 \rightarrow Y_1$.

Thus, where the baseline outcome Y_0 is a mediator for the effect of X_0 on Y_1, the effects targeted by the change-score analysis and the follow-up adjusted for baseline analysis are expected to differ by $p_{Y_0X_0} \cdot (p_{Y_1Y_0} - 1)$. As the causal effect of X_0 on Y_0 strengthens (i.e. as $p_{Y_0X_0}$ increases), the magnitude of this difference will increase. Only under two specific circumstances are the two methods of analyses expected to agree:

1. Where X_0 and Y_0 are uncorrelated, such that $p_{Y_0X_0} = 0$; or
2. Where Y_0 and Y_1 are perfectly correlated, such that $p_{Y_1Y_0} - 1 = 0$.

The first condition corresponds to the randomised experimental setting (i.e. Scenario 1), whereas the second corresponds to a setting in which Y_0 and Y_1 are related deterministically.

Table 4.3 Total association between X_0 and each of ΔY and Y_1, subdivided into causal and confounding associations, for the path diagram depicted in Figure 4.4

Outcome	Path		Association size	Total association
ΔY	Causal:	$X_0 \rightarrow Y_1 \rightarrow \Delta Y$	$p_{Y_1X_0} \cdot 1$	$p_{Y_1X_0} + p_{Y_0X_0} \cdot p_{Y_1Y_0} - p_{Y_0X_0}$
		$X_0 \rightarrow Y_0 \rightarrow Y_1 \rightarrow \Delta Y$	$p_{Y_0X_0} \cdot p_{Y_1Y_0} \cdot 1$	
		$X_0 \rightarrow Y_0 \rightarrow \Delta Y$	$p_{Y_0X_0} \cdot -1$	
	Confounding:	n/a		
Y_1	Causal:	$X_0 \rightarrow Y_1$	$p_{Y_1X_0}$	$p_{Y_1X_0} + p_{Y_0X_0} \cdot p_{Y_1Y_0}$
		$X_0 \rightarrow Y_0 \rightarrow Y_1$	$p_{Y_0X_0} \cdot p_{Y_1Y_0}$	
	Confounding:	n/a	n/a	

4.5 Follow-up adjusted for baseline analyses are not always the best solution for the analysis of change

In the previous sections, we have argued that change scores are of limited utility for causal analyses (§4.3) and demonstrated why change-score analyses will differ from follow-up adjusted for baseline analyses in non-randomised scenarios (§4.4).

It may thus be tempting to conclude that follow-up adjusted for baseline analyses represent the best solution for analyses of change in observational (i.e. non-randomised) scenarios, as has been concluded previously by several authors (150, 152). However, such a general conclusion is unwarranted for two reasons, both of which stem from a failure to consider situations in which the exposure X_0 causes the baseline outcome Y_0 (i.e. Scenario 3). In such situations, a follow-up adjusted for baseline analysis targets the *direct effect* of X_0 on Y_1 only. However, estimation of direct effects is notoriously difficult using standard regression methods due to their potential susceptibility to collider bias (159). Moreover, the direct effect is itself arguably less useful than the *total effect* in such situations, which can be obtained by a 'follow-up *unadjusted* for baseline' analysis.

We detail these two issues in greater detail in the following subsections.

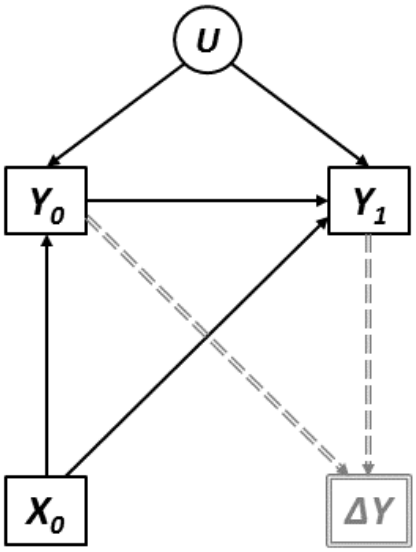
4.5.1 The issue of collider bias in the analysis of change

Where the exposure X_0 causes the baseline outcome Y_0 (i.e. Scenario 3), conditioning on Y_0 implies conditioning on a *mediator*, which can introduce collider bias if there exist unknown and/or unmeasured common causes of Y_0 and Y_1. Although we purposely omitted confounders in our simplified example scenarios (§4.4), in reality there are likely to be numerous common causes of Y_0 and Y_1 since they represent the same variable measured at different times.

To illustrate briefly, we consider the scenario depicted in Figure 4.5, in which Y_0 is a mediator for the effect of X_0 on Y_1 and there exists a latent variable U which confounds the relationship between Y_0 and Y_1. Here, adjusting for Y_0 (as in a follow-up adjusted for baseline analysis)

creates a spurious association between X_0 and Y_1 via the open path $X_0 \rightarrow Y_0 \leftarrow U \rightarrow Y_1$. This path could be closed by additionally conditioning on U, but this is impossible by definition because U is unobserved.

Figure 4.5 DAG representing the hypothesised data-generating process for an exposure X measured once at baseline (i.e. X_0), two measurements of a longitudinal outcome Y (i.e. Y_0 and Y_1), and one unobserved/latent variable U



In the setting of mediation analyses, U is frequently referred to as a 'mediator-outcome confounder' (159) since it confounds the relationship between the mediator (i.e. Y_0) and the outcome (i.e. Y_1).

4.5.2 Follow-up *unadjusted* for baseline analysis

Not only are direct effects more difficult to estimate in situations where the exposure X_0 causes the baseline outcome Y_0, they are arguably less useful. The experimental context is unique for ensuring that Y_0 is measured at or before the time of X_0 (as in Scenarios 1 and 2, respectively), thereby ensuring that all changes in Y that are caused by X_0 are fully realised by the effect of X_0 on the follow-up outcome Y_1. In other words, the experimental setting ensures that the effect of X_0 on exogenous change in Y is equal to the total causal effect of X_0 on Y_1. Indeed, this is underlined by Senn, S. (152), who argues that 'one should focus clearly on "outcomes" as being the only values influenced by treatment.'

However, where X_0 causes Y_0 (i.e. Scenario 3), the effects of an intervention targeting X_0 will be realised via its effects on *both* Y_0 and Y_1. In this case, a follow-up *unadjusted* for baseline analysis may be more appropriate, as in:

$$\hat{Y}_1 = \hat{\gamma}_0 + \hat{\gamma}_1 X_0 \qquad \textbf{Equation 4.3}$$

In this formulation, $\hat{\gamma}_1$ represents the effect of interest – the *total causal effect* of X_0 on Y_1.

4.6 The importance of defining the most useful estimand

A 'one-size-fits-all' approach to the analysis of change is unwarranted, since such an approach fails to consider the potential causal structures which exist. While change-score analyses

should be avoided in non-randomised contexts, the decision of whether to adjust for the baseline outcome (i.e. Y_0) when analysing effect of the exposure (i.e. X_0) the follow-up outcome (i.e. Y_1) should be informed by the context under consideration.

When the baseline outcome Y_0 causes X_0 (i.e. Scenario 2), a follow-up *adjusted* for baseline analysis should be conducted, since Y_0 is a *confounder* for the total effect of X_0 on Y_1. However, when the baseline outcome Y_0 is instead caused by X_0 (i.e. Scenario 3), adjusting for Y_0 targets the *direct* effect only. This effect may be useful in some contexts, including where the exposure X_0 is immutable and/or cannot hypothetically be targeted for intervention; however, estimation of this effect is likely to involve additional difficulties when there are unmeasured and/or unobserved common causes of Y_0 and Y_1 (as outlined in Section 4.5.1). Thus, it might alternately be determined that the *total* effect of X_0 on Y_1 is a more useful estimand, in which case adjustment for the baseline Y_0 is unwarranted.

Across all contexts, using DAGs to consider the causal structures involved can help to clarify these issues and to identify appropriate adjustment to target the most useful estimand. To this end, we present a simple simulated example in the following subsection.

4.6.1 Simulated example

To illustrate the degree of inferential bias that might be introduced by a change-score analysis, and to emphasise the importance of using DAGs to help determine the most useful analytical strategy, we consider a simple simulated example involving the effect of baseline waist circumference (WC_0) on the longitudinal exposure serum insulin concentration (IC), measured at baseline (i.e. IC_0) and follow-up (i.e. IC_1).

4.6.1.1 Methods

Data were simulated according to eight causal scenarios, each of which are depicted in Figure 4.6:

Scenario 1: Baseline waist circumference (WC_0) and baseline serum insulin concentration (IC_0) are causally unrelated, i.e. IC_0 is a *competing exposure* for the effect of WC_0 on IC_1.

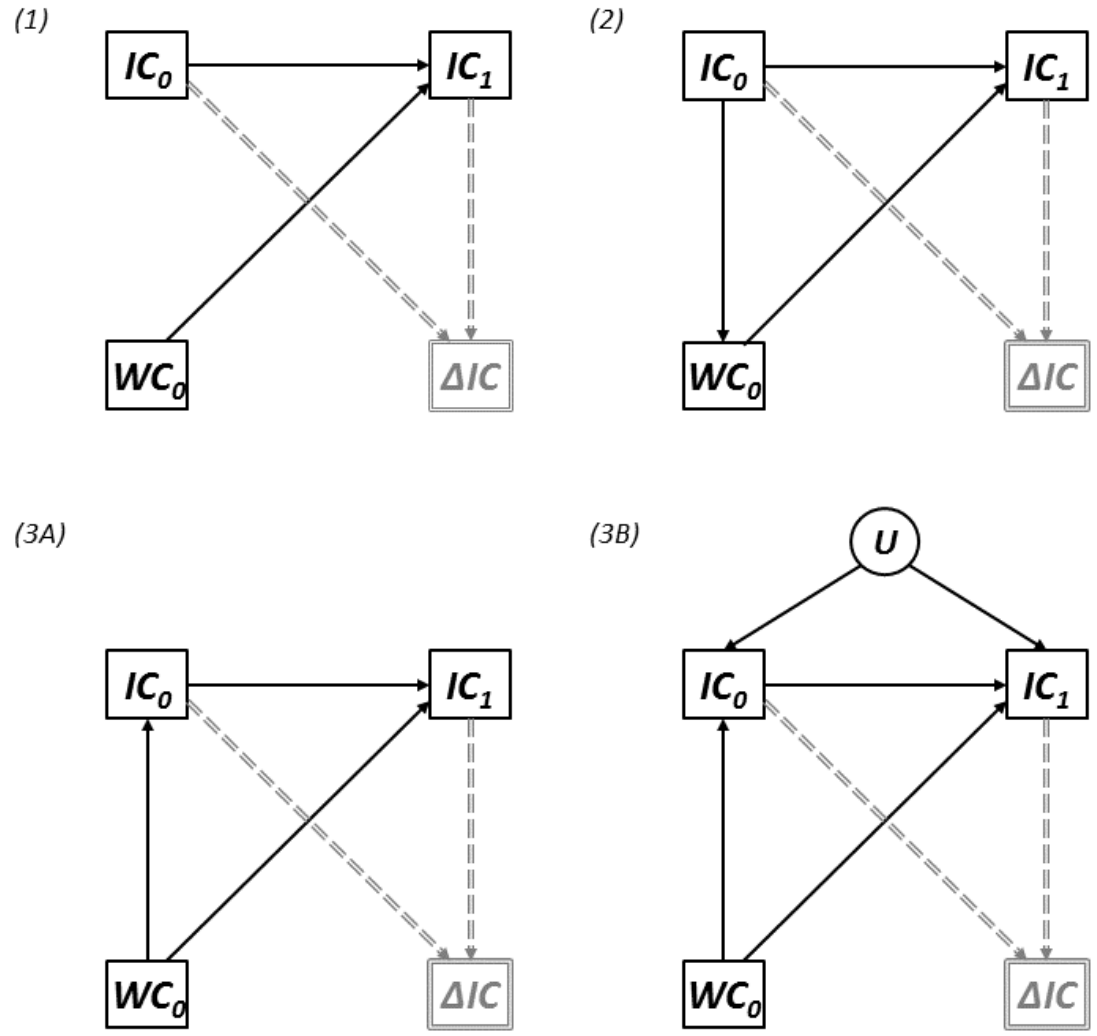
Scenario 2: Baseline waist circumference (WC_0) is caused by baseline serum insulin concentration (IC_0), i.e. IC_0 is a *confounder* for the effect of WC_0 on IC_1.

Scenario 3: Baseline waist circumference (WC_0) causes baseline serum insulin concentration (IC_0), i.e. IC_0 is a *mediator* for the effect of WC_0 on IC_1.

A. No unmeasured confounding.

B. Unmeasured variable (U) affecting IC_0 and IC_1 (i.e. mediator-outcome confounding (159)).

Figure 4.6 DAG representing four distinct hypothesised data-generating processes for the exposure waist circumference (WC) measured once at baseline (i.e. WC_0) and two measurements of the outcome serum insulin concentration (IC, i.e. IC_0 and IC_1)



In Scenario (1), IC_0 is a competing exposure for the effect of WC_0 on IC_1. In Scenario (2), IC_0 is a confounder for the effect of WC_0 on IC_1. In Scenario (3), IC_0 is a mediator for the effect of WC_0 on IC_1; U represents an unobserved or unmeasured variable that confounds the relationship between IC_0 and IC_1 (i.e. a mediator-outcome confounder).

Parameter values and path coefficients were informed by data on US adults aged 18-49 collected by the US National Health and Nutrition Examination Survey (NHANES), from the years 2019-2014 (160). The total causal effect of WC_0 on IC_1 was fixed at 0.200 Log(mmol/L)/dm; when mediated through IC_0, this was partitioned into an indirect causal effect of 0.150 Log(mmol/L)/dm and a direct causal effect of 0.050 Log(mmol/L)/dm. Additional details relating to this simulation, including parameters and code, can be found in Appendix A (§A.2).

For each Scenario, we simulated 10,000 non-overlapping random samples of 1,000 observations from a multivariate normal distribution based upon the relevant DAG in Figure 4.6, using the 'dagitty' package (v. 0.2-2)(46, 47) in R (v. 3.3.2)(161). For each sample, we conduct three analyses:

1. A change-score analysis: $\widehat{\Delta IC} = \hat{\alpha}_0 + \hat{\alpha}_1 WC_0$.
2. A follow-up adjusted for baseline analysis: $\widehat{IC_1} = \hat{\beta}_0 + \hat{\beta}_1 WC_0 + \hat{\beta}_2 IC_0$.
3. A follow-up *un*adjusted for baseline analysis: $\widehat{IC_1} = \hat{\gamma}_0 + \hat{\gamma}_1 WC_0$.

The median value across all 1000 samples for the coefficient of WC_0 in each analysis (i.e. $\hat{\alpha}_1$, $\hat{\beta}_1$, or $\hat{\gamma}_1$) is reported along with its 95% simulation limits (i.e. 2.5 and 97.5 centile estimates). We then consider the implications of interpreting these coefficients as the desired causal effect on IC_1. Note that coefficient units (i.e. Log(mmol/L)/dm) are omitted to aid readability.

4.6.1.2 Results

The simulation results are summarised in Table 4.4.

4.6.1.2.1 Scenario 1: WC_0 and IC_0 are causally unrelated

Scenario 1 is analogous to a large, well-conducted RCT. The total association between WC_0 and ΔIC consists entirely of the causal effect of WC_0 on IC_1 since there is no confounding or mediation by IC_0. All methods therefore provide an unbiased estimate of the total causal effect of WC_0 on 'change' in IC (i.e. $\hat{\alpha}_1, \hat{\beta}_1, \hat{\gamma}_1$ = 0.200).

4.6.1.2.2 Scenario 2: WC_0 is caused by IC_0

In Scenario 2, the total association between WC_0 and ΔIC consists of both the causal effect of WC_0 on IC_1 and confounding by IC_0. Therefore, only the follow-up adjusted for baseline analysis provides an unbiased estimate of the total causal effect of WC_0 on IC_1 ($\hat{\beta}_1$ = 0.200), whereas the change-score analysis and the follow-up *un*adjusted for baseline analysis are biased ($\hat{\alpha}_1$ = 0.119 and $\hat{\gamma}_1$ = 0.350, respectively).

4.6.1.2.3 Scenario 3: WC_0 causes IC_0

In Scenario 3, the total association between WC_0 and ΔIC consists of both the direct and indirect effects of WC_0 on IC_1, in addition to the biasing path from WC_0 to ΔIC through the baseline outcome IC_0. Thus, in Scenarios 3A and 3B the change-score analysis provides a biased estimate ($\hat{\alpha}_1$ = -0.031) of both the total and direct causal effects of WC_0 on IC_1; moreover, this estimate is of *opposite sign* to the true effects. The follow-up adjusted for baseline analysis provides an unbiased estimate of the *direct* causal effect of WC_0 on IC_1 ($\hat{\beta}_1$ = 0.050, Scenario 3A), though it becomes biased in the presence of unmeasured mediator-outcome confounding by U ($\hat{\beta}_1$ = 0.025, Scenario 3B). The follow-up *un*adjusted for baseline analysis, however, provides an unbiased estimate of the *total* causal effect of WC_0 on IC_1 ($\hat{\gamma}_1$ = 0.200, Scenario 3A); this estimate remains robust in the presence of mediator-outcome confounding (i.e. Scenario 3B).

4.6.1.3 Implications

In this simulated example, we explored the seemingly simple context of change in an outcome (i.e. insulin concentration) with respect to a baseline exposure (i.e. waist circumference) for

four different causal scenarios. Using change-score analyses, misleading coefficients – sometimes of opposite sign to the true causal effects – were observed in all scenarios where the baseline outcome and exposure were correlated at baseline (i.e. Scenarios 2 and 3). In such scenarios, determining the most appropriate adjustment for the baseline outcome when analysing the effect of the baseline exposure on the follow-up outcome was aided greatly by the use of DAGs.

Although our simulations were deliberately simplified and made several distributional assumptions that may not be entirely realistic, they clearly demonstrate the potential problems associated with analysing 'change' using change scores, and the benefits of using DAGs to understand and identify the most appropriate analytical strategies. We additionally considered the four causal scenarios in Figure 4.6 with an unmeasured baseline confounder U_2 affecting each of WC_0, IC_0, and IC_1. Under such circumstances, all three methods unsurprisingly provided biased estimates of the total causal effect of WC_0 on IC_1. However, across all scenarios, the results were broadly consistent with those of the original simulation; a follow-up adjusted for baseline analysis appeared to be the least biased for Scenarios 1, 2, and 3A, whereas a follow-up unadjusted for baseline analysis was preferred for Scenario 3B. All details related to this additional simulation and analysis are located in Appendix A (§A.2).

Table 4.4 Median regression coefficient of WC_0 (and 95% simulation limits) for each method of analysis, for each causal scenario depicted in Figure 4.6

	IC_0 **is:**	**Competing exposure**	**Confounder**	**Mediator**	
Method of analysis: ↓	**Scenario:**	**1**	**2**	**3A**	**3B**
Change-score		**0.200**	0.119	-0.031	-0.031
$(\widehat{\Delta IC} = \hat{\alpha}_0 + \hat{\boldsymbol{\alpha}}_{\mathbf{1}} WC_0)$		(0.180, 0.220)	(0.106, 0.132)	(-0.053, -0.009)	(-0.050, -0.012)
Follow-up adjusted for baseline		**0.200**	**0.200**	**0.050**	0.025
$(\widehat{IC_1} = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}_{\mathbf{1}} WC_0 + \hat{\beta}_1 IC_0)$		(0.183, 0.218)	(0.189, 0.211)	(0.027, 0.073)	(0.005, 0.046)
Follow-up unadjusted for baseline		**0.200**	0.351	**0.200**	**0.200**
$(\widehat{IC_1} = \hat{\gamma}_0 + \hat{\boldsymbol{\gamma}}_{\mathbf{1}} WC_0)$		(0.174, 0.226)	(0.333, 0.369)	(0.174, 0.227)	(0.174, 0.226)

Numbers in green indicate unbiased estimates of either the total or direct effect of WC_0 on IC_1. The true total effect was simulated to be 0.200; where this was when mediated through IC_0 (i.e. Scenarios 3A and 3B), the true direct effect was simulated to be 0.050. In Scenario 3B, an unobserved/unmeasured variable U was simulated to confound the relationship between IC_0 and IC_1.

4.7 Examining 'Lord's Paradox'

The previous analysis offers a compelling lens through which to view 'Lord's Paradox' (147), a peculiarity that has evaded statisticians since its original formulation in 1967 (154-156, 162). The paradox is summarised below:

A university is interested in investigating whether the diet provided in the dining halls has an effect on students' weight over the course of the year, and whether there are any sex differences in these effects. It hires two statisticians to answer this question. The first statistician examines the mean weight of the girls at the beginning of the year and at the end of the year and finds that these are identical; moreover, the frequency distribution of weight for girls at the end of the year is the same as it was at the beginning. He finds the same to be true for boys, and thus concludes that there is no evidence for any differential effect on the two sexes. The second statistician conducts an analysis of covariance (i.e. a follow-up adjusted for baseline analysis). He finds that the slope of the regression line of final weight on initial weight is essentially the same for the two sexes, but that the difference between the intercepts is statistically highly significant. The second statistician therefore concludes that boys showed significantly more gain in weight than the girls when proper allowance is made for differences in initial weight between the two sexes.[16]

Which statistician is correct? The conclusions of the two statisticians appear to contradict one another, leading Lord to conclude that 'there simply is no logical or statistical procedure that can be counted on to make proper allowances for uncontrolled pre-existing differences between groups' (147). However, the causal lens adopted in the previous sections can help to resolve this question.

4.7.1 Considering the paradox within a causal framework

One of the primary challenges in interpreting Lord's paradox stems from the fact that baseline weight is a *mediator* for the effect of sex on final weight. Thus, although baseline weight represents a 'pre-existing difference' between boys and girls according to Lord, it is actually a *consequence* of the exposure rather than a *cause* of it. Therefore, it is fundamentally different from the experimental setting that has historically been considered in the analysis of change – a setting in which pre-existing differences occur before the exposure.[17]

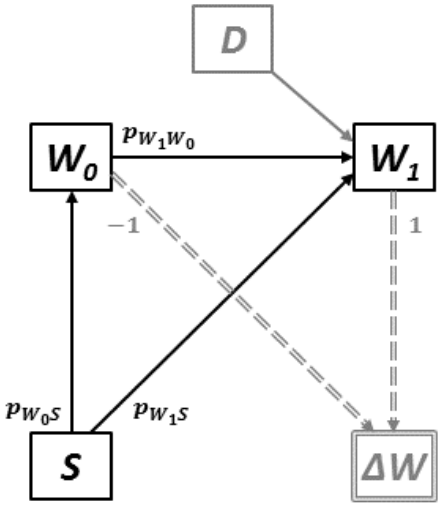
---

[16] We note that the research question is itself ill-defined, since the diet is a fixed condition that is applied to all students, male and female. Therefore, the diet can have no identifiable causal effect on the students' weights. Consequently, the question appears to actually be about the differential effect of *sex* on weight change.

[17] Additional confusion has been created by the fact that in several subsequent reinterpretations of Lord's Paradox, the baseline outcome is in fact a *confounder* rather than a mediator (27, 163).

In Figure 4.7 we draw the scenario described by Lord as a path diagram, where S represents sex, W_0 represents initial weight, W_1 represents final weight, and D represents diet. This path diagram is equivalent to the one considered in Scenario 3 (§4.4.3).

Figure 4.7 Path diagram representing Lord's Paradox (147)



The exposure sex (S) is measured once at baseline, and the outcome weight (W) is measured once at baseline (W_0) and once at follow-up (W_1). The change-score ΔW is constructed from $W_1 - W_0$. The diet (D) is depicted in grey because, although it affects W_1, it is not truly a variable; all students are subjected to it, and it thus does not have any identifiable causal effect on W_1.

The first statistician, in comparing the average change scores between boys and girls, has essentially conducted a change-score analysis and found the effect of sex on 'change' in weight to be zero. In other words, the first statistician estimated the *total* effect of sex on *weight change-score* (i.e. $p_{W_1S} + p_{W_0S} \cdot p_{W_1W_0} - p_{W_0S}$, in Figure 4.7). By contrast, the second statistician conducted a straightforward follow-up adjusted for baseline analysis, and thus estimated the *direct* effect of sex on *final weight* (i.e. p_{W_1S}, in Figure 4.7).

It is therefore not surprising that the two statisticians came to different conclusions, and Lord himself did not see how this problem might be resolved, as the effects estimated are expected to differ by $p_{W_0S} \cdot (p_{W_1W_0} - 1)$. Only under one of the following conditions would the two agree:

1. Where S and W_0 were uncorrelated (i.e. the randomised experimental setting), such that $p_{W_0S} = 0$; or
2. Where W_0 and W_1 were perfectly correlated (i.e. the deterministic case in which *no exogenous change exists*), such that $p_{W_1W_0} - 1 = 0$.

4.7.2 Identifying the most useful estimand

Identifying the most useful and/or important estimand in this scenario can help us to identify which statistician was correct.

In conducting a change-score analysis, Statistician 1 did not in fact estimate the effect of sex on 'change' in weight; this is because 'change' is fully encapsulated in the follow-up weight whereas a change score conflates information from *both* baseline and follow-up (as argued in Section 4.3).[18] Therefore, Statistician 2 can claim to be in possession of a more meaningful answer to the original query, since the direct effect represents a valid estimand in this scenario.

However, we might consider an equally valid solution (presented by a fictional 'Statistician 3') to be the *total* effect of sex on final weight, obtained via a follow-up *unadjusted* for baseline analysis. This effect captures *all* changes in weight which result either directly or indirectly from sex. Moreover, this estimate is not susceptible to the potential biases introduced by conditioning on a mediator (i.e. baseline weight, which is likely to share many causes with follow-up weight).

4.8 Comparison with Glymour, M.M. et al. (158) and Kim, Y. and P.M. Steiner (148)

Our conclusions appear to fundamentally contradict those of Glymour, M.M. et al. (158) and Kim, Y. and P.M. Steiner (148) who – as was mentioned in Section 4.2.3 – are notable for having considered the analysis of change using DAGs. Here, we attempt to briefly explain the reasons behind our differing conclusions.

Glymour, M.M. et al. (158) examine the analysis of change in the context of the effect of education on change in cognitive function in an elderly cohort. This context is most closely approximated by our Scenario 3, in which the baseline outcome (i.e. baseline cognitive function) is a mediator for the effect of the exposure (i.e. education) on the follow-up outcome (i.e. follow-up cognitive function). However, a critical difference is that the authors do not recognise the deterministic nature of the change score, and that it is fundamentally distinct from true exogenous (or *modifiable*) change. This represents a philosophical difference that cannot be resolved by mathematics, but for which both we (in Section 4.3.2) and Shahar, E. and D.J. Shahar (151) have argued.

More problematically, the conclusions of Glymour, M.M. et al. (158) are supported only by an analysis of data from the Assets and Health Dynamics Among the Oldest Old (AHEAD) study, in which the true causal structure is not known. The authors *assume* there is no causal effect of education on change in cognitive function, but do not simulate data for which this is sure to be the case. Their change-score analysis produces an estimated effect of -0.02 (-0.05, 0.01) and their follow-up adjusted for baseline analysis produces an estimated effect of 0.20 (0.17, 0.23),[19] which they interpret as suggesting that the change-score analysis is unbiased.

---

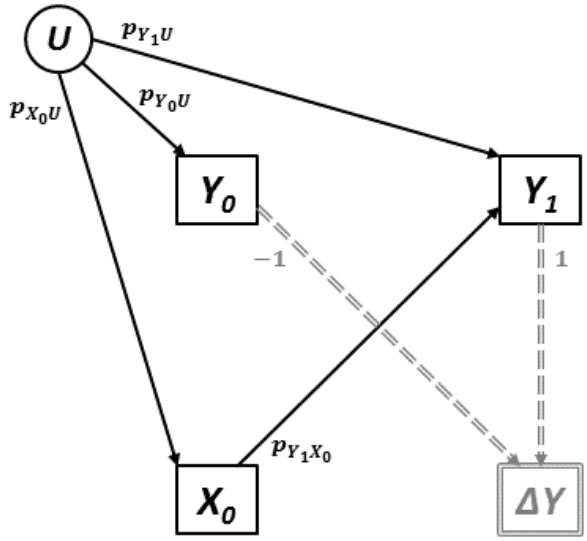
[18] We note that this is in contrast to Pearl, J. (67), who claims that the first statistician did in fact estimate the total effect of sex on gain (i.e. change).

[19] For change in cognitive function between the years 1993 and 1995.

However, we note that their results are broadly consistent with those produced for Scenario 3 in our simulated example (Table 4.4), in which we simulated a *true causal effect*.

Kim, Y. and P.M. Steiner (148) do acknowledge the deterministic nature of the change score, but fail to recognise that this is not equal to true exogenous change. Of note, the authors assume that any relationship between the baseline and follow-up outcome is due to a (possibly latent) preceding common cause which also affects the exposure; this causal structure is depicted in Figure 4.8. Using path tracing, the authors show that where $p_{Y_1U} = p_{Y_0U}$ (i.e. where U affects both Y_0 and Y_1 to the same extent) the confounding bias caused by U cancels out. While there may be specific instances in which this holds, we do not believe that it can be assumed to hold in general. Moreover, where there exists a causal relationship between the baseline outcome and the exposure (i.e. $Y_0 \rightarrow X_0$ in Figure 4.8), the authors acknowledge that this assumption is likely untenable.

Figure 4.8 Path diagram representing the analysis of change as depicted by Kim, Y. and P.M. Steiner (148)



The exposure X_0 is measured once at baseline, and the outcome Y is measured once at baseline (Y_0) and once at follow-up (Y_1). The unobserved variable U affects each of X_0, Y_0, and Y_1. The change-score ΔY is constructed from $Y_1 - Y_0$.

4.9 Implications

Using DAGs, we have demonstrated that change scores do not in general represent exogenous change, and thus are of limited utility in causal analyses despite their seemingly intuitive formulation. Change-score analyses treat two separate events (i.e. Y_0 and Y_1) as one (i.e. ΔY), thus conflating the causal pathways involved and potentially leading to inferential bias. Only under baseline randomisation can change scores be used without bias; in non-randomised data, change-score analyses do not generally estimate causal effects. Previous studies which have conducted change-score analyses in observational data should be scrutinised, and their results viewed with caution.

The use of a DAGs also clarifies the situations in which adjustment for the baseline outcome (i.e. Y_0) may or may not be warranted, when analysing the effect of the exposure (i.e. X_0) on the follow-up outcome (i.e. Y_1). Statistical adjustment for Y_0 (as in follow-up adjusted for baseline analyses) has historically been considered necessary, in order to 'standardise' Y_1 by Y_0. However, this may not actually be desirable across all contexts. This highlights the importance of using DAGs to identify the most plausible causal structure on a case-by-case basis, and to determine appropriate adjustment for Y_0 according to the most useful estimand.

4.10 Summary

Studies of 'change' are common in the epidemiological literature, yet rarely has the concept of change been formally considered within a causal framework. This chapter demonstrates that DAGs are useful for clarifying the distinction between change scores and true exogenous change, which is the concept of change most useful from a causal perspective. DAGs are also useful for considering the potential causal structures that may arise in analyses of change, and consequently in understanding why (and the degree to which) change-score analyses may differ from follow-up adjusted for baseline analyses. Across all contexts, using DAGs to consider the causal structures involved can help to identify the most useful estimand to target in analyses of change, which may necessitate follow-up *unadjusted* for baseline analyses in certain situations.

Chapter 5
Regression with 'unexplained residuals'

5.1 Introduction

Time-varying exposures present analytical challenges above and beyond those of time-fixed exposures. Consequently, time-varying exposures are often reconceptualised as a series of time-fixed exposures, *each* of which has a total causal effect on the outcome of interest that can be estimated using a standard regression model. The necessity of multiple models led to the introduction of 'unexplained residuals' (UR) models by Keijzer-Veen, M.G. et al. (93) as a way of estimating the total causal effect of multiple measurements of a time-varying exposure within a single model.[20] However, this method presents other unrecognised analytical challenges, particularly in the presence of confounding by both baseline and time-dependent covariates (which were not formally considered when UR models were first introduced, nor subsequently).

Chapter 5 considers UR models within a formal causal framework. The basis of this chapter has three primary benefits. First, it clarifies why the method works (i.e. why it is equivalent to standard regression methods) for estimating the total causal effect of multiple measures of a time-varying exposure on an outcome in the absence of any confounding. Second, it allows us to consider how the method may be extended robustly to account for confounding by both baseline and time-dependent covariates, since UR models are of limited utility if they may only be used in situations in which no confounding exists. Third, it provides a general framework for considering how the method may be extended robustly to more complex longitudinal scenarios. With this information, we are then able to more comprehensively evaluate the benefits of UR models across a wide variety of longitudinal scenarios.

5.1.1 Chapter overview

A general chapter overview is provided below.

In Section 5.2 we introduce the example scenario originally considered by Keijzer-Veen, M.G. et al. (93), and depict this scenario using a DAG. We consider how standard regression models and UR models may be used to estimate total causal effects in this setting.

In Section 5.3 we use the method of path coefficients to illustrate the unique properties of UR models within a causal framework.

---

[20] UR models have alternately been referred to as 'unexplained residuals regression' (164), the 'method of unexplained residuals' (165), 'conditional linear regression' (164), 'conditional (regression) models' (24, 166), 'conditional (regression) analysis' (167-171), 'regression with conditional growth measures' (166), 'conditional growth models' (172-175), and 'conditional weight models' (176).

In Section 5.4 we separately consider confounding by a baseline (§5.4.1) and time-dependent covariate (§5.4.2). In each scenario, we use a DAG to consider correct confounder adjustment and to explain why a UR model with correct adjustment for confounding will continue to satisfy the original properties of UR models.

In Section 5.5 we extend the method of UR models to a scenario involving T measurements of a time-varying exposure, and additionally consider how to adjust confounding by a baseline (§5.5.1.1) and time-dependent covariate (§5.5.1.2) in this extended context.

In Section 5.6 we demonstrate that the standard error(s) of the estimated effect sizes are artificially reduced when using UR models – a previously unrecognised caveat regarding their use and implementation.

In Section 5.7 we discuss the implications of our findings.

5.1.2 Related publications

This chapter contains work based on the following publication:

Arnold, K.F., Ellison, G.T.H., Gadd, S.C., Textor, J., Tennant, P.W.G., Heppenstall, A. and Gilthorpe, M.S. Adjustment for time-invariant and time-varying confounders in 'unexplained residuals' models for longitudinal data within a causal framework and associated challenges. *Statistical Methods in Medical Research.* 2019, 28(5), pp.1347-1364. (5)

5.2 Estimating the total causal effect of multiple measurements of a time-varying exposure on a future outcome

The total causal effect of an exposure on a subsequent outcome comprises both its direct effect and any indirect effects on the outcome. Where an exposure is time-varying, the total effect of *each* measurement may be desired.

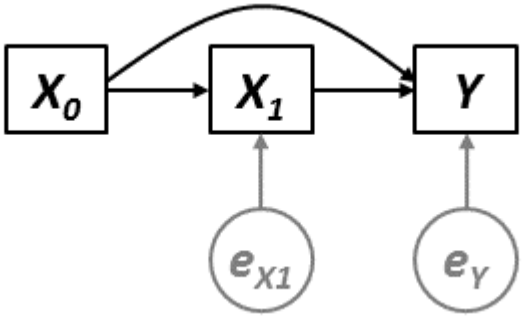
In the following sections, we introduce a simple example scenario involving two measurements of a time-varying exposure and subsequent outcome. We then describe the 'standard' regression method for estimating the total causal of each measurement and the associated 'unexplained residuals' (UR) regression method.

Throughout, all DAGs are drawn forwardly saturated (i.e. where each node may causally affect all future nodes), and all unexplained causes of endogenous nodes are represented by the variable *e* and depicted as independent (i.e. we assume no unobserved confounding).

5.2.1 Example scenario

We consider a time-varying exposure X measured at two time points (i.e. X_0 and X_1) and a subsequent outcome Y, where all variables are continuous. This scenario is depicted using a DAG in Figure 5.1, in which all unexplained causes of the endogenous nodes X_1 and Y are represented by the variables e_{X1} and e_Y, respectively.

Figure 5.1 DAG depicting the hypothesised data-generating process for two measurements of a time-varying exposure X (i.e. X_0 and X_1) and one outcome Y



The terms e_{X1} and e_Y represent all unexplained causes of X_1 and Y, respectively, and are included to explicitly reflect uncertainty in all endogenous nodes (whether modelled or not).

5.2.2 Standard regression method

To estimate the total causal effect on Y of each measurement of the exposure (i.e. X_0 and X_1), each must be treated as a separate entity that is potentially subject to confounding by any previous measurement(s) of that variable. Therefore, two distinct regression models are necessary, respectively:

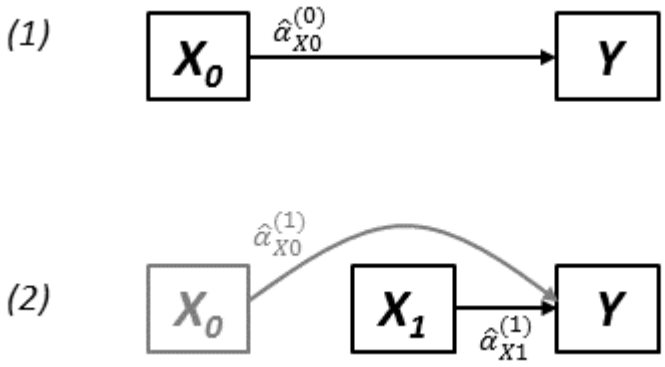
$$\hat{Y}_S^{(0)} = \hat{\alpha}_0^{(0)} + \hat{\boldsymbol{\alpha}}_{\boldsymbol{X0}}^{(\mathbf{0})} X_0 \qquad \textbf{Equation 5.1}$$

$$\hat{Y}_S^{(1)} = \hat{\alpha}_0^{(1)} + \hat{\alpha}_{X0}^{(1)} X_0 + \hat{\boldsymbol{\alpha}}_{\boldsymbol{X1}}^{(\mathbf{1})} X_1 \qquad \textbf{Equation 5.2}$$

In Equation 5.1, the total causal effect of X_0 on Y is represented by the coefficient $\hat{\alpha}_{X0}^{(0)}$. In Equation 5.2, the total causal effect of X_1 on Y is represented by the coefficient $\hat{\alpha}_{X1}^{(1)}$; no interpretation of $\hat{\alpha}_{X0}^{(1)}$ is possible (nor should it be attempted) for X_0 in Equation 5.2, as it acts purely as a confounder for the effect of X_1 on Y.

A visual depiction of these equations is given in Figure 5.2.

Figure 5.2 Path diagrams depicting the two standard regression models that would be constructed to estimate the total causal effect of each of X_0 and X_1 on Y (i.e. Equation 5.1 and Equation 5.2, respectively)

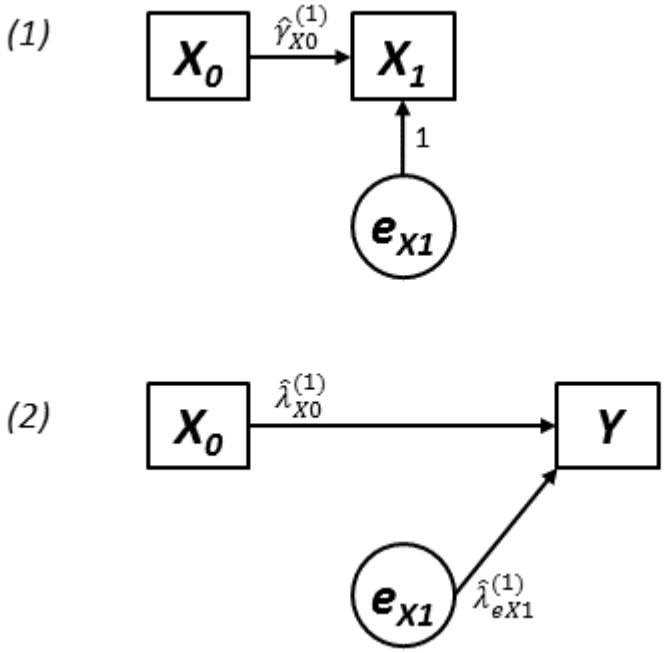


For each model, only the final coefficient may be interpreted as a total causal effect; all other coefficients are greyed to illustrate that no such interpretation should be made for them.

5.2.3 Unexplained residuals (UR) method

A UR model allows us to quantify the total effects of *both* the initial measurement of X (i.e. X_0) and subsequent change in X (i.e. X_1) on the outcome Y. To achieve this, a two-step process is implemented, summarised in Figure 5.3.

Figure 5.3 Path diagrams depicting the two steps of constructing a UR model



Step (1): the preparatory regression of X_1 on X_0 (Equation 5.3); and Step (2): the UR model (Equation 5.4).

First, the second measurement of the exposure X_1 is regressed on the initial measurement X_0, as in:

$$X_1 = \hat{\gamma}_0^{(1)} + \hat{\gamma}_{X0}^{(1)} X_0 + e_{X1} \qquad \textbf{Equation 5.3}$$

This produces a measure of each observation's 'expected' value of X_1 as predicted by its value of X_0. The difference between the expected value of X_1 and the observed value of X_1 equals the residual term e_{X1}. Put another way, e_{X1} represents the part of X_1 'unexplained' by X_0.

Second, Y is regressed on both the initial exposure X_0 and subsequent residual term e_{X1}:

$$\hat{Y}_{UR}^{(1)} = \hat{\lambda}_0^{(1)} + \hat{\boldsymbol{\lambda}}_{\boldsymbol{X0}}^{(\mathbf{1})} X_0 + \hat{\boldsymbol{\lambda}}_{\boldsymbol{eX1}}^{(\mathbf{1})} e_{X1} \qquad \textbf{Equation 5.4}$$

Keijzer-Veen, M.G. et al. (93) have previously demonstrated that the UR model in Equation 5.4 is algebraically equivalent to the final standard regression model (Equation 5.2), whilst allowing for the interpretation of *both* coefficients (i.e. $\hat{\lambda}_{X0}^{(1)}$ and $\hat{\lambda}_{eX1}^{(1)}$) as the total causal effects on Y of X_0 and X_1, respectively. Indeed, this is perceived to be a key advantage of UR models (177), since the final standard regression model may be prone to misinterpretation of the coefficient $\hat{\alpha}_{X0}^{(1)}$ (which represents the *direct effect* of X_0 on Y *only*).

5.2.3.1 Key properties of UR models

The key properties of UR models are described and formally expressed mathematically in Table 5.1.

Table 5.1 Description of key properties of UR models for a time-varying exposure X measured at two time points (i.e. X_0 and X_1) and one outcome Y

Property	Description	Mathematical formulation
(i)	The outcome values predicted by the final standard regression model (i.e. for exposure X_1) are equal to those predicted by the UR model.	$\hat{Y}_S^{(1)} = \hat{Y}_{UR}^{(1)}$
(ii)	The estimated coefficient for X_0 in the initial standard regression model (i.e. for exposure X_0) is equal to the estimated coefficient for X_0 in the UR model.	$\hat{\alpha}_{X0}^{(0)} = \hat{\lambda}_{X0}^{(1)}$
(iii)	The estimated coefficient for X_1 in its individual standard regression model (i.e. for exposure X_1) is equal to the estimated coefficient for the corresponding UR term e_{X1} in the UR model.	$\hat{\alpha}_{X1}^{(1)} = \hat{\lambda}_{eX1}^{(1)}$

Formal proofs of Properties (i) – (iii) are provided in Appendix B (§B.4, with $T = 2$), though intuitive explanations are given below.

Property (i) follows from the fact that the UR model (Equation 5.4) is simply a reparameterisation of the final standard regression model (Equation 5.2). Whereas the final standard regression model represents Y as a function of X_0 and X_1, the UR model represents Y

as a function of X_0 and e_{X1}. However, owing to the fact that e_{X1} is a function of X_0 and X_1 (Equation 5.3), the UR model itself is also a function of X_0 and X_1.

Property (ii) follows from the fact that e_{X1} is orthogonal to X_0 by construction. Thus, the estimated effect of X_0 on Y is the same regardless of whether or not e_{X1} is included in the model.

Property (iii) follows from the previous two. Adjustment for X_0 in the standard regression model (Equation 5.4) amounts to testing the relationship between Y and the part of X_1 *unexplained by* X_0 (i.e. the *unexplained residual*). Thus, the two coefficients are in fact equal because they mean the same thing (94), the only difference being that e_{X1} is orthogonal to X_0 by construction but X_1 is not.

5.3 Understanding UR models using DAGs

The unique properties of UR models can be easily visualised and understood using DAGs (Figure 5.1). If one were to naively model both measurements of the exposure (i.e. X_0 and X_1) simultaneously, only the coefficient for X_1 could be interpreted as a total causal effect on Y; the coefficient of X_0 would represent only the *direct* effect of X_0 on Y, because X_1 mediates the effect of X_0 on Y.

However, the UR modelling process relies upon and exploits the independence (i.e. orthogonality) of the UR term e_{X1}, which is independent of X_0 by construction. Thus, e_{X1} does not act as a mediator for the effect of X_0 on Y. In fact, the UR term e_{X1} can be understood as an instrumental variable (178) for X_1 that has been produced by the modelling process.[21]

More formally, we may apply the method of path coefficients (35, 36) to demonstrate that the 'true' total causal effect of each measurement of the exposure (i.e. X_0 and X_1) in the data-generating process is equal to the total causal effect of the associated terms in the UR model (i.e. X_0 and e_{X1}, respectively). Because both the standard regression models and their corresponding UR models are based on linear regression (where the causal relationships between variables are assumed to be linear functions), we may parameterise the DAG in Figure 5.1 by assigning a single coefficient to every arrow and assuming all variables have a variance of one. We use the notation p_{ba} to represent the coefficient of the arrow $a \rightarrow b$.

Table 5.2 gives the total effects of the model covariates in both the standard regression model and the corresponding UR model; each effect is decomposed into direct and indirect effects. As is evident, the total effects of X_0 and e_{X1} in the UR model are equivalent to the effects of X_0 and X_1, respectively, in the standard regression model. This is because there are no direct paths between e_{X1} and Y, and the only indirect path passes through X_1 (with $p_{X_1 e_{X1}}$ being equal to one, as in Figure 5.3).

---

[21] The process shares similarities with the two-stage least squares regression method (179), a form of instrumental variable analysis commonly encountered in economics research.

Table 5.2 Comparing standard regression models and UR models using the method of path coefficients

Exposure	Path		Effect size	Total effect
Standard regression models (Equation 5.1 ***and Equation*** 5.2***):***				
X_0	Direct:	$X_0 \to Y$	p_{YX_0}	$p_{YX_0} + p_{X_1X_0} \cdot p_{YX_1}$
	Indirect:	$X_0 \to X_1 \to Y$	$p_{X_1X_0} \cdot p_{YX_1}$	
X_1	Direct:	$X_1 \to Y$	p_{YX_1}	p_{YX_1}
	Indirect:	n/a		
'Unexplained residuals' (UR) model (Equation 5.4***):***				
X_0	Direct:	$X_0 \to Y$	p_{YX_0}	$p_{YX_0} + p_{X_1X_0} \cdot p_{YX_1}$
	Indirect:	$X_0 \to X_1 \to Y$	$p_{X_1X_0} \cdot p_{YX_1}$	
e_{X1}	Direct:	n/a		p_{YX_1}
	Indirect:	$e_{X1} \to X_1 \to Y$	$p_{X_1e_{X1}} \cdot p_{YX_1}$	

Total effects estimated by individual standard regression models (Equation 5.1 and Equation 5.2) compared to total effects estimated by a single composite UR model (Equation 5.4).

5.4 Confounding adjustment within UR models

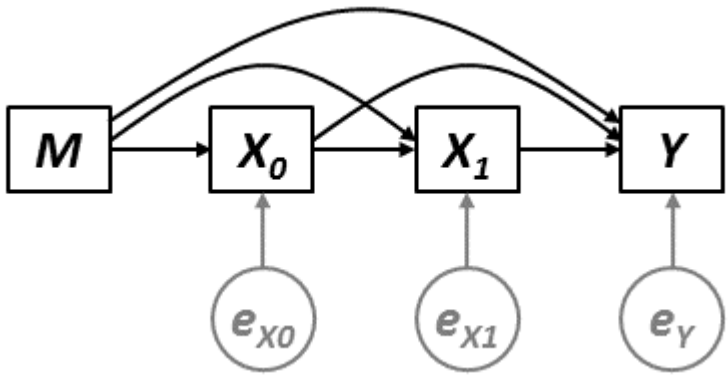
Keijzer-Veen, M.G. et al. (93) did not address confounding variables in their original paper, and there has been little to no discussion or analysis of this issue by subsequent authors using UR models. Consequently, *ad-hoc* methods for confounding adjustment within UR models have arisen in the absence of any formal guidance. For instance, Horta, B.L. et al. (173) made no adjustments for potential confounders when deriving their UR terms, but did make adjustments within their UR model. In contrast, Gandhi, M. et al. (175) adjusted for one potential confounder when creating their UR terms, but also made further adjustments in the UR model. Different procedures for confounder adjustment abound in the literature.

Because confounding is fundamentally a causal concept, considering how to correctly adjust for confounding within UR models is aided greatly by the use of a causal framework. To this end, we use DAGs to justify how to adjust for confounding variables in UR models, and prove mathematically in Appendix B that the resulting models satisfy the three properties of UR models (from Table 5.1).

5.4.1 Baseline confounding

We first consider the scenario in Figure 5.4, in which a baseline covariate M confounds the relationship between each of X_0, X_1 and Y.

Figure 5.4 Directed acyclic graph (DAG) depicting the hypothesised data-generating process for two measurements of a time-varying exposure X (i.e. X_0 and X_1), one outcome Y, and one baseline confounder M



The terms e_{X0}, e_{X1} and e_Y represent all unexplained causes of X_0, X_1 and Y, respectively, and are included to explicitly reflect uncertainty in all endogenous nodes (whether modelled or not).

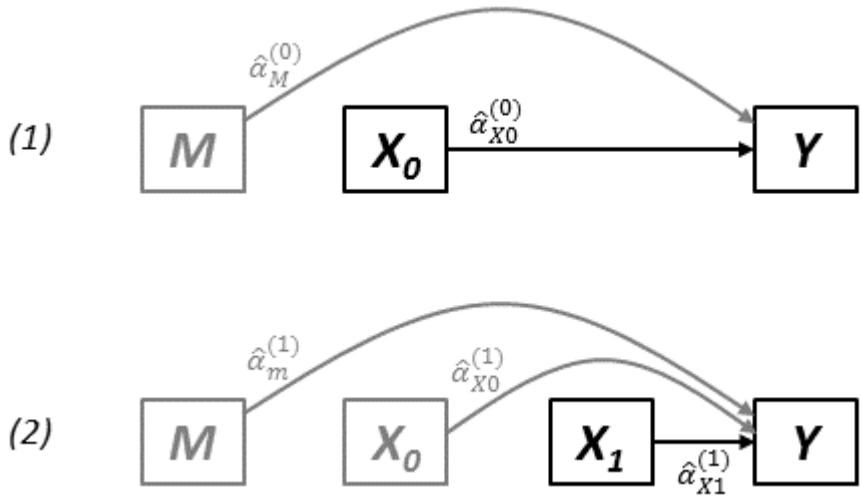
Because the relationship between each of X_0 and X_1 and Y is confounded by M *in addition to* any potential confounding by a previous measurement of X, M must be included in the standard regression models, as in:

$$\hat{Y}_S^{(0)} = \hat{\alpha}_0^{(0)} + \hat{\alpha}_M^{(0)} M + \widehat{\boldsymbol{\alpha}}_{\boldsymbol{X0}}^{(\mathbf{0})} X_0 \qquad \textbf{Equation 5.5}$$

$$\hat{Y}_S^{(1)} = \hat{\alpha}_0^{(1)} + \hat{\alpha}_M^{(1)} M + \hat{\alpha}_{X0}^{(1)} X_0 + \widehat{\boldsymbol{\alpha}}_{\boldsymbol{X1}}^{(\mathbf{1})} X_1 \qquad \textbf{Equation 5.6}$$

The total casual effects of X_0 and X_1 are represented by the coefficients $\hat{\alpha}_{X0}^{(0)}$ (Equation 5.5) and $\hat{\alpha}_{X1}^{(1)}$ (Equation 5.6), respectively; a visual depiction of these equations is given in Figure 5.5.

Figure 5.5 Path diagrams depicting the two standard regression models that would be constructed to estimate the total causal effect of each of X_0 and X_1 on Y in the presence of a baseline confounder M (i.e. Equation 5.5 and Equation 5.6, respectively)



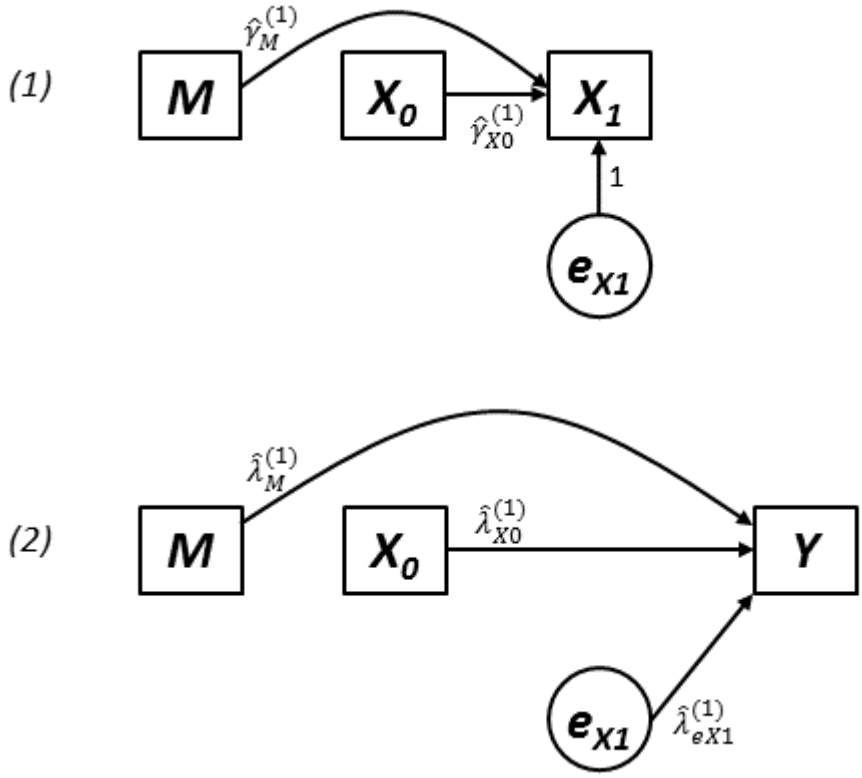
For each model, only the final coefficient may be interpreted as a total causal effect; all other coefficients are greyed to illustrate that no such interpretation should be made for them.

Fully adjusting for M in the UR modelling process requires that M be adjusted *both* when creating the UR term *and* in the UR model, as summarised in Figure 5.6. As is evident from the DAG in Figure 5.4, M confounds the relationship between X_0 and X_1; therefore, when X_1 is regressed on X_0 to produce the UR term, M must also be included as a covariate:

$$X_1 = \hat{\gamma}_0^{(1)} + \hat{\gamma}_M^{(1)} M + \hat{\gamma}_{X0}^{(1)} X_0 + e_{X1} \qquad \textbf{Equation 5.7}$$

In this way, the UR term e_{X1} represents the difference between the actual value of X_1 and the value of X_1 as predicted by X_0, *adjusted for the confounding effect of* M.

Figure 5.6 Path diagrams depicting the two steps of constructing a UR model in the presence of a baseline confounder M



Step (1): the preparatory regression of X_1 on X_0 (Equation 5.7); and Step (2): the UR model (Equation 5.8).

Moreover, because M confounds the effect of X_0 on Y, it is necessary to adjust for M in the subsequent UR model:

$$\hat{Y}_{UR}^{(1)} = \hat{\lambda}_0^{(1)} + \hat{\lambda}_M^{(1)} M + \hat{\boldsymbol{\lambda}}_{\boldsymbol{X0}}^{(\mathbf{1})} X_0 + \hat{\boldsymbol{\lambda}}_{\boldsymbol{eX1}}^{(\mathbf{1})} e_{X1} \qquad \textbf{Equation 5.8}$$

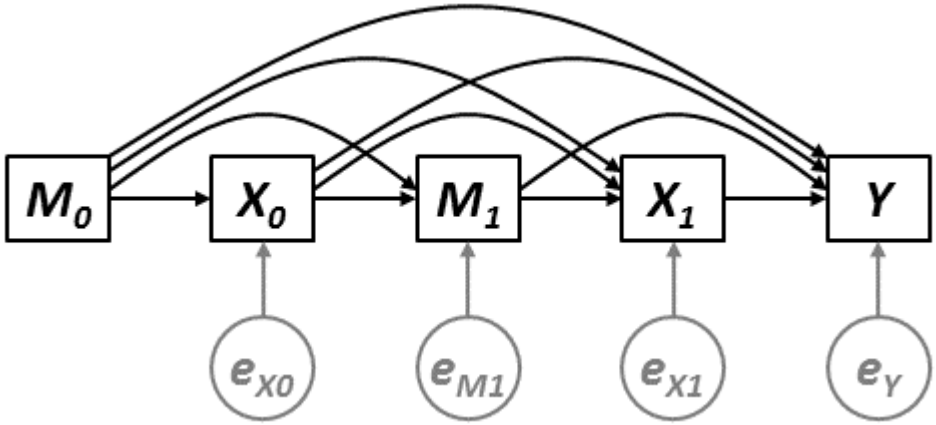
The UR model given in Equation 5.8 thus represents the outcome Y as function of the initial value of the exposure X_0, the subsequent 'unexplained' increase e_{X1}, and the baseline confounder M. It can be proven (Appendix B, §B.5, with $T = 2$) that this model satisfies the three properties of UR models.

By considering the DAG in Figure 5.4 as a path diagram (as in §5.3), we can again visualise the properties of a UR model correctly adjusted for a baseline confounder M. A regression model containing M, X_0, and X_1 (as in Equation 5.6) would only allow for the interpretation of the coefficient for X_1 as a total causal effect on Y; the coefficient for X_0 would represent only the direct effect on Y, because X_1 blocks the indirect path while M blocks the backdoor path between X_0 and Y. However, within the UR model, the independence of the UR term e_{X1} ensures no indirect paths are blocked and the only backdoor path between X_0 and Y is blocked by M.

5.4.2 Time-dependent confounding

We finally consider the scenario in Figure 5.7, in which a time-varying covariate M_0, M_1 confounds the relationship between each of X_0, X_1 and Y.

Figure 5.7 Directed acyclic graph (DAG) depicting the hypothesised data-generating process for two measurements of a time-varying exposure X (i.e. X_0 and X), one outcome Y, and two measurements of a time-dependent confounder M (i.e. M_0 and M_1)



The terms e_{X0}, e_{M1}, e_{X1} and e_Y represent all unexplained causes of X_0, M_1, X_1 and Y, respectively, and are included to explicitly reflect uncertainty in all endogenous nodes (whether modelled or not).

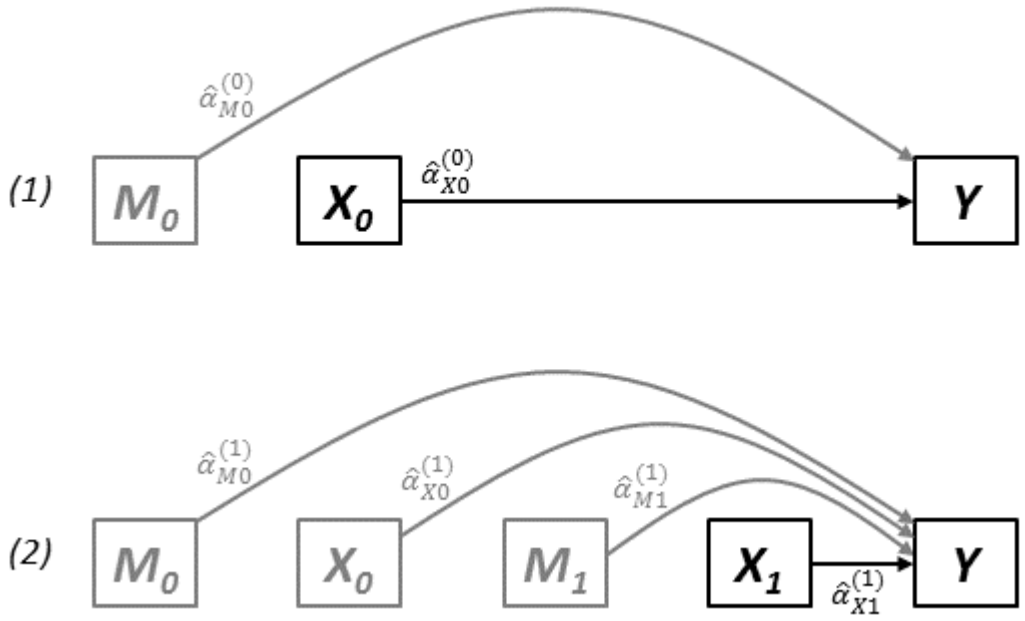
The relationship between each of X_0, X_1 and Y is not only potentially confounded by previous a measurement of X, but also by the current measurement and all previous measurements of the time-dependent confounder M. Thus, the standard regression models necessitate adjustment for M_0 and M_1 in the following way:

$$\hat{Y}_S^{(0)} = \hat{\alpha}_0^{(0)} + \hat{\alpha}_{M0}^{(0)} M_0 + \hat{\boldsymbol{\alpha}}_{\boldsymbol{X0}}^{(\mathbf{0})} X_0 \qquad \textbf{Equation 5.9}$$

$$\hat{Y}_S^{(1)} = \hat{\alpha}_0^{(1)} + \hat{\alpha}_{M0}^{(1)} M_0 + \hat{\alpha}_{X0}^{(1)} X_0 + \hat{\alpha}_{M1}^{(1)} M_1 + \hat{\boldsymbol{\alpha}}_{\boldsymbol{X1}}^{(\mathbf{1})} X_1 \qquad \textbf{Equation 5.10}$$

The total causal effects of X_0 and X_1 are represented by the coefficients $\hat{\alpha}_{X0}^{(0)}$ (Equation 5.9) and $\hat{\alpha}_{X1}^{(1)}$ (Equation 5.10), respectively; visual depictions are provided in Figure 5.8.

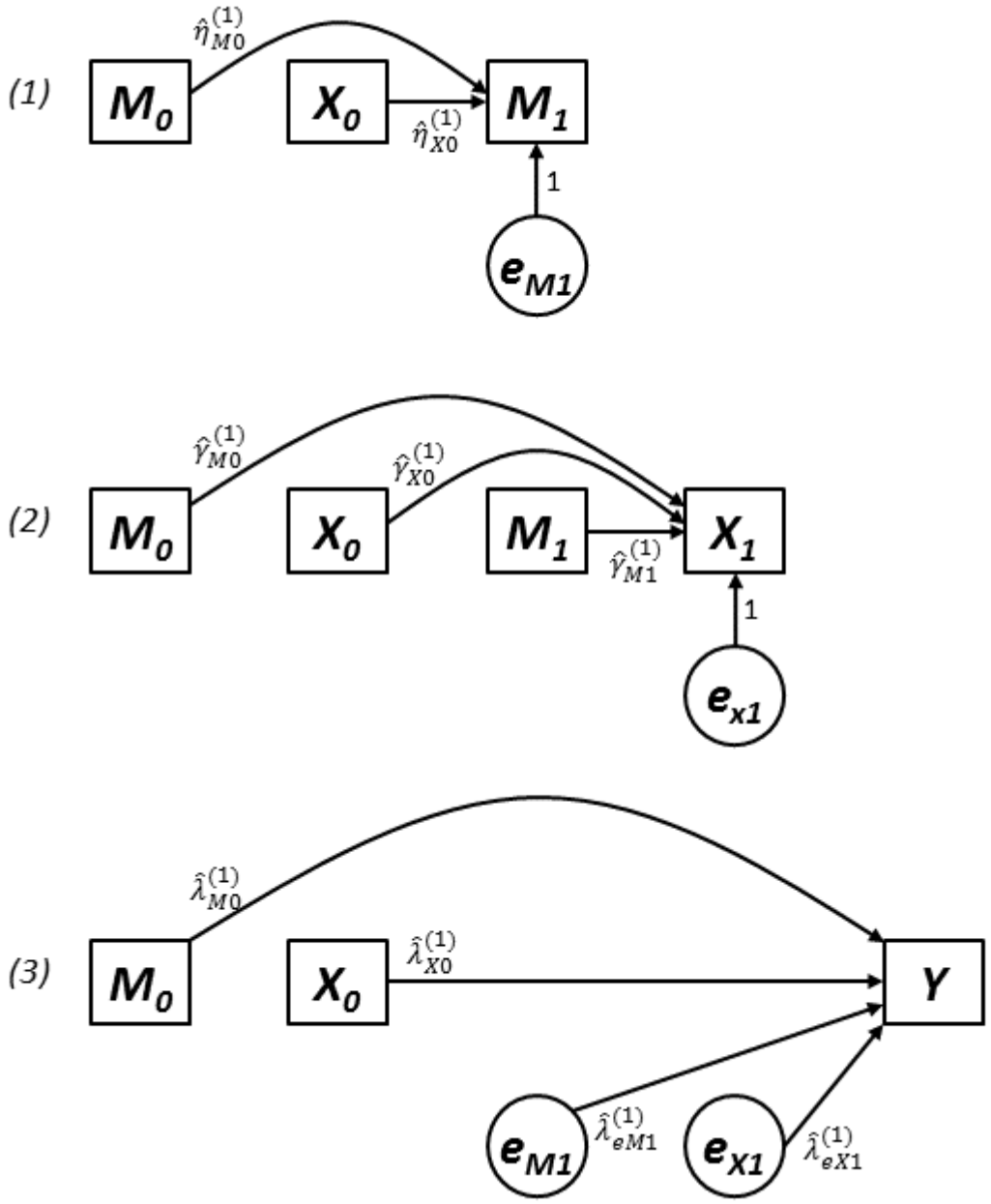
Figure 5.8 Path diagrams depicting the two standard regression models that would be constructed to estimate the total causal effect of each of X_0 and X_1 on Y in the presence of time-dependent confounders M_0 and M_1 (i.e. Equation 5.9 and Equation 5.10, respectively)



For each model, only the final coefficient may be interpreted as a total causal effect; all other coefficients are greyed to illustrate that no such interpretation should be made for them.

Whereas adjustment for the time-fixed covariate M is relatively straightforward using standard regression methods, extending the UR modelling process to accommodate M_0 and M_1 requires a significant extension to the original process formulated by Keijzer-Veen, M.G. et al. (93). This process is summarised visually in Figure 5.9.

Figure 5.9 Path diagrams depicting the three steps of constructing a UR model in the presence of time-dependent confounders M_0 and M_1



Step (1): the preparatory regression of M_1 on M_0 and X_0 (Equation 5.11); Step (2): the preparatory regression of X_1 on M_0, X_0, and M_1 (Equation 5.12); and Step (3): the UR model (Equation 5.13).

The introduction of a time varying confounder necessitates the creation of UR terms for *both* the confounder M_1 *and* the exposure X_1 (i.e. e_{M1} and e_{X1}, respectively). This is because UR models rely upon the orthogonality of the terms included in the model post-baseline. To this end, the UR term e_{M1} is constructed by regressing M_1 on all previous variables M_0 and X_0, as in:

$$M_1 = \hat{\eta}_0^{(1)} + \hat{\eta}_{M0}^{(1)} M_0 + \hat{\eta}_{X0}^{(1)} X_0 + e_{M1} \qquad \textbf{Equation 5.11}$$

Similarly, the UR term e_{X1} is constructed by regressing X_1 on all previous variables M_0, X_0, and M_1, as in:

$$X_1 = \hat{\gamma}_0^1 + \hat{\gamma}_{M0}^{(1)} M_0 + \hat{\gamma}_{X0}^{(1)} X_0 + \hat{\gamma}_{M1}^{(1)} M_1 + e_{X1} \qquad \textbf{Equation 5.12}$$

The UR terms e_{M1} and e_{X1} thus represent the difference between the observed value of their respective variables and the value of those variables as predicted by all previous measurements.

The UR model, then, represents Y as a function of the initial value of the confounder M_0 and the exposure X_0, and the subsequent 'unexplained' increases e_{M1} and e_{X1}, respectively:

$$\hat{Y}_{UR}^{(1)} = \hat{\lambda}_0^{(1)} + \hat{\lambda}_{M0}^{(1)} M_0 + \hat{\boldsymbol{\lambda}}_{\boldsymbol{X0}}^{(\mathbf{1})} X_0 + \hat{\lambda}_{eM1}^{(1)} e_{M1} + \hat{\boldsymbol{\lambda}}_{\boldsymbol{eX1}}^{(\mathbf{1})} e_{X1} \qquad \textbf{Equation 5.13}$$

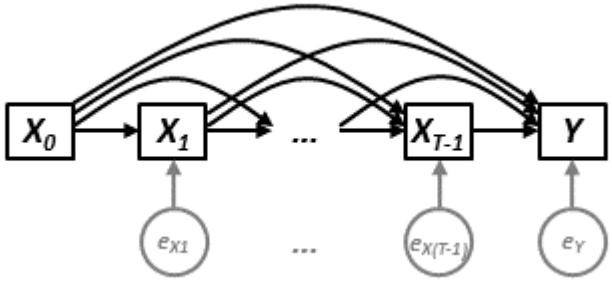
It can be proven (Appendix B, §B.6, with $T = 2$) that this model satisfies the three properties of UR models.

While it may not seem immediately obvious as to why UR terms must be created for both the exposure and the time-dependent confounder in this case, considering the DAG in Figure 5.7 as a path diagram (as previously in §5.3 and §5.4.1) sheds light on this. A regression model containing all of M_0, X_0, M_1, X_1 (as in Equation 5.10) does not allow for the interpretation of the coefficient of X_0 as a total causal effect, because the indirect paths between X_0 and Y are blocked by both M_1 and X_1. However, when we create the UR terms e_{M1} and e_{X1} which are independent of X_0 by construction, they do not block any of the effect of X_0 on Y.

5.5 Extension of UR models to a time-varying exposure measured at T time points

Although the method of using unexplained residuals was originally formulated and introduced for a scenario involving just two measurements of a time-varying exposure (i.e. Figure 5.1), many authors have also extended the method *ad-hoc* to accommodate scenarios involving T measurements of a time-varying exposure, as in Figure 5.10. This has resulted in different methods for deriving the UR terms $e_{X1}, e_{X1}, \dots, e_{X(T-1)}$, and uncertainty about which is in fact correct. For example, Horta, B.L. et al. (173), Gandhi, M. et al. (175), and Toemen, L. et al. (168) derived each UR term e_{Xt} by regressing each measured value of the exposure X_t on all previous measurements $X_0, X_1, \dots, X_{t-1}$, for $1 \leq t \leq (T-1)$. In contrast, Hardy, R. *et al.* (180) derived each UR term e_{Xt} by regressing each measured value of the exposure X_t on only the previous measure X_{t-1}, for $1 \leq t \leq (T-1)$.

Figure 5.10 DAG depicting the hypothesised data-generating process for T measurements of a time-varying exposure X (i.e. $X_0, X_1, \dots, X_{T-1}$) and one outcome Y.



The terms $e_{X1}, \dots, e_{X(T-1)}, e_Y$ represent all unexplained causes of $X_1, \dots, X_{T-1}, Y$, respectively, and are included to explicitly reflect uncertainty in all endogenous nodes (whether modelled or not).

Determining how to extend the method of unexplained residuals to a time-varying exposure measured at T time points (as in Figure 5.10) is greatly aided by the use of DAGs, as previously.

To create a UR model for the scenario depicted in Figure 5.10, each UR term e_{Xt} should be derived from the regression of each measured value of the exposure X_t on all previous measurements $X_0, X_1, \dots, X_{t-1}$, for $1 \leq t \leq (T-1)$, in order to maintain the orthogonality of the UR terms with all other terms in the model. The outcome Y should then be regressed on X_0 and all subsequent UR terms $e_{X1}, e_{X1}, \dots, e_{X(T-1)}$.

This can be formally proven mathematically, as in Appendix B (§B.4).

5.5.1 Confounding adjustment

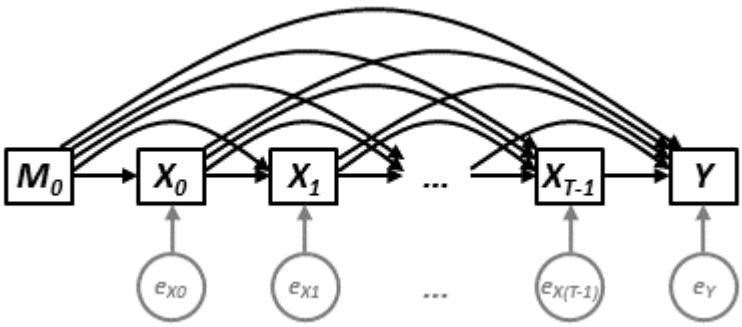
Here, we also summarise how to adjust for both baseline and time-dependent confounders in the extended UR models. These results follow naturally from the original scenarios considered previously (i.e. in §5.4.1 and §5.4.2), and so we only briefly outline the results.

Formal proofs are provided in Appendix B (§B.5 and §B.6, respectively).

5.5.1.1 Baseline confounding

Where there exists a baseline covariate M which confounds the relationship between each of $X_0, X_1, \dots, X_{T-1}$ and Y (Figure 5.11), each UR term e_{Xt} should be derived from the regression of each measured value of the exposure X_t on all previous measurements $X_0, X_1, \dots, X_{t-1}$, for $1 \leq t \leq (T-1)$, and on M. The UR model should then be constructed by regressing Y on M, X_0, and all subsequent UR terms $e_{X1}, e_{X2}, \dots, e_{X(T-1)}$.

Figure 5.11 DAG depicting the hypothesised data-generating process for T measurements of a time-varying exposure X (i.e. $X_0, X_1, \dots, X_{T-1}$), one outcome Y, and one baseline confounder M

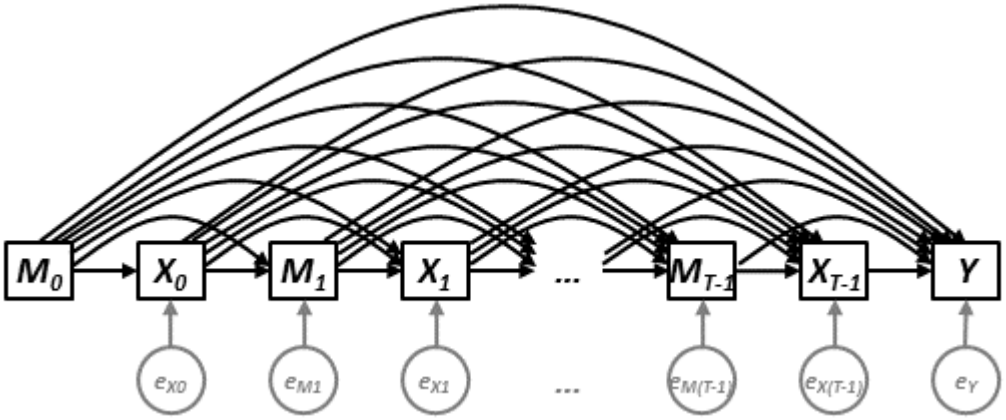


The terms $e_{X0}, e_{X1}, \dots, e_{X(T-1)}, e_Y$ represent all unexplained causes of $X_0, X_1, \dots, X_{T-1}, Y$, respectively, and are included to explicitly reflect uncertainty in all endogenous nodes (whether modelled or not).

5.5.1.2 Time-dependent confounding

Where there exists a time-varying covariate $M_0, M_1, \dots, M_{T-1}$ which simultaneously confounds and mediates the relationship between distinct measurements of X and Y (Figure 5.12), UR terms e_{Xt} and e_{Mt} must be created for all post-baseline measurements of the exposure X_t and confounder M_t, respectively, for $1 \leq t \leq (T-1)$. Each UR term e_{Mt} should be derived from the regression of each measured value of the time-dependent confounder M_t on all previous measurements of the confounder $M_0, M_1, \dots, M_{t-1}$ and on all previous measurements of the exposure $X_0, X_1, \dots, X_{t-1}$. Each UR term e_{Xt} should be derived from the regression of each measured value of the exposure X_t on all previous measurements of the exposure $X_0, X_1, \dots, X_{T-1}$ and on all previous measurements of the confounder $M_0, M_1, \dots, M_t$. Finally, the UR model should be constructed by regressing Y on M_0, X_0, and all subsequent UR terms $e_{M1}, \dots, e_{M(T-1)}, e_{X1}, \dots, e_{X(T-1)}$.

Figure 5.12 DAG depicting the hypothesised data-generating process for T measurements of a time-varying exposure X (i.e. $X_0, X_1, \dots, X_{T-1}$), one outcome Y, and T measurements of a time-varying exposure M (i.e. $M_0, M_1, \dots, M_{T-1}$)



The terms $e_{X0}, e_{M1}, e_{X1}, \dots, e_{M(T-1)}, e_{X(T-1)}, e_Y$ represent all unexplained causes of $X_0, M_1, X_1, \dots, M_{T-1}, X_{T-1}, Y$, respectively, and are included to explicitly reflect uncertainty in all endogenous nodes (whether modelled or not).

5.6 Artefactual standard error reduction using UR models

While a UR model is algebraically equivalent to its associated standard regression model (as in Equation 5.4 and Equation 5.2, respectively), a previously unexamined issue surrounding their use and implementation is that of an artefactual reduction in coefficient standard errors (SEs). Although focus on *statistical* significance by way of p-values and confidence intervals is not in and of itself justifiable within a causal framework (where focus is on effect sizes and likely *functional* significance, e.g. the absolute risk posed), the artificial precision of estimated effect sizes within a UR model must nevertheless be considered.

By definition, the SE of an estimated regression coefficient is a point estimate of the standard deviation of an (infinitely) large sampling distribution of estimated regression coefficients. Because standard regression and UR models elicit identical point estimates of the total causal effect of each measure of a time-varying exposure on the outcome of interest, it follows that the associated SEs should themselves be equal. However, this is not the case.

Standard regression models estimate the total causal effect of a particular measurement of the exposure using information from the past only (i.e. any past measures of the exposure plus any potential confounders). In contrast, UR modelling process generates (orthogonal) residuals for the entire exposure period and combines these into a single model, thereby using information that is from both the past and the future. If we possessed data pertaining to any true independent causes of future measurements of the exposure, this would be valid; however, the UR terms are simply estimated using prior measurements of the exposure. Moreover, the UR terms are *estimates* which thus consequently contain additional variation that is not accommodated by traditional regression methods. As a result, the SEs of estimated causal effects derived from UR models are artificially reduced and should not be inferred as robust.

Indeed, when the SEs within the UR models are estimated via bootstrapping, they are similar to those within the standard regression models.

5.6.1 Simulated example

To briefly demonstrate the artefactual standard error reduction that results from the use of UR models, we consider the simple example depicted in Figure 5.1, which involves two measurements of a time-varying exposure X (i.e. X_0 and X_1) and a subsequent outcome Y.

5.6.1.1 Method

1,000 non-overlapping random samples of 1,000 observations from a multivariate normal distribution were simulated based upon the DAG in Figure 5.1 using the 'dagitty' package (v. 0.2-2) in R (v. 3.3.2) (46, 47, 161).

For each sample, the following steps were carried out:

1. The two standard regression models necessary for estimating the total causal effect of each of X_0, X_1 on Y (Equation 5.1 and Equation 5.2, respectively) were created;
2. The UR term e_{X1} was derived by regressing X_1 on X_0 (Equation 5.3); and
3. The UR model in which Y is regressed on X_0 and e_{X1} (Equation 5.4) was created.

For each standard regression model, the reported SE of the regression coefficient for the exposure (i.e. X_0 and X_1, respectively) was stored. For each UR model, the SE of the regression coefficients for each of X_0 and e_{X1} was stored in two forms: (1) as reported in the UR model summary output; and (2) as estimated by bootstrapping 1000 samples and calculating the standard deviation of the distribution of estimated coefficients.

Additional details relating to this simulation, including parameters and code, can be found in Appendix B (§B.7).

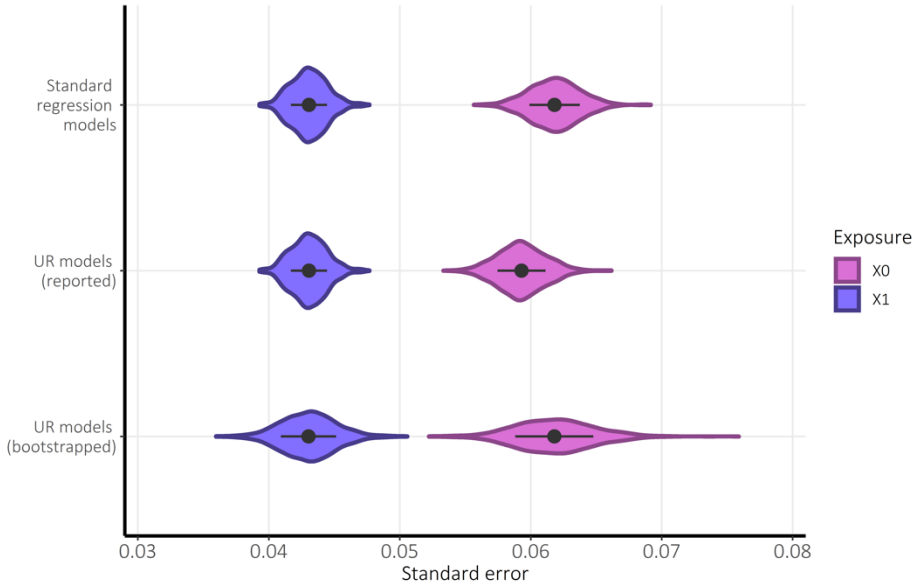
5.6.1.2 Results and discussion

Violin plots of the SEs estimated for each coefficient representing a total causal effect across the 1000 simulations are displayed in Figure 5.13 for each method considered.

As is evident from Figure 5.13, the *reported* SEs within the UR models are reduced in comparison to those within the first standard regression models (for exposure X_0) and equal to those within the final standard regression models (for exposure X_1).

Although the magnitude of bias in estimated SEs is small in this simulated example, it will always be present due to the way in which UR models are constructed. Quantifying the magnitude of this bias is not trivial and is beyond the scope of the present research, but it is worth noting that the degree of bias will increase as the number of measurements of the time-varying exposure and/or time-dependent confounders increases (i.e. as more orthogonal terms are added to the UR model).

Figure 5.13 Violin plots comparing the standard errors (SEs) associated with equivalent coefficients estimated in standard regression vs. UR models



Data were simulated based upon the scenario depicted in Figure 5.1. Horizontal bars within each distribution represent the mean ± 1 standard deviation.

5.7 Implications

We have demonstrated that UR models are able to quantify the total causal effect of multiple measurements of a time-varying exposure on a subsequent outcome *in a single model*, even in the presence of baseline and time-dependent confounding. However, the modelling process is substantially more complex to implement than standard regression methods. Although only one UR model need ultimately be presented, the necessity of generating orthogonal UR terms for all post-baseline variables requires that multiple models be created. In fact, the total number of models created by the UR process will always be either equal to or greater than the total number of models created by the standard regression process.

For an exposure X measured at T points in time, the standard regression approach necessitates T separate models for estimating the total causal effect of each measurement on the outcome *regardless of the number of confounders*. In the case of one time-invariant confounder (Figure 5.11), T models are also created (i.e. $T-1$ models to generate all UR terms and one UR model); for a time-dependent confounder (Figure 5.12), $2T-1$ models are created (i.e. $2T-2$ models to generate all UR terms and one UR model). Where there exist multiple, causally-linked confounders, the number of intermediate or 'preparatory' models increases by orders of magnitude.

If the additional complexity of UR models were offset by true gains in insight into the scenario under consideration, the method may in fact be preferred to standard regression methods. However, as has been demonstrated, they offer no additional insights compared to standard regression methods, and indeed the additional challenges associated with implementing them may result in additional errors. Moreover, UR models may create unwarranted confidence in

the precision of estimated effect sizes. Previous research that has utilised UR models without undertaking sufficient adjustment for confounders and correcting standard errors via bootstrapping should not be considered robust.

5.8 Summary

Regression with 'unexplained residuals' was introduced as a method to circumvent the need for multiple standard regression models to estimate the total causal effects of multiple measurements of a time-varying exposure on a subsequent outcome, and to 'solve' the potential interpretational challenges associated with multiple models. This chapter demonstrates that DAGs are useful for understanding the properties of UR models and determining how to correctly adjust for confounding, which has allowed for the benefits and drawbacks of the method to be fully evaluated against traditional regression approaches. Using DAGs, the benefits of UR models have been demonstrated to be little more than illusory, as the method provides no additional insight compared to standard regression methods. Moreover, the additional complexity required to implement them (particularly in the case of time-dependent confounding) makes them more vulnerable to analytical and interpretational issues.

Chapter 6
Microsimulation modelling

6.1 Introduction

Estimating the causal effect of a time-varying exposure on a subsequent outcome is both theoretically and computationally challenging. Standard regression methods and UR models (considered previously in Chapter 5) are able to estimate causal effects in this context; however, they do so by considering separate measurements of the exposure as if they were distinct entities, which remains unsatisfactory in many situations. Microsimulation models (MSMs) have been identified as promising tools for considering multiple measurements of the exposure together as an *exposure regime*, and parallels between microsimulation and the g-formula have been recognised previously (99).[22] However, the conditions under which MSMs provide robust estimates of causal effects are not well understood, nor are the unique challenges presented by simulation approaches fully appreciated.

Chapter 6 considers MSMs within a formal causal framework. This allows us to draw explicit parallels between the data-generating processes modelled in MSMs and those represented by DAGs, in order to consider the different issues associated with modelling this process using microsimulation compared to the g-formula. To this end, we simulate a longitudinal population for which the data-generating process is known, and interrogate it from the perspective of both methods. We demonstrate how the process of constructing an MSM might be improved by using DAGs, and investigate how reliable and/or robust microsimulation and the g-formula might be for longitudinal studies where the data-generating process is mis-specified to varying degrees. This chapter establishes a framework and simulation template for the evaluation of longitudinal methods from a sensitivity perspective for methods robustness.

6.1.1 Chapter overview

In Section 6.2, we introduce the key features and concepts of microsimulation. We emphasise important similarities between the data-generating processes modelled in MSMs and those represented by DAGs (§6.2.1), and we illustrate how an MSM can be represented as a DAG in the context of a specific example scenario (§6.2.1.1). Additionally, we highlight key differences between the g-formula and microsimulation (§6.2.2).

In Section 6.3, we outline the implications which result from the differences between the g-formula and microsimulation, specifically those which pertain to the relative importance of faithfully modelling the true data-generating process of the target population. This sets the

[22] MSMs are also commonly referred to as 'state transition models' (136), 'decision analytic models' (136, 146), and Markov models (181), particularly in clinical decision analysis and health-economic evaluation.

stage for (and provides the rationale behind) the simulations which are presented in the following section.

In Section 6.4, we describe and present the results of a simulation in which we evaluated how methodological differences between the g-formula and microsimulation affected estimation of causal effects. We simulate a longitudinal population for which the data-generating process is known (§6.4.1), and interrogate it from the perspective of both the g-formula and microsimulation (§6.4.2).

In Section 6.5, we discuss the results of our findings, including limitations and areas for future work (§6.5.1).

6.2 Microsimulation models (MSMs)

MSMs simulate an artificial population of heterogeneous individuals, typically over a long time horizon that extends into the future.[23] Each individual in the model possesses a set of attributes or 'states' (e.g. physical, socio-demographic, geographic), which may be updated throughout the simulation; in particular, individuals are often defined as belonging to one of a finite number of mutually exclusive and collectively exhaustive states, and events of interest are modelled as transitions from one state to another that occur according to a set of deterministic and/or stochastic rules (i.e. 'transition probabilities') (136, 141, 181, 182). The parameters governing transitions between states often depend on an individual's characteristics and possibly on previous history, and these parameters are typically estimated from a wide range of data sources, such as cohort studies, population-based epidemiological studies, and RCTs (136, 183).

MSMs may be either case-based or time-based (184). In a case-based model, individuals are simulated one at a time through all time points; in a time-based model, all simulated individuals are transitioned simultaneously through the model. Both methods produce equivalent results where there are no interactions amongst individuals, but time-based models tend to be more computationally efficient since they can easily be vectorised (185). Additionally, MSMs may be modelled in either discrete or continuous time (184). In a discrete-time model, transitions between states occur at discrete time steps; in a continuous-time model, the duration between state transitions is modelled in continuous time. In this chapter, we focus only on discrete-time MSMs, as they share natural parallels with the causal data structures considered throughout this thesis.[24]

---

[23] Note that the term 'microsimulation' may also refer to the process by which a cross-sectional snapshot of a population is created by generating a synthetic set of individuals whose characteristics match aggregate, area-level statistics; this type of microsimulation is referred to as '*spatial microsimulation*' (95). However, the focus of this chapter relates primarily to microsimulation which is explicitly longitudinal.

[24] The depiction of continuous time and competing events using DAGs is complicated by their discretised nature, in which variables are required to have a clear time ordering. There is new

An MSM may simply be used to model the 'natural history' of the population, which describes the progression of the population under no exogenous intervention (182); such a model might be used for the purposes of population projection, for instance (184). Additionally, an MSM may also be used to model 'counterfactual histories', which describe the progression of the population under various hypothetical interventions (98). This has historically made MSMs important tools for policy evaluation (3).

6.2.1 Representing an MSM as a DAG

A key aspect of microsimulation is *evolution*, which is a concept closely related to data-generating processes. For instance, Ryder, N.B. (188) note that the focus is on 'events rather than things, processes rather than states'. This is echoed by van Imhoff, E. and W. Post (189), who argue that an MSM should not only specify *what* the population will look like at some future point in time, but also *how* it gets there.

There exist clear parallels between the data-generating processes modelled in MSMs and those represented by DAGs. At every time point in an MSM, each individual's characteristics may be updated according to some specified probabilities, which may themselves be conditional on any number of current and/or past characteristics; each characteristic may thus be thought of as having a conditional probability (or distribution) associated with it. Similarly, each variable in a DAG is hypothesised to have a probability (or distribution), conditional on the variables which directly cause it.

These similarities make representation of an MSM as a DAG useful and informative, as it helps to draw explicit parallels between the two processes and enables us to understand the conditions under which MSMs may provide valid causal effect estimates. In the following subsection, we illustrate how this might be done in the context of a specific example scenario.

6.2.1.1 Example scenario

We consider an example scenario involving eleven time periods (i.e. $T = 11$) and three variables: (1) sex (female or male); (2) obesity status (non-obese or obese); and (3) diabetes status (non-diabetic or diabetic). At baseline (i.e. $t = 0$), individuals possess a value for each of the three attributes. At each time t, for $1 \leq t \leq 10$, each individual's obesity and diabetes states may be updated according to some conditional probability. Specifically, obesity status at time t is conditional on sex, obesity status at time $t - 1$, and diabetes status at time $t - 1$; diabetes status at time t is conditional on sex, diabetes status at time $t - 1$, and obesity status at time t.

A DAG offers a useful way to visually summarise the aforementioned process, as in Figure 6.1. Panel (1) depicts the full data-generating process (i.e. for $0 \leq t \leq 10$). While correct, this representation may nevertheless be difficult to interpret due to the number of time points.

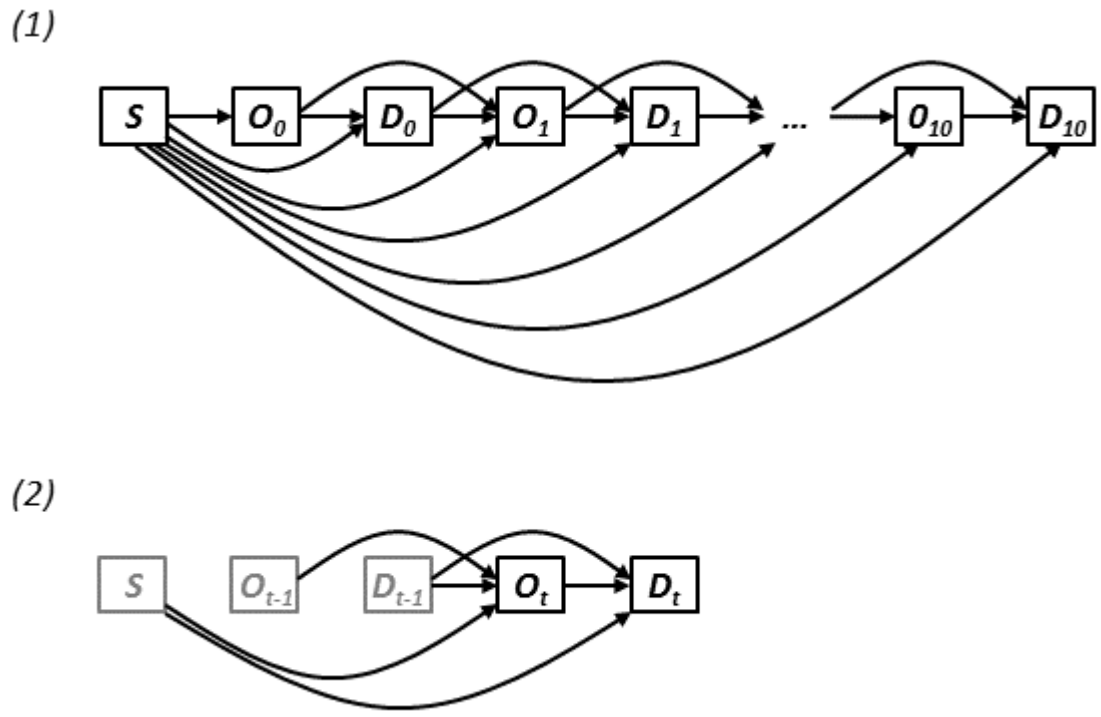
---

research being done in the area of causal inference in the presence of competing events (186, 187), but this is beyond the scope of the present research.

Therefore, in panel (2) we exploit the repeated nature of the data-generating process to produce a simplified representation for time t; variables depicted in grey are those which affect variables at time t but whose causes are not themselves represented in the graph. Panel (2) allows for easier visualisation of the data-generating process and identification of the conditional probabilities which govern it.

The data-generating process described in Figure 6.1 may be thought of as the 'natural history' of the population, as it represents the population under no exogenous intervention.

Figure 6.1 DAG representing the data-generating process for the variables sex (S), obesity (O), and diabetes (D), for $0 \leq t \leq 10$



In (1), the full DAG (i.e. for $0 \leq t \leq 10$) is depicted. In (2), only the data-generating process for time t is depicted; variables in grey are those which affect variables at time t but whose causes are not themselves represented in the graph.

6.2.2 Key differences between the g-formula and microsimulation

Parallels between the g-formula and microsimulation have been recognised by Murray, E.J. et al. (99), who describe the use of a 'similar mathematical approach: construction of a sequential model that is the basis for a Monte Carlo simulation of a (counterfactual) population under each treatment strategy of interest.' The g-formula involves modelling the observed joint distribution of the data, and then estimating the (counterfactual) distributions under various interventions to calculate causal effects. This can be related to an MSM, which models the 'natural history' of the population and then estimates the 'counterfactual histories' under various interventions. However, although the two are methodologically similar, there exist key differences which arise from their distinct historical evolutions (as outlined in Chapter 3).

The joint distribution of the data is generally unknown and cannot be directly estimated in a microsimulation model (189). This is because microsimulation models are often used to make general inferences about a population, and often in the future. This is in contrast to the g-formula, which makes inferences about a specific (often highly-selected) population from a single retrospective dataset (99). Thus, using the g-formula, the conditional probability of every variable *at every time point* in Figure 6.1 can be estimated from a single dataset. MSMs, however, do not have direct access to these probabilities. In this way, the g-formula may be thought of as a special case of microsimulation, in which we have access to the entire joint distribution of all variables and in which all parameters come from a single dataset.

6.3 The importance of faithfully modelling data-generating processes

The differences between the g-formula and microsimulation which were highlighted in the previous section (§6.2.2) have implications for the relative importance of faithfully modelling the data-generating process of the target population using each method.

Using microsimulation, the future distributions of states in the population must be generated by the repeated processes specified in the model. As described by van Imhoff, E. and W. Post (189), 'microsimulation models can be regarded as models which generate their own explanatory variables', thus underscoring the importance of modelling the true data-generating process. Likewise, DAG-based methods – including the g-formula – emphasise the importance of understanding and modelling data-generating processes in order to make causal inferences. However, the g-formula is potentially more robust to mis-specifications of the data-generating processes because it uses data from the entire distribution of data to estimate the observed (i.e. natural) and counterfactual distributions. Mis-specifications in the data-generating process in an MSM are likely to have more consequential onward effects which result in biased estimates of both natural and counterfactual histories.

Moreover, an MSM is arguably more likely to simplify the data-generating process being modelled due to the way in which it is constructed. Whereas implementing the g-formula requires *estimating* the parameters governing the conditional probability of each variable (at each time point) from a single dataset, implementing an MSM requires *specifying* the parameters governing these conditional probabilities to produce a plausible population. The additional challenges of specifying such parameters, which frequently must be combined from multiple datasets, may encourage simplification of the data-generating process. Indeed, Murray, E.J. et al. (183) highlight some of the challenges associated with parameterising direct effects in MSMs, including the lack of clear guidance on how to use published effect estimates to inform model construction.

In the following section, we present the results of a simple simulated example, in which we evaluated how methodological differences between the g-formula and microsimulation affected estimation of natural and counterfactual histories. In particular, we investigated how

robust each method was to varying degrees of mis-specifications of the data-generating process.

6.4 Simulated example

We conducted a simple simulation based on the example scenario described previously in Section 6.2.1.1, in order to demonstrate the importance of faithfully modelling true data-generating processes, and to assess the relative performance of the g-formula and microsimulation for estimating natural and counterfactual histories when the data-generating process was mis-specified to varying degrees. A brief overview of this simulation is given below.

First, we simulated a population for which the data-generating process was known. We simulated both the 'natural history' of this population and the 'counterfactual histories' of this population under six hypothetical interventions. These histories represent the *true* histories of the population, which any method should aim to faithfully replicate in order to estimate the true intervention effects. This simulation is described and its results presented in Section 6.4.1

We then used both the g-formula and microsimulation to model the true natural and counterfactual histories of the population in order to estimate the causal effects of the interventions in the population, using hypothesised data-generating processes which differ from the true data-generating process to varying degrees. When both methods correctly modelled the true data-generating process of the population, we expected them to perform equally well at replicating both the true natural and counterfactual histories of the population under intervention. Conversely, we expected both methods to be biased for estimating the true natural and counterfactual histories when the data-generating process was mis-specified; however, we expected the g-formula to be more robust to mis-specification since it utilised data from the population across all time points. These simulations are described and their results are presented in Section 6.4.2.

The results of our simulations are synthesised and discussed in Section 6.4.3, and the potential for substantive changes to our methodological findings through specific sensitivity analyses are explored in Section 6.4.4.

6.4.1 Simulation of a population according to the true data-generating process

We simulated a population for which the data-generating process was known, according to that which is depicted in the DAG in Figure 6.1. For simplicity, we simulated a closed population (i.e. where no individuals dropped out of the simulation and no new individuals entered it post-baseline).

In the following subsections, we describe our simulations relating to the 'natural history' of this population (§6.4.1.1) and the 'counterfactual histories' under six hypothetical interventions which targeted obesity (the 'exposure') (§6.4.1.2). We also calculate the true

population average causal effect of each intervention on diabetes prevalence at time 10 (the 'outcome') (§6.4.1.3).

6.4.1.1 Natural history

Longitudinal data for 5 million individuals were simulated using a discrete time, time-based microsimulation model in R (v.3.5.2) (190), according to the data-generating process depicted in Figure 6.1. The population size was chosen to be sufficiently large that all possible exposure and covariate histories would be represented, in order to capture the important features in the data-generating process.

Simulation parameters were chosen to produce a population whose characteristics approximately tracked the true overall and disaggregated (by sex) prevalence of obesity and diabetes in England, as reported by the Health Survey for England (HSE) for the years 1994 to 2004 (191, 192). Briefly, males were simulated to have a higher overall probability of obesity compared to females at baseline, and a higher probability of developing obesity (i.e. incident probability) at each time t; the effect of previous diabetes on the probability of obesity at time t was positive but modest for both males and females. Males were also simulated to have a higher overall probability of diabetes compared to females at baseline, and a higher probability of developing diabetes at each time t; obesity substantially increased the risk of developing diabetes for both males and females. Once an individual developed diabetes, he/she maintained that status for all subsequent time points.

Appendix C contains the simulation parameters (§C.2.1.1.1), a fuller description of the characteristics of the simulated population (§C.2.1.1.2), a comparison of the simulated population with HSE statistics (§C.2.1.1.3), and all annotated R code relating to this simulation (§C.2.1.1.4).

6.4.1.2 Counterfactual histories under hypothetical interventions

The effects on diabetes prevalence as observed at time 10 due to the following population interventions were simulated:

Intervention 1: Prevent anyone from being obese (i.e. reduce the incident and prevalent probabilities of obesity to zero).

Intervention 2: Make everyone obese (i.e. increase the incident and prevalent probabilities of obesity to one).

Intervention 3: Prevent any new individuals from becoming obese (i.e. reduce the incident probability to zero).

Intervention 4: Reduce the probability of becoming obese by 15% (i.e. reduce the incident probability by 0.15).

Intervention 5: Reduce the probability of remaining obese by 10% (i.e. reduce the prevalent probability by 0.10).

Intervention 6: Reduce the probability of becoming obese by 15% and reduce the probability of remaining obese by 10% (i.e. both Interventions 4 and 5 were combined).

As is evident, all interventions targeted diabetes prevalence indirectly – that is, they altered diabetes prevalence by altering obesity. Because diabetes could not be reversed in our simulation, and no individuals were 'lost to follow-up', no intervention on obesity could be expected to reduce diabetes prevalence to zero. Nevertheless, each intervention could be expected to modify diabetes prevalence to some degree.

All interventions were applied to the population at each time point post-baseline (i.e. for $1 \leq t \leq 10$). Because the simulation progresses via a series of stochastic events, each intervention was simulated fifty times, and the mean of all simulation runs was calculated to be the 'true' counterfactual population average history under the specified intervention.

Figure 6.2 displays obesity prevalence at each time point under each intervention, compared to the natural history of obesity prevalence in the population. Figure 6.3 displays diabetes prevalence at each time point under each intervention, compared to the natural history of diabetes prevalence in the population. Figure 6.4 displays the proportion of individuals with each combination of characteristics at each time point under intervention, compared to that of the natural history.

Appendix C contains a more thorough discussion of the effects of each intervention on obesity and diabetes prevalence (§C.2.1.2.2) and all annotated R code relating to this simulation (§C.2.1.2.3).

Figure 6.2 Obesity prevalence in the simulated population under Interventions 1 through 6, compared to the natural history

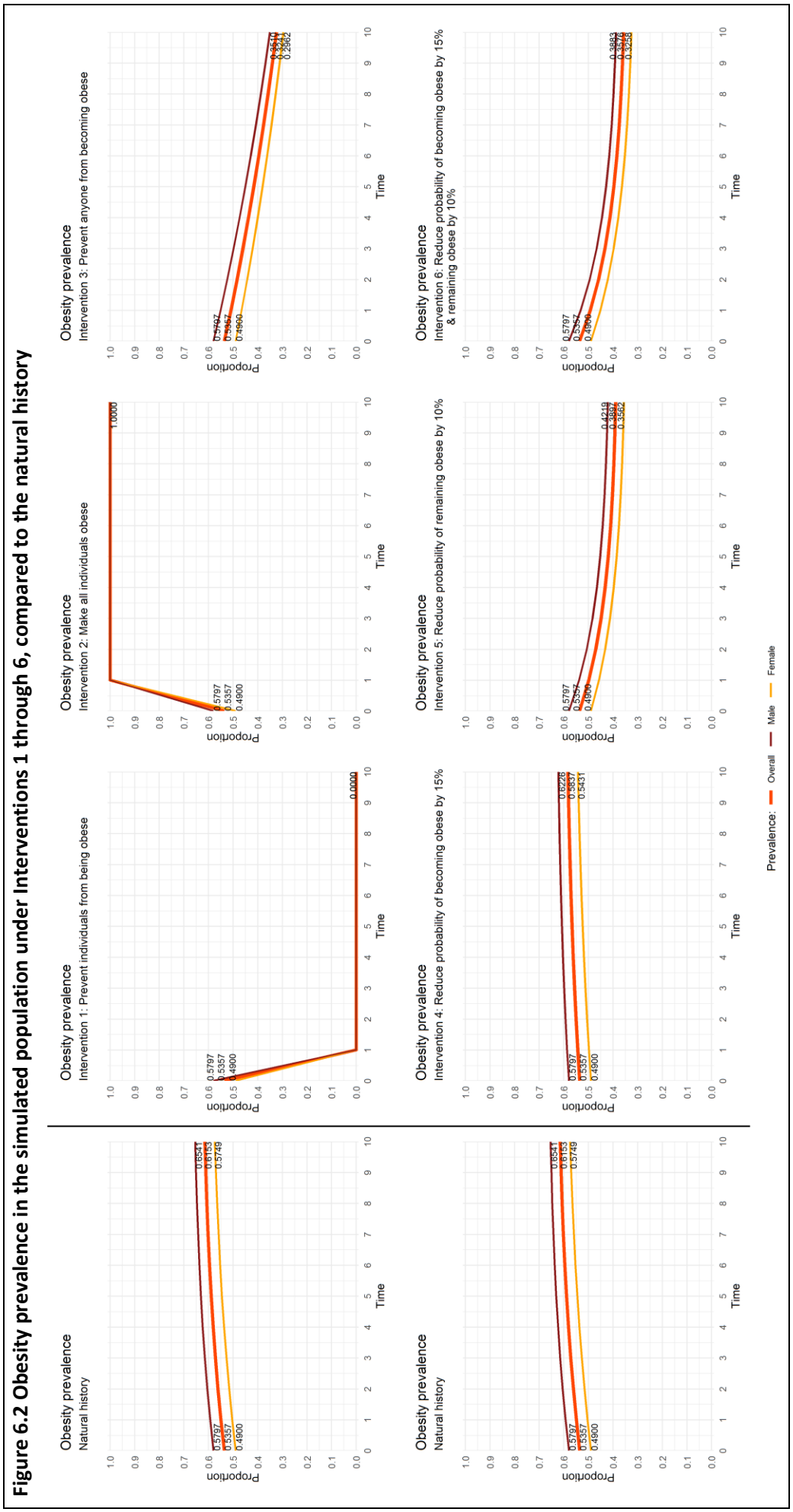


Figure 6.3 Diabetes prevalence in the simulated population under Interventions 1 through 6, compared to the natural history

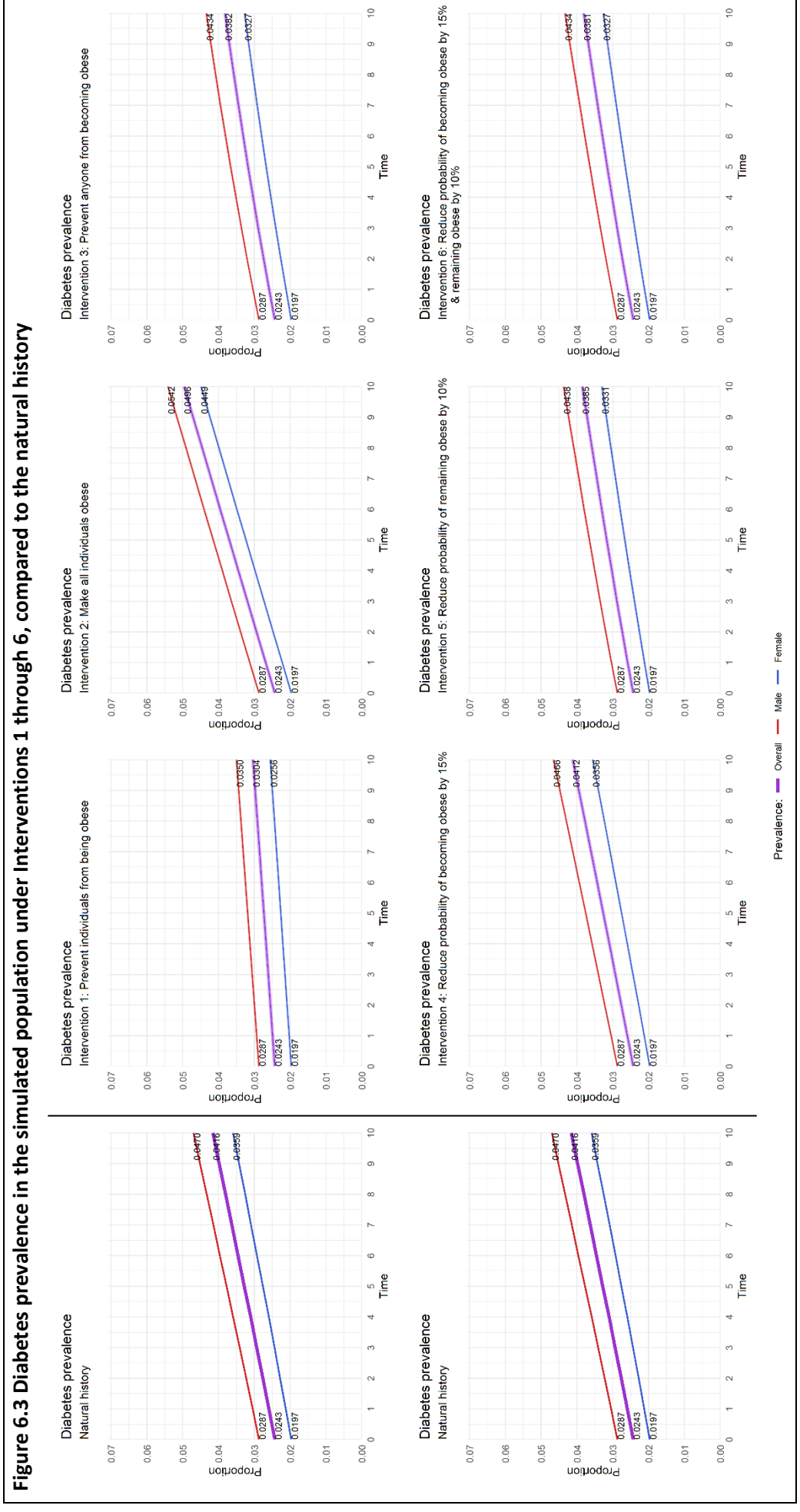
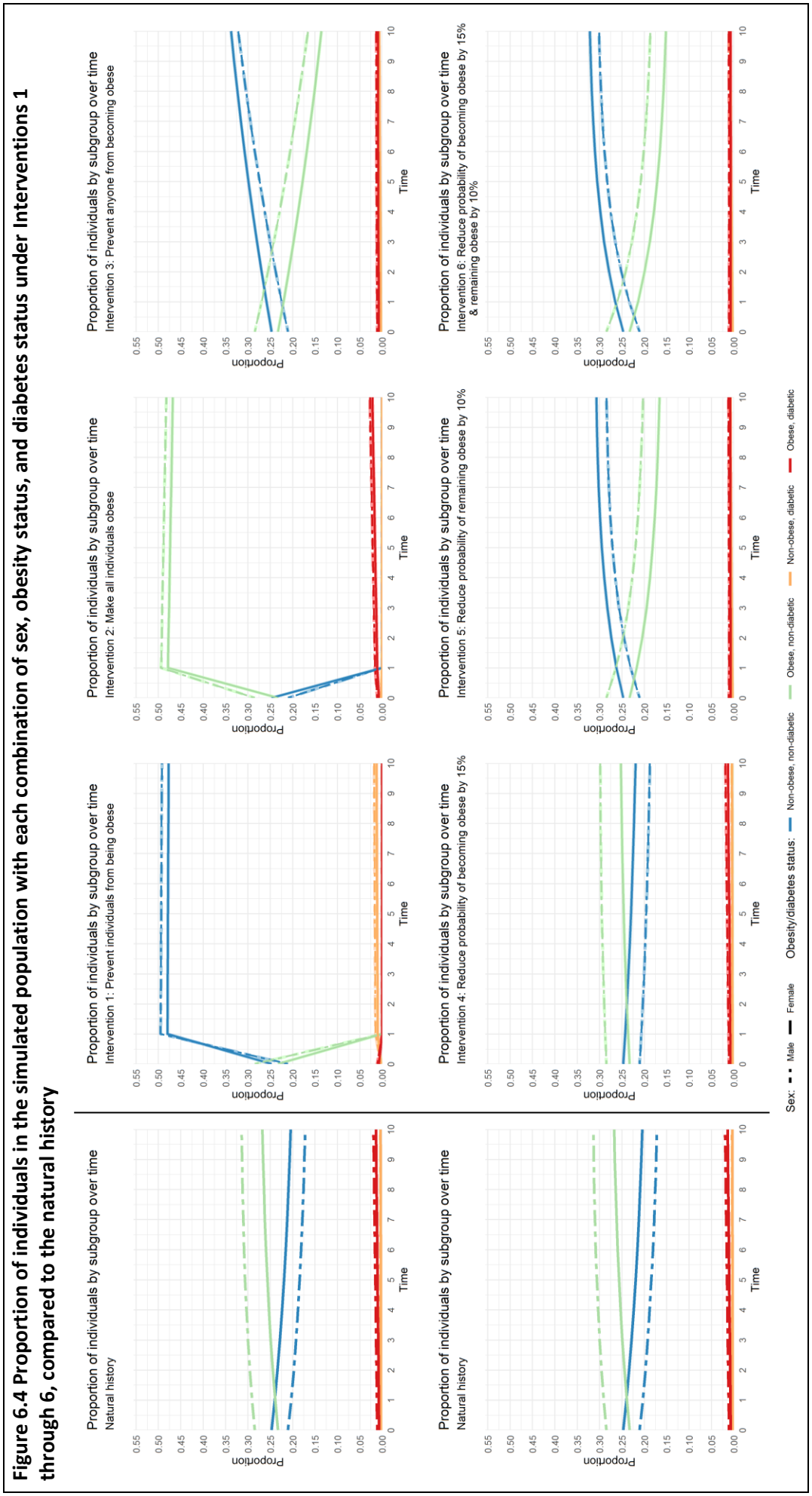


Figure 6.4 Proportion of individuals in the simulated population with each combination of sex, obesity status, and diabetes status under Interventions 1 through 6, compared to the natural history



6.4.1.3 True causal effects of interventions

We defined the true population average causal effect of each intervention on diabetes prevalence to be the difference between *natural* diabetes prevalence at time 10 (i.e. under the natural history) and mean diabetes prevalence at time 10 under that intervention (i.e. under the counterfactual history). For example, the true causal effect of Intervention 1 was calculated by subtracting the observed diabetes prevalence at time 10 under the natural history from the average diabetes prevalence at time 10 that was observed when Intervention 1 was applied to the population.

In addition, we defined the true population average total causal effect (TCE) of obesity on diabetes to be the difference between average diabetes prevalence at time 10 if everyone were obese at all time points post-baseline (i.e. for $1 \leq t \leq 10$) and average diabetes prevalence at time 10 if no one were obese at any time point post-baseline. The true TCE of obesity on diabetes was therefore calculated by subtracting the average diabetes prevalence at time 10 that was observed when Intervention 1 was applied to the population from that which was observed when Intervention 2 is applied.[25]

Table 6.1 contains the true causal effects of each intervention on diabetes prevalence, as well as the TCE of obesity on diabetes.

[25] This represents the standard definition of the TCE in the setting of a time-varying exposure (6). We acknowledge that it is an unrealistic effect to estimate from a practical or policy perspective, but we nevertheless believe it is important to calculate in our analyses because it is often of calculated in standard epidemiological analyses of time-varying exposures.

Table 6.1 Table describing the true population average causal effect of each intervention on diabetes prevalence in the simulated population

Effect	Value
Effect of Intervention 1 *(prevent individuals from being obese)*	-0.0112 *(-26.9%)*
Effect of Intervention 2 *(make all individuals obese)*	0.0080 *(19.3%)*
Effect of Intervention 3 *(prevent anyone from becoming obese)*	-0.0034 *(-8.2%)*
Effect of Intervention 4 *(reduce probability of becoming obese by 15%)*	-0.0004 *(-0.9%)*
Effect of Intervention 5 *(reduce probability of remaining obese by 10%)*	-0.0031 *(-7.4%)*
Effect of Intervention 6 *(reduce probability of becoming obese by 15% and remaining obese by 10%)*	-0.0035 *(-8.3%)*
Total causal effect (TCE)	0.0192 *(63.2%)*

The true causal effect of each intervention (1 through 6) on diabetes prevalence was calculated by subtracting the observed diabetes prevalence at time 10 under the natural history from the average diabetes prevalence at time 10 that was observed when the given intervention is applied to the population. The true TCE was calculated by subtracting the average diabetes prevalence at time 10 that was observed when Intervention 1 was applied to the population from that which is observed when Intervention 2 was applied. All effects are additionally expressed as percentage changes.

6.4.2 Comparison of the g-formula versus microsimulation for estimating true causal effects in the population

We investigated the robustness of the g-formula and microsimulation for replicating the true natural history of the population and the true counterfactual histories of the population under six hypothetical interventions which targeted obesity (described in §6.4.1.2), and the resulting robustness of each method for estimating the true population average causal effects (from Table 6.1).

A general description of each simulation follows. For each autocorrelation structure, a random sample of 20,000 individuals was drawn from the population, and the conditional probabilities of obesity and diabetes were estimated according to the autocorrelation structure under consideration. Using the g-formula, these conditional probabilities were estimated at each time t, for $1 \leq t \leq 10$; however, using microsimulation, these conditional probabilities were estimated at time 1 only. A random sample of 20,000 was then simulated through time 10 by applying the estimated conditional probabilities to estimate the natural history. The sample of 20,000 was also simulated under Interventions 1 through 6 by modifying the estimated

conditional probabilities accordingly. This process was repeated 100 times, and the mean history calculated to estimate the intervention effects.

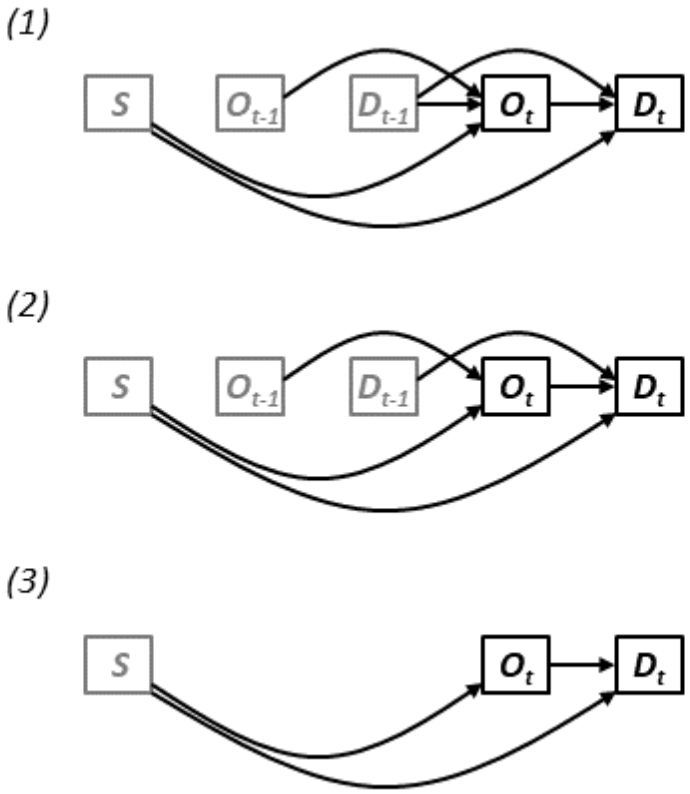
Because the g-formula utilises data about the conditional probability of each state at every time point, it was expected to produce a population whose *natural history* appears consistent with the true population for most of the autocorrelation structures considered. In contrast, microsimulation was expected to produce a population whose natural history appears consistent with the true population *only* when the true data-generating process was modelled, since it was unable to utilise data about the conditional probability of each state at every time point in the population. It was also expected that mis-specification of the data-generating process would adversely affect estimation of *counterfactual histories* using both methods, and therefore produce inaccurate estimates of intervention effects.

In the following subsections, we describe the three hypothesised data-generating processes which were considered (§6.4.2.1). We then describe results obtained by using the g-formula (§6.4.2.2) and microsimulation (§6.4.2.3) to model the natural and counterfactual histories of the population, according to each of the hypothesised data-generating processes.

6.4.2.1 Description of hypothesised data-generating processes

Three distinct autocorrelation structures (representing three distinct hypothesised data-generating processes) were considered, with each summarised visually as a DAG in Figure 6.5.

Figure 6.5 DAGs representing three hypothesised data-generating processes at time t for the time-varying variables obesity (O) and diabetes (D)



Each panel represents a different autocorrelation structure modelled by using the g-formula or microsimulation (AS1 through AS3, respectively). Variables in grey are those which affect variables at time t but whose causes are not themselves represented in the graph.

Autocorrelation structure 1 (AS1) represents the true data-generating structure of the population. Updated obesity status at time t is conditional on sex, current obesity status (O_{t-1}), and current diabetes status (D_{t-1}). Updated diabetes status at time t is conditional on sex, current diabetes status (D_{t-1}), and current obesity status (O_t).

Autocorrelation structure 2 (AS2) is nearly identical to AS1, but does not fully model the true data-generating process in the population. Updated obesity status at time t is conditional on sex and current obesity status (O_{t-1}), but not on current diabetes status (D_{t-1}). Thus, it does not fully model the time-dependent confounding that exists.

Autocorrelation structure 3 (AS3) does not model any dependence between adjacent time points. Updated obesity status at time t is conditional on sex only, whereas diabetes status at time t is conditional on sex and current obesity status (O_t) only. Thus, modelling AS3 is equivalent to randomly sampling from the cross-sectional joint distribution of sex, obesity, and diabetes (i.e. $P(S, O_t, D_t) = P(D_t|O_t, S) \cdot P(O_t|S) \cdot P(S)$) in the population.

These autocorrelation structures were chosen to represent a broad range of possible data-generating processes, though we acknowledge that not all are equally plausible in practice. For example, AS3 does not model any dependence between time points (except that which exists

due to sex), which contradicts clinical and biological knowledge about the conditions of obesity and diabetes. However, AS2 does represent a plausible hypothesis for the data-generating process, since the potentially small magnitude of time-varying confounding may be judged as trivial and not to warrant the added complexity introduced by additional model parameters. This is particularly relevant for MSMs, which rely on published estimates to inform model parameterisation (183).

Hereafter, we alternately refer to AS1 as the *correctly specified* data-generating process, AS2 as the *slightly mis-specified* data-generating process, and AS3 as the *highly mis-specified* data-generating process.

6.4.2.2 The g-formula

This subsection describes the results of using the g-formula to estimate the true natural and counterfactual histories of the population, according to the process described previously (§6.4.2).

6.4.2.2.1 Estimated causal effects of interventions

The estimated causal effect on diabetes prevalence of each intervention is given in Table 6.2, for each of AS1 through AS3, as modelled by the g-formula; the true effects in the population (from Table 6.1) are also given for comparison.

Both the correctly-specified and slightly mis-specified autocorrelation structures (i.e. AS1 and AS2, respectively) appeared to perform well in estimating the true effects of Interventions 1 through 6, while the highly mis-specified autocorrelation structure (i.e. AS3) performed relatively poorly. In the following subsections, we briefly discuss these results further by examining how well each autocorrelation structure replicated the true natural and counterfactual histories.

Table 6.2 Table describing the estimated causal effect of each intervention on diabetes prevalence, for each of A1 through A3 modelled using the g-formula, compared to the true effect in the population

Effect	True	AS1	AS2	AS3
Effect of Intervention 1 *(prevent individuals from being obese)*	**-0.0112** ***(-26.9%)***	-0.0115 *(-27.5%)*	-0.0115 *(-27.6%)*	-0.0206 *(-49.1%)*
Effect of Intervention 2 *(make all individuals obese)*	**0.0080** ***(19.3%)***	0.0080 *(19.3%)*	0.0079 *(19.0%)*	0.0124 *(-29.7%)*
Effect of Intervention 3 *(prevent anyone from becoming obese)*	**-0.0034** ***(-8.2%)***	-0.0034 *(-8.1%)*	-0.0036 *(-8.6%)*	-0.0207 *(-49.3%)*
Effect of Intervention 4 *(reduce probability of becoming obese by 15%)*	**-0.0004** ***(-0.9%)***	-0.0005 *(-1.2%)*	-0.0003 *(-0.8%)*	-0.0032 *(-7.6%)*
Effect of Intervention 5 *(reduce probability of remaining obese by 10%)*	**-0.0031** ***(-7.4%)***	-0.0031 *(-7.6%)*	-0.0032 *(-7.6%)*	-0.0024 *(-5.7%)*
Effect of Intervention 6 *(reduce probability of becoming obese by 15% and remaining obese by 10%)*	**-0.0035** ***(-8.3%)***	-0.0036 *(-8.8%)*	-0.0037 *(-8.8%)*	-0.0050 *(-12.0%)*
Total causal effect (TCE)	**0.0192** ***(63.2%)***	0.0195 *(64.6%)*	0.0194 *(64.4%)*	0.0330 *(154.7%)*

The estimated causal effect of each intervention (1 through 6) on diabetes prevalence was calculated by subtracting the average observed diabetes prevalence at time 10 under the natural history from the average diabetes prevalence at time 10 that was observed when the given intervention was applied to a random sample of 20,000 individuals. The TCE was calculated by subtracting the average diabetes prevalence at time 10 that was observed when Intervention 1 was applied from that which was observed when Intervention 2 was applied. All effects are additionally expressed as percentage changes.

Figure 6.6 Natural history of obesity and diabetes prevalence for each of the autocorrelation structures modelled using the g-formula, compared to the true natural history

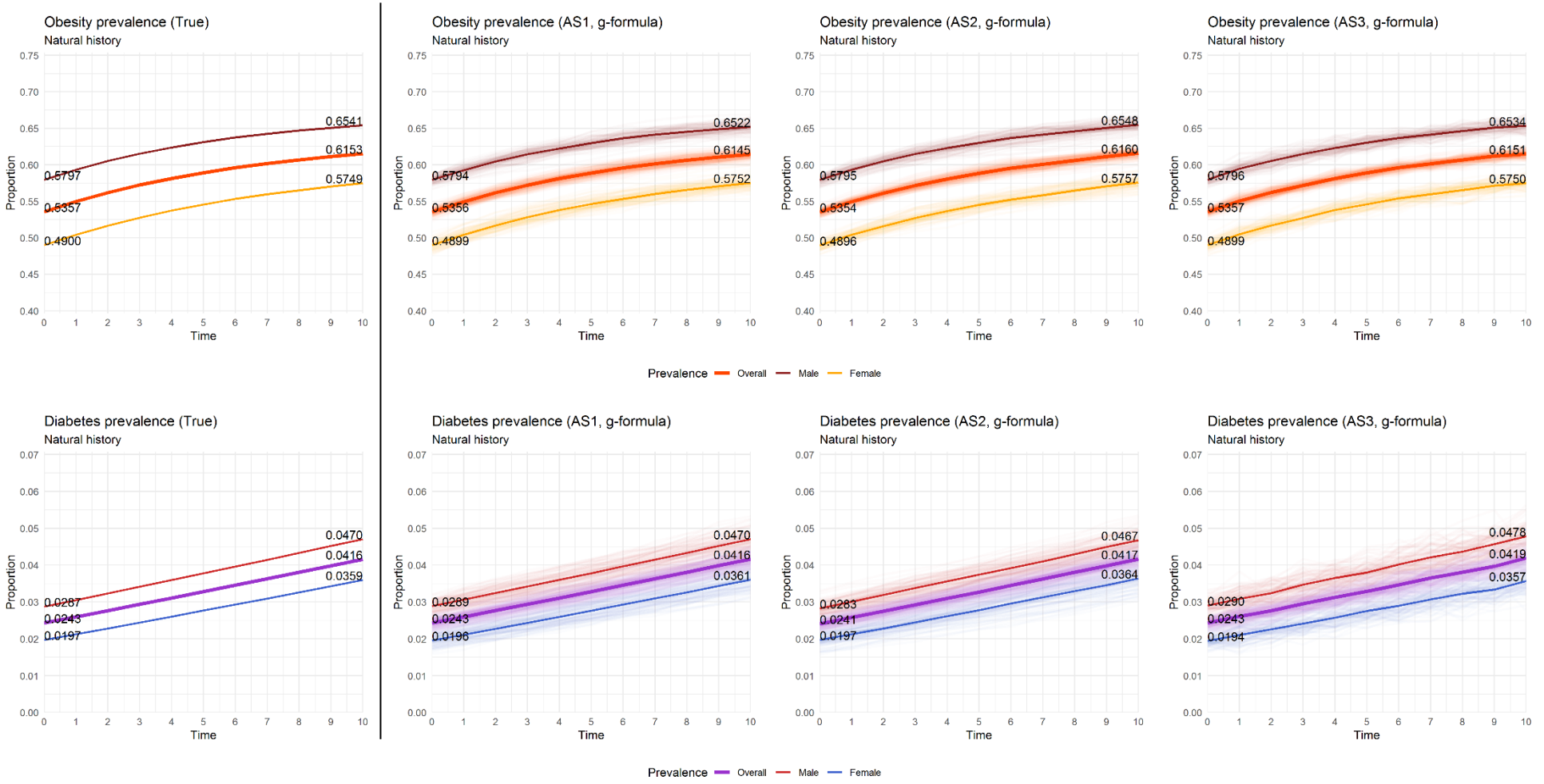


Figure 6.7 Natural history of the cross-sectional prevalence of sex, obesity, and diabetes, for each of AS1 through AS3 modelled using the g-formula, compared to the true natural history

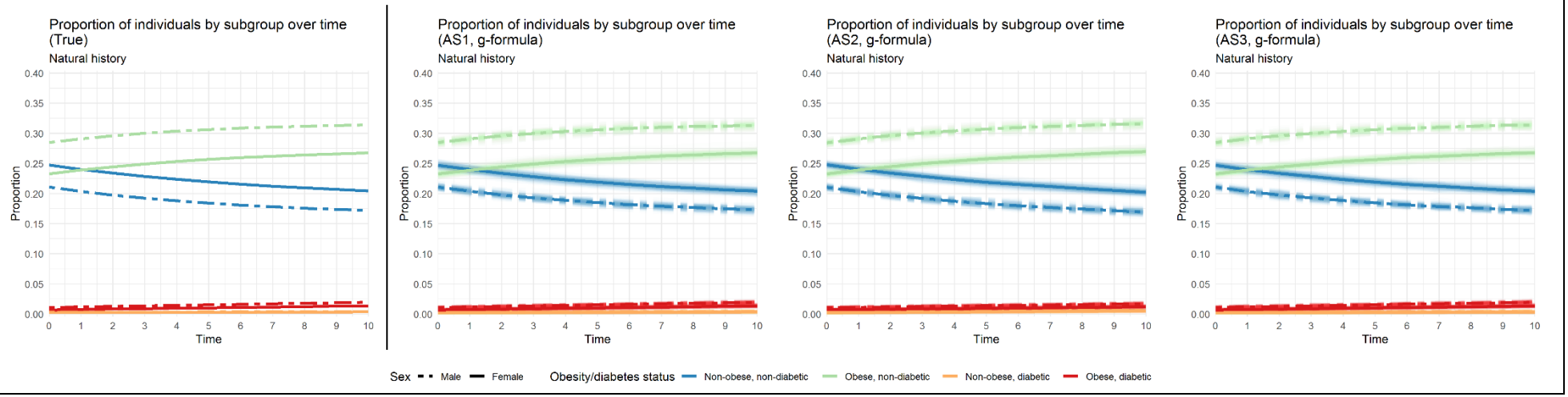
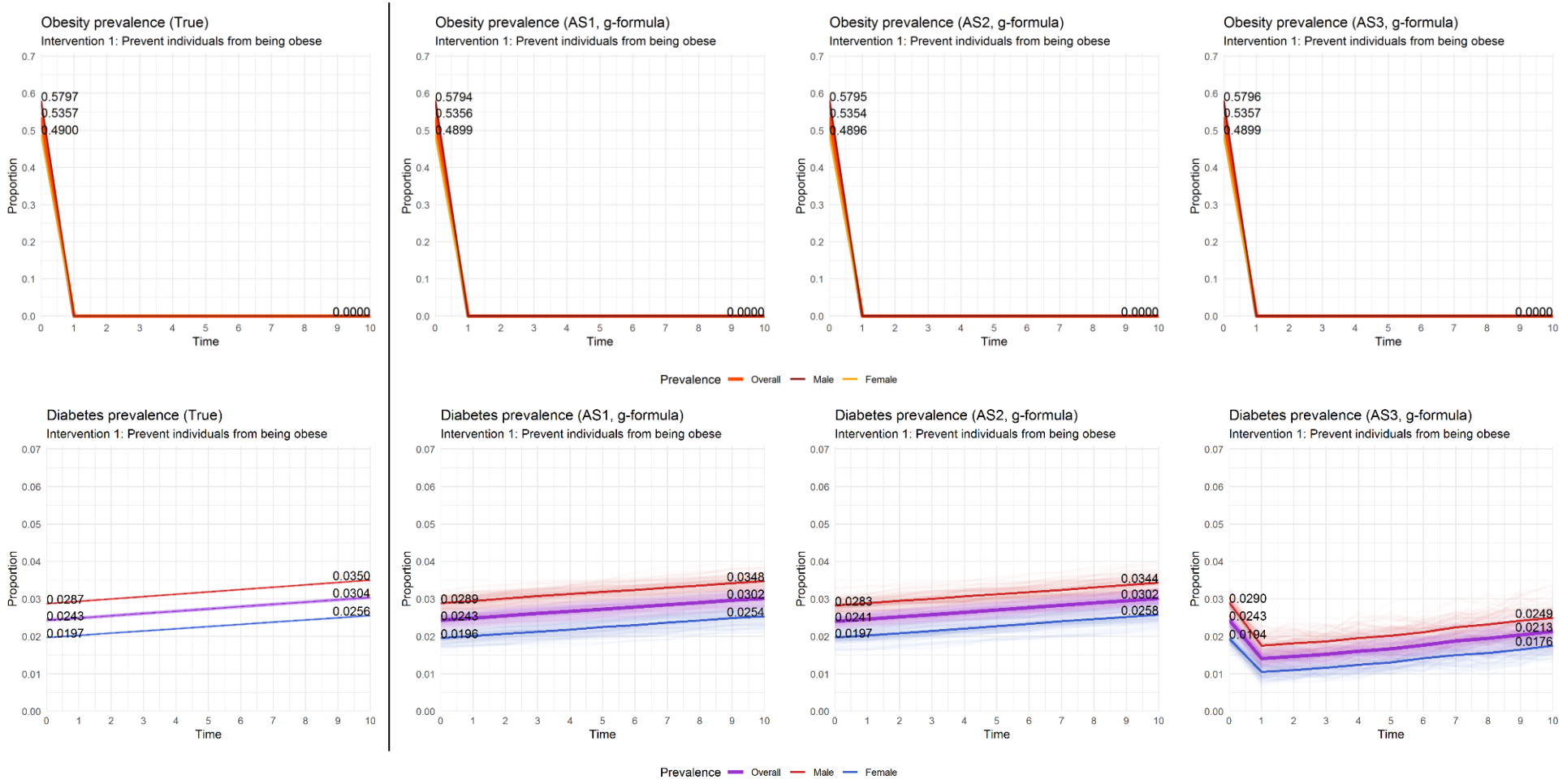


Figure 6.8 Counterfactual histories of obesity and diabetes prevalence under Intervention 1 for each of AS1 through AS3 modelled using the g-formula, compared to the true counterfactual history



6.4.2.2.2 Natural history

Plots which compare the true natural history of the population and those modelled by the g-formula according to AS1 through AS3 are shown in Figure 6.6 (obesity and diabetes prevalence) and Figure 6.7 (cross-sectional prevalence of sex, obesity, and diabetes).

All three autocorrelation structures appeared to reasonably reflect the true overall and disaggregated (by sex) prevalence of obesity and diabetes, as well as the proportion of individuals with each combination of characteristics at each time point under the natural history.

6.4.2.2.3 Counterfactual histories under hypothetical interventions

Although each autocorrelation structure was able to produce a population whose aggregated characteristics were consistent with those of the true population under its *natural history*, the same cannot be said for the *counterfactual histories*. Across all interventions, the highly mis-specified autocorrelation structure (i.e. AS3) performed poorly, thus explaining its poor performance relative to the more correctly-specified autocorrelation structures (i.e. AS1 and AS2) at estimating the causal effects of each intervention (from Table 6.2).

As an example, we consider how well each autocorrelation structure modelled the effects of Intervention 1. Each of AS1, AS2, and AS3 accurately modelled the effects of Intervention 1 on *obesity* prevalence (Figure 6.8), because the probability of obesity is (counterfactually) reduced to zero for all individuals regardless of the true probability of obesity. However, all autocorrelation structures were not equally good at modelling the effects of Intervention 1 on *diabetes* prevalence. For example, AS3 dramatically underestimated diabetes prevalence; this is because according to AS3, diabetes status at each time point was dependent upon only sex and obesity status (and not on previous diabetes status); thus, there were very few prevalent cases and the dramatic reduction (to zero) in the probability of obesity had the effect of substantially reducing incident cases.

The results for all interventions are presented in Appendix C (§C.2.2.1).

6.4.2.3 Microsimulation

This subsection describes the results of using microsimulation to estimate the true natural and counterfactual histories of the population, according to the process described previously (§6.4.2).

6.4.2.3.1 Estimated causal effects of interventions

The estimated causal effect on diabetes prevalence of each intervention is given in Table 6.3, for each of AS1 through AS3, as modelled using microsimulation; the true effects in the population (from Table 6.1) are also given for comparison.

The correctly-specified autocorrelation structure (i.e. AS1) performed well in estimating the true effects of Interventions 1 through 6, while the highly mis-specified autocorrelation structure (i.e. AS3) performed poorly, as expected. However, the slightly mis-specified autocorrelation structure (i.e. AS2) appeared to perform as well as AS1 in estimating the intervention effects of interest. In the following subsections, we briefly discuss these results further by examining how well each autocorrelation structure replicated the true natural and counterfactual histories.

Table 6.3 Table describing the estimated causal effect of each intervention on diabetes prevalence, for each of the autocorrelation structures modelled using microsimulation, compared to the true effect in the population

Effect	True	AS1	AS2	AS3
Effect of Intervention 1 *(prevent individuals from being obese)*	**-0.0112** ***(-26.9%)***	-0.0110 *(-26.6%)*	-0.0118 *(-28.3%)*	-0.0121 *(-46.2%)*
Effect of Intervention 2 *(make all individuals obese)*	**0.0080** ***(19.3%)***	0.0077 *(18.5%)*	0.0080 *(19.3%)*	0.0097 *(37.3%)*
Effect of Intervention 3 *(prevent anyone from becoming obese)*	**-0.0034** ***(-8.2%)***	-0.0032 *(-7.7%)*	-0.0037 *(-9.0%)*	-0.0121 *(-46.3%)*
Effect of Intervention 4 *(reduce probability of becoming obese by 15%)*	**-0.0004** ***(-0.9%)***	-0.0005 *(-1.1%)*	-0.0004 *(-1.0%)*	-0.0019 *(-7.3%)*
Effect of Intervention 5 *(reduce probability of remaining obese by 10%)*	**-0.0031** ***(-7.4%)***	-0.0031 *(-7.4%)*	-0.0032 *(-7.8%)*	-0.0014 *(-5.2%)*
Effect of Intervention 6 *(reduce probability of becoming obese by 15% and remaining obese by 10%)*	**-0.0035** ***(-8.3%)***	-0.0035 *(-8.5%)*	-0.0038 *(-9.2%)*	-0.0030 *(-11.4%)*
Total causal effect (TCE)	**0.0192** ***(63.2%)***	0.0187 *(61.5%)*	0.0198 (*66.5%)*	0.0218 *(155.4%)*

The estimated causal effect of each intervention (1 through 6) on diabetes prevalence was calculated by subtracting the average observed diabetes prevalence at time 10 under the natural history from the average diabetes prevalence at time 10 that was observed when the given intervention was applied to a random sample of 20,000 individuals. The TCE was calculated by subtracting the average diabetes prevalence at time 10 that was observed when Intervention 1 was applied from that which was observed when Intervention 2 was applied. All effects are additionally expressed as percentage changes.

Figure 6.9 Natural history of obesity and diabetes prevalence for each of AS1 through AS3 modelled using microsimulation, compared to the true natural history

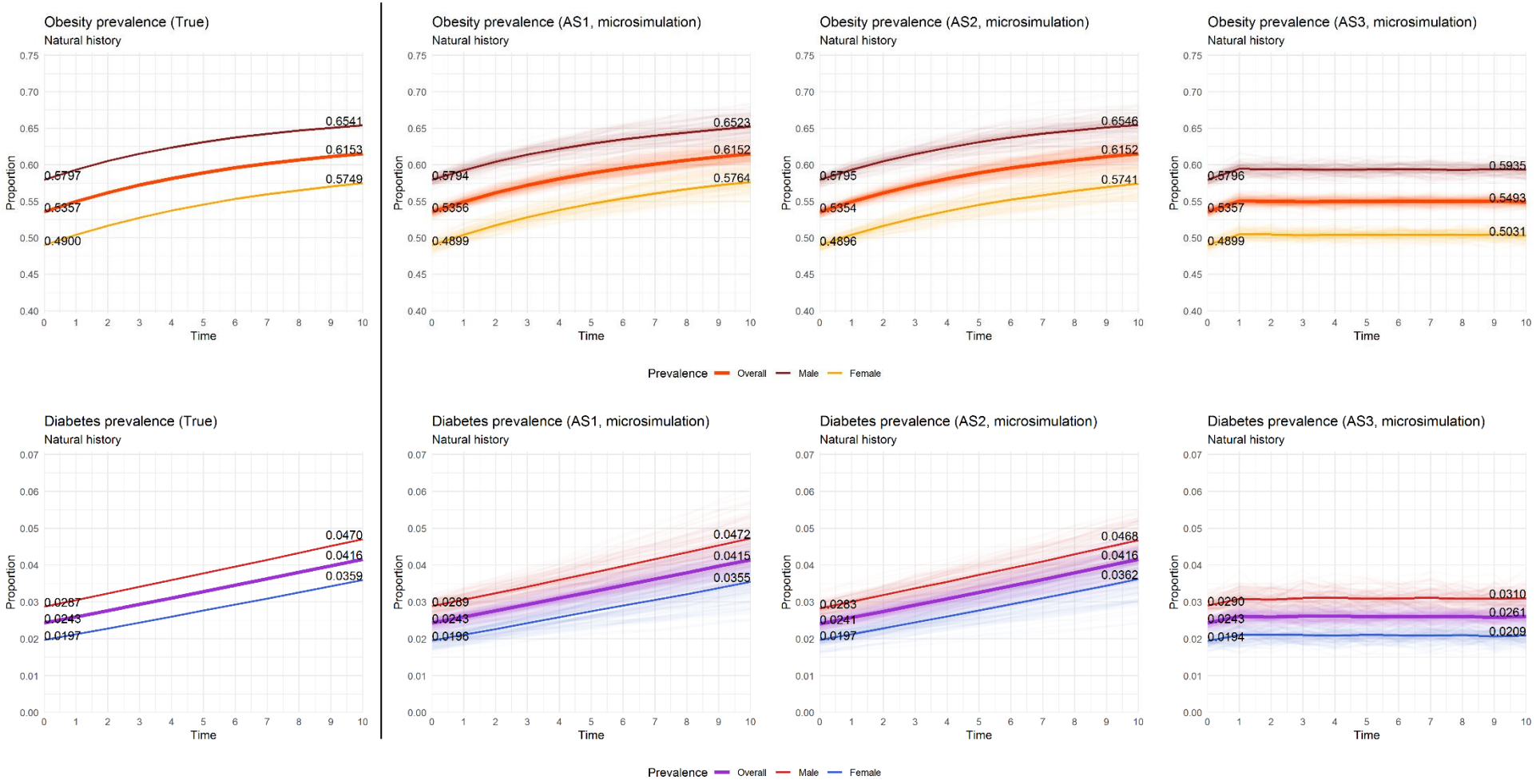


Figure 6.10 Natural history of the cross-sectional prevalence of sex, obesity, and diabetes, for each of AS1 through AS3 modelled using microsimulation, compared to the true natural history

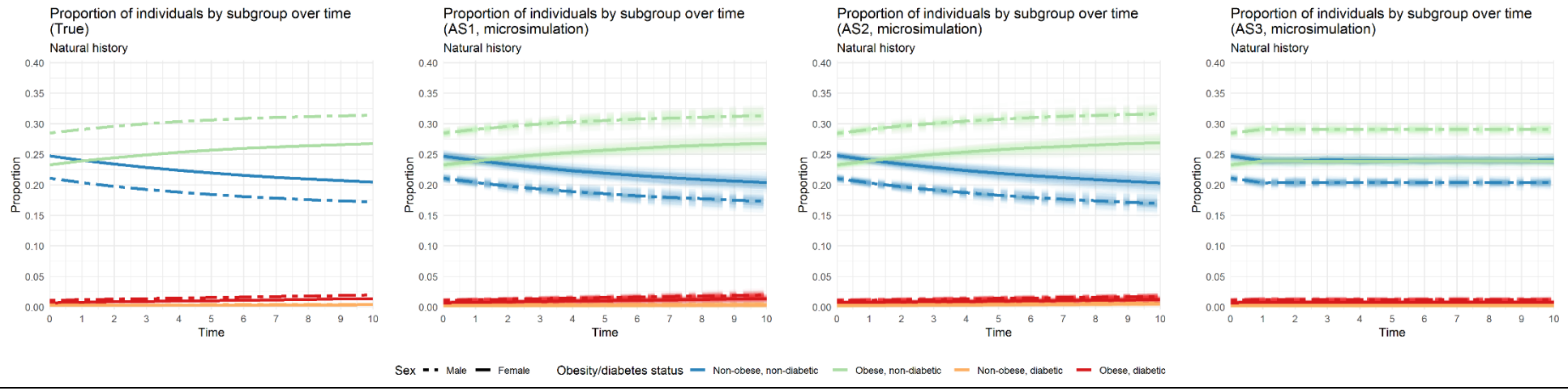
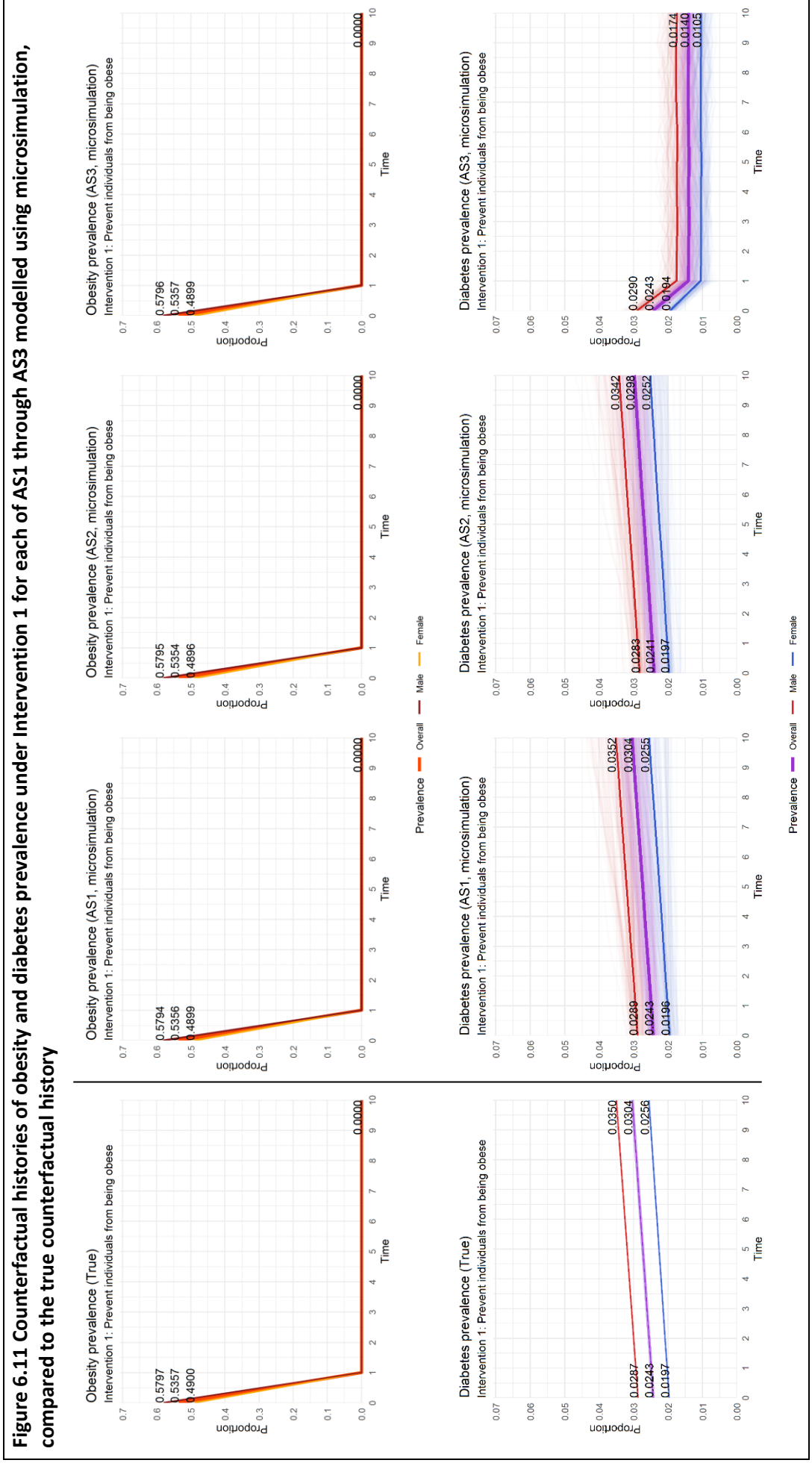


Figure 6.11 Counterfactual histories of obesity and diabetes prevalence under Intervention 1 for each of AS1 through AS3 modelled using microsimulation, compared to the true counterfactual history



6.4.2.3.2 Natural history

Plots which compare the true natural history of the population and those modelled by microsimulation according to AS1 through AS3 are shown in Figure 6.9 (obesity and diabetes prevalence) and Figure 6.10 (cross-sectional prevalence of sex, obesity, and diabetes).

As is evident, both the correctly specified and slightly mis-specified autocorrelation structures (i.e. AS1 and AS2, respectively) produced populations whose characteristics at every time point were broadly consistent with those of the true natural history. However, the highly mis-specified autocorrelation structure (i.e. AS3), which was unable to model the increasing conditional probabilities of obesity and diabetes over time, produced a population which diverged substantially from that of the true population under the natural history.

6.4.2.3.3 Counterfactual histories under hypothetical interventions

As with the natural history, modelling the correctly-specified autocorrelation structure (i.e. AS1) and the slightly mis-specified autocorrelation structure (i.e. AS2) appeared able to accurately replicate the true counterfactual histories under Interventions 1 through 6. However, across all interventions, modelling the highly mis-specified autocorrelation structure (i.e. AS3) using microsimulation performed poorly at estimating the counterfactual histories of obesity and/or diabetes, which resulted in poor estimates of the true causal effects of each intervention (from Table 6.2).

We again consider how well each autocorrelation structure models the effects of Intervention 1 as an example (depicted in Figure 6.11). Each of AS1, AS2, and AS3 accurately modelled the effects of Intervention 1 on *obesity* prevalence, because the probability of obesity was (counterfactually) reduced to zero for all individuals regardless of the true probability of obesity.[26] AS3 nevertheless dramatically underestimated diabetes prevalence, since the dramatic reduction (to zero) in the probability of obesity substantially reduced incident cases of diabetes and the autocorrelation structure implies very few prevalent cases.

The results for all interventions are presented in Appendix C (§C.2.2.2).

6.4.3 Discussion of findings

Both the g-formula and microsimulation performed equally well at replicating the true natural history and the true counterfactual histories of the population under various interventions when they correctly modelled the true data-generating process of the population (i.e. AS1), as expected. Moreover, both methods performed poorly when the data-generating process of the population was highly mis-specified (i.e. AS3); even so, the g-formula performed better at

[26] We note that for interventions which reduce the probability of obesity by a factor which is itself dependent on the true probability of obesity (i.e. Interventions 3 through 6), AS3 is not able to accurately model their effects on obesity prevalence. This is discussed further in Appendix C, where the full results of these simulations are presented (§C.2.2.2.1).

estimating the true natural and counterfactual histories, relative to microsimulation, since it utilised data from the true population at all time points.

However, using both methods, the slightly mis-specified autocorrelation structure (i.e. AS2) appeared to perform as well as the correctly specified autocorrelation structure (i.e. AS1) at modelling the natural history of the population, as well as the counterfactual histories of the population under Interventions 1 through 6, suggesting that in our example scenario both methods are relatively robust to minor mis-specifications of the data-generating process.

6.4.4 Sensitivity analyses

While the results of our simulation were broadly consistent with expectations, the slightly mis-specified data-generating process (i.e. AS2) appeared to perform as well as the correctly specified data-generating process (i.e. AS1) at replicating the true natural and counterfactual histories when using both the g-formula and microsimulation, despite the fact that AS2 did not fully model the time-dependent confounding that was present in the true data-generating process.

We speculated that this might be due to two factors specific to the example scenario chosen: (1) the low overall incidence and prevalence of diabetes in the population, or (2) the low degree of time-dependent confounding in the population (i.e. the small effect of obesity status on diabetes incidence, which was not modelled under AS2).

Therefore we performed the following five sensitivity analyses:

Sensitivity analysis 1:	Increase baseline diabetes prevalence.
Sensitivity analysis 2:	Increase diabetes incidence.
Sensitivity analysis 3:	Increase the effect of previous diabetes on obesity incidence/prevalence (i.e. increase the magnitude of time-dependent confounding).
Sensitivity analysis 4:	Both (1) and (3).
Sensitivity analysis 5:	Both (2) and (3).

6.4.4.1 Method

For each sensitivity analysis, the natural history of the initial population was simulated according to the same process described in Section 6.4.1.1, but with different simulation parameters. The counterfactual history under each of the six hypothetical interventions on obesity was simulated according to the same process described in Section 6.4.1.2, and the true causal effect of each intervention was calculated as previously (§6.4.1.3). We then used the g-formula and microsimulation to estimate the causal effects on diabetes prevalence of each intervention, according to the same processes described in Section 6.4.2.

Appendix C contains the simulation parameters for each sensitivity analysis and a fuller description of the characteristics of the simulated populations (§C.2.3).

6.4.4.2 Results

In sensitivity analysis 5, we saw the largest divergence between the results of the g-formula and microsimulation when modelling autocorrelation structures which differed from the true (i.e. AS2 and AS3); we also saw the largest divergence between the results of the true (i.e. AS1) and the slightly mis-specified data-generating processes (i.e. AS2), though this was still relatively modest. Table 6.4 describes the effect of each intervention on diabetes prevalence in sensitivity analysis 5, as estimated by the g-formula and microsimulation; the true effects in the population are given for comparison.

In Table 6.4, we can see that when the data-generating process is correctly specified (i.e. AS1), both the g-formula and microsimulation perform equally well at estimating the true intervention effects, as in the original simulations. However, when the data-generating process is slightly mis-specified (i.e. AS2), both the g-formula and microsimulation perform poorly compared to when they are correctly specified, though the magnitude of divergence is still relatively modest. Furthermore, the g-formula performs better than microsimulation for estimating the effects of the most dramatic interventions (i.e. Interventions 1 through 3) but there is little difference with respect to more modest interventions (i.e. Interventions 4 through 6). When the data-generating process is highly mis-specified (i.e. AS3), both the g-formula and microsimulation perform poorly at estimating the true intervention effects, as in the original simulations.

Results from all sensitivity analyses are provided in Appendix C (§C.2.3.2).

Table 6.4 Table describing the estimated causal effect of each intervention on diabetes prevalence for each of AS1 through AS3 modelled using the g-formula and microsimulation, compared to the true effect in the population (Sensitivity analysis 5)

		AS1		AS2		AS3	
Effect	**True**	**g-formula**	**MSM**	**g-formula**	**MSM**	**g-formula**	**MSM**
Effect of Intervention 1 *(prevent individuals from being obese)*	**-0.0432** ***(-47.2%)***	-0.0434 *(-47.3%)*	-0.0434 *(-47.3%)*	-0.0443 *(-47.9%)*	-0.0450 *(-48.9%)*	-0.0829 *(-90.7%)*	-0.0217 *(-69.8%)*
Effect of Intervention 2 *(make all individuals obese)*	**0.0304** ***(33.1%)***	0.0309 *(33.7%)*	0.0308 *(33.6%)*	0.0295 *(31.9%)*	0.0302 *(32.8%)*	0.0482 *(52.8%)*	0.0172 *(55.4%)*
Effect of Intervention 3 *(prevent anyone from becoming obese)*	**-0.0134** ***(-14.6%)***	-0.0129 *(-14.1%)*	-0.0130 *(-14.2%)*	-0.0138 *(-14.9%)*	-0.0142 *(-15.4%)*	-0.0829 *(-90.7%)*	-0.0217 *(-69.9%)*
Effect of Intervention 4 *(reduce probability of becoming obese by 15%)*	**-0.0017** ***(-1.8%)***	-0.0015 *(-1.7%)*	-0.0016 *(-1.8%)*	-0.0013 *(-1.4%)*	-0.0013 *(-1.4%)*	-0.0123 *(-13.5%)*	-0.0034 *(-10.9%)*
Effect of Intervention 5 *(reduce probability of remaining obese by 10%)*	**-0.0118** ***(-12.9%)***	-0.0116 *(-12.6%)*	-0.0116 *(-12.6%)*	-0.0119 *(-12.9%)*	-0.0118 *(-12.8%)*	-0.0081 *(-8.9%)*	-0.0023 *(-7.3%)*
Effect of Intervention 6 *(reduce probability of becoming obese by 15% and remaining obese by 10%)*	**-0.0134** ***(-14.6%)***	-0.0133 *(-14.5%)*	-0.0133 *(-14.5%)*	-0.0137 *(-14.8%)*	-0.0137 *(-14.8%)*	-0.0191 *(-20.9%)*	-0.0051 *(-16.4%)*
Total causal effect (TCE)	**0.0736** ***(151.9%)***	0.0743 *(153.8%)*	0.0743 *(153.5%)*	0.0738 *(153.1%)*	0.0752 *(159.6%)*	0.1311 *(1539.9%)*	0.0389 *(415.2%)*

The estimated causal effect of each intervention (1 through 6) on diabetes prevalence was calculated by subtracting the average observed diabetes prevalence at time 10 under the natural history from the average diabetes prevalence at time 10 that was observed when the given intervention was applied to a random sample of 20,000 individuals. The TCE was calculated by subtracting the average diabetes prevalence at time 10 that was observed when Intervention 1 was applied from that which was observed when Intervention 2 was applied. All effects are additionally expressed as percentage changes.

6.5 Discussion

For our example scenario, our simulations broadly aligned with our expectations. That is, both the g-formula and microsimulation faithfully replicated the true natural and counterfactual histories of the population when they correctly modelled the data-generating process of the population. Our results also suggested that small mis-specifications in this context don't make substantial differences for either the g-formula or microsimulation, but that more serious mis-specifications were more likely to negatively impact MSMs. It can be interpreted with cautious optimism that the most accurate results were produced by the most plausible hypothesised autocorrelation structures (i.e. AS1 and AS2). However, our simulations were deliberately simplified and thus the magnitude of any biases should not be assumed to be transferrable to other contexts.

Our sensitivity analyses produced a larger divergence between the correctly specified and slightly mis-specified autocorrelation structures, and provided evidence for the g-formula being more robust to small mis-specifications in the data-generating process. However, the magnitude of these differences was still relatively modest, suggesting they also may be the result of some other structural factor(s) present in the example context chosen. For instance, both obesity and diabetes were simulated to have a strong serial correlation, reflecting that they are conditions which are difficult to transition out of; the probability of becoming non-obese at any given time point ranged between 0.03 and 0.05 in the original simulation, whereas the probability of becoming non-diabetic was zero. Moreover, diabetes incidence was simulated to be very low in absolute terms, both in the original simulation and subsequent sensitivity analyses. Across other contexts, in which individuals can more easily transition in and out of different states, the differences might become more pronounced.

Despite the g-formula being potentially more robust to mis-specifications than microsimulation, it is worth keeping in mind that the utility of MSMs lies in their ability to produce estimates of a future population, which are inherently uncertain; thus, those employing MSMs may be more willing to sacrifice a certain degree accuracy and/or precision for the sake of utility. Nevertheless, where possible, researchers would benefit from modelling different plausible data-generation processes as sensitivity analyses.

6.5.1 Limitations and future work

Our simulations were deliberately simplified in several respects. First, we considered only three binary variables (i.e. sex, obesity, and diabetes), when in reality there are many others which are likely relevant to the causal processes of interest. This simplification also meant that the conditional probabilities of each variable could be nonparametrically estimated using both methods. Second, the true data-generating process (as depicted in Figure 6.1) had only first-order autocorrelation, i.e. where variables at one time point did not affect any future variables except for those in the immediately subsequent time point. For example, obesity status at time

t was dependent only on variables at time $t-1$ and not on any variables at time $t-2$. Third, as simulated, the true probabilities governing transitions in and out of obesity were the same for every time point, which could be interpreted as representing no change in the underlying obesogenic environment. We did not consider a situation involving transition problems which changed over time. We suspect that had the true data-generating process been more complex, and had the true transition probabilities been simulated to change over time, mis-specifications in the hypothesised data-generating process when using the g-formula and microsimulation would have been more consequential.

Other limitations include that we did not consider interventions which varied over the course of the simulation, and that one of the hypothesised autocorrelation structures considered (i.e. AS3, in Figure 6.5) was so simple that it is unlikely to be encountered in practical applications. Nevertheless, it was chosen as part of a broad range of possible data-generating processes. We also did not consider the added complexity of parameterising our MSMs with estimates from different datasets, as this was not the primary focus of this particular research.[27] Future simulations are warranted to explore these issues, with the current simulation providing a foundation for doing so.

6.6 Summary

Microsimulation provides a promising method for estimating causal effects in a longitudinal setting via the simulation counterfactual scenarios. This chapter demonstrates the utility of DAGs for understanding how specification of data-generating processes impacts on estimation of both natural and counterfactual histories. DAGs are also demonstrated to be an invaluable tool for clearly explicating the assumptions made about the causal structure of an MSM, thereby aiding interpretability and reproducibility. The simulations presented in this chapter provide a framework for evaluating individual-based simulation methods intended for causal inference, and inform how the robustness and reliability of such methods may be improved by accurately capturing data-generating processes.

[27] Murray, E.J. et al. (99) have begun to explore this issue, and provide a useful starting point for considering some of the potential issues arising from the combination of parameter estimates which have come from populations which differ in their distribution of unmeasured confounders.

Chapter 7
Conclusion

7.1 Introduction

This thesis set out to explore how counterfactual thinking, encoded in the language of DAGs, could be integrated into established methods for longitudinal data analysis by considering three specific methods – the analysis of change, regression with 'unexplained residuals', and microsimulation modelling. Each of these methods is typically applied in a distinct longitudinal context, and DAGs have been demonstrated to be useful tools for thinking through causal processes and informing causal analyses in each. This highlights the utility and promise of DAGs for informing a wide variety of methods for longitudinal data in a robust causal framework – a task that has become increasingly necessary in the era of 'big data', where the familiar biases associated with observational data are likely amplified.

While longitudinal data are of great interest to epidemiologists and data scientists, they present additional challenges for causal inference. Even the seemingly simple idea of 'change' can pose difficulties, since changes are in fact captured by follow-up events conditional on baseline events rather than a conflated summary of the two, as demonstrated in Chapter 4. Indeed, this conceptualisation of change is that which is exploited by UR models and fundamental to why the method works, as demonstrated in Chapter 5. The importance of understanding and faithfully modelling data-generating processes in order to make robust causal inferences in longitudinal contexts has been demonstrated throughout this thesis.

Chapter 7 summarises and critically evaluates the findings of this thesis.

7.1.1 Chapter overview

A general chapter overview is provided below.

In Section 7.2, we summarise the key findings of the individual pieces of research contained in this thesis and discuss their implications.

In Section 7.3, we highlight the contributions made to the literature by this research, including details of related publications.

In Section 7.4, we discuss the limitations of this research, and outline potential areas for future research.

7.2 Summary of findings

Chapter 3 provided a foundation for understanding the contexts in which the three methods considered in this thesis might be used. For each of methods, DAGs were first used to depict the longitudinal context under consideration. The principles of graphical model theory were then applied so that robust conclusions could be drawn about how the method ought to be deployed in order to estimate causal effects. In each scenario, the application of DAGs was also

useful for identifying potential problems and/or biases that might arise if the method was deployed incorrectly or without proper consideration for causal structures.

A general summary of the findings of each chapter follows, each of which includes fuller details of how the objectives of this thesis have been fulfilled. Additionally, key messages for epidemiological and public health researchers are summarised in Table 7.1.

Table 7.1 Key messages for epidemiological and public health researchers

1. DAGs are useful tools for thinking through causal processes and informing causal analyses involving longitudinal data; where possible, they should be incorporated into standard practices for answering causal questions.
2. DAGs should be used to consider the causal structures and data-generating processes governing a given scenario; this enables researchers to identify appropriate covariate adjustment to target the most useful estimand, and to clearly communicate causal assumptions when presenting results.
3. The misapplication of methods for causal inference is likely to invoke *inferential bias*, i.e. where the numerical estimate obtained does not correspond to a sensible or interpretable causal quantity, thereby leading to incorrect interpretation and/or inference. This is very different to *statistical bias*.

7.2.1 Statistical versus individual-based simulation methods for causal inference

Chapter 3 considered several important methods for estimating causal effects in longitudinal data, including DAG-informed regression modelling, microsimulation, and agent-based modelling. The distinct historical evolutions of these three methods have given rise to distinct features of the methods themselves, which provide a foundation for critical comparison. Of note are the differing levels of emphasis they place on data versus theory, which in turn informs the types of causal questions which they are well-suited to answering, their relative focus on fixed versus random effects, and the timescales upon and timeframes in which they operate.

DAG-informed regression modelling is well-suited to analyses in which the query of interest can be explicated in the traditional language of ‘exposures’ and ‘outcomes’, for which sufficient individual-level data are available on a suitable timescale to capture the causal processes of interest, and for which spillover effects and interference are negligible. With regards to their practical utility for policy-making decisions, this type of modelling is appropriate for exposures and/or interventions whose effects may be safely assumed to be transportable across time. When such conditions are met, statistical DAG-informed approaches provide a robust method for causal inference whilst requiring relatively few assumptions, and they offer a transparent means for communicating those assumptions.

In contrast, agent-based modelling provide a means for modelling greater complexity (e.g. in the form of individual interactions and spillover effects) though they do so by relying on a greater number of assumptions. Moreover, ABMs inherently contain greater uncertainty about the validity of their causal effect estimates because they are typically applied in situations in which key variables may not be represented numerically, or in which observed data are not sufficiently granular in timescale to fully inform parameterisation and/or enable effective validation. In this context, microsimulation offers a useful halfway house, since they may be able to utilise the robust foundations of graphical causal models whilst also exploring the effects of complex (i.e. multiple) interventions that occur over long periods of time (which possibly extend well into the future).

7.2.2 The analysis of change

Chapter 4 considered the analysis of change using DAGs, a context which involves quantifying the relationship between a single exposure and subsequent 'change' in a longitudinal outcome. In doing so, we demonstrated that change scores do not in general represent exogenous change, and that the follow-up outcome should be the true target of any analysis of change. Moreover, we used path tracing to demonstrate why, and the degree to which, change-score analyses differ from follow-up adjusted for baseline analyses (i.e. ANCOVA) in non-randomised data.

However, follow-up adjusted for baseline analyses are not always the best solution for the analysis of change, since the estimand targeted by such analyses differs according to the causal structure of the data. Where the exposure is caused by the baseline outcome, a follow-up adjusted for baseline analysis targets the total causal effect; where the exposure causes the baseline outcome, this method of analysis targets the direct causal effect only. Thus, determining whether to adjust for the baseline outcome is context-dependent, and there may exist scenarios in which follow-up *unadjusted* for baseline analyses are more appropriate. This has not previously been considered by other authors examining the analysis of change because scenarios in which the exposure causes the baseline outcome are not generally encountered in experimental contexts; nevertheless, they arise frequently in observational contexts and therefore this research has the potential for substantively improving analyses of change.

Using a simple simulated example, the degree of inferential bias that might be introduced by a change-score analysis was illustrated. Furthermore, the importance of using DAGs to help determine the most useful analytical strategy was emphasised. This has applicability across other methods and contexts, where DAGs can be used to consider the causal structures involved and to identify appropriate adjustment to target the most useful estimand.

7.2.3 Regression with 'unexplained residuals'

Chapter 5 considered regression with 'unexplained residuals' using DAGs, a context which involves quantifying the relationship between *separate* measurements of a longitudinal exposure and a subsequent outcome. Using path tracing, we demonstrated why the method

(as originally formulated) is able to quantify the total causal effects of multiple measurements of an exposure on a subsequent outcome within a single model. We also demonstrated how UR models must be implemented in order to robustly accommodate confounding by both baseline and time-dependent covariates; this is an issue that has not previously been explored but is nevertheless crucial for practical applications of the method.

Despite their perceived advantages, UR models are significantly more complex to implement than standard regression methods, particularly in the presence of time-dependent confounding. UR models rely on the orthogonality of the constructed UR terms; where there exists time-dependent confounding, this necessitates the creation of UR terms for the confounder as well as the exposure. Moreover, bootstrapping is necessary to obtain robust estimates of standard errors (SEs) for all UR models, since SEs are artefactually reduced when using UR models; this was demonstrated through a simple simulated example.

Taking all results together, we were able to conclude that DAGs are useful for understanding the properties of UR models, and for determining correct adjustment for confounding. However, the additional complexity required to implement UR models makes them more vulnerable to analytical and interpretational problems, and thus they offer little to no benefit compared to standard regression models.

7.2.4 Microsimulation modelling

Chapter 6 considered microsimulation modelling using DAGs, a context which involves quantifying the relationship between *multiple* measurements of a longitudinal exposure and a subsequent outcome. By specifically contrasting microsimulation with the g-formula, we demonstrated some of the unique challenges faced by individual-based simulation approaches and the importance of faithfully modelling the data-generating processes in order to estimate causal effects. Using a simulated example, we demonstrated how varying degrees of mis-specification adversely impacted estimation of natural and counterfactual histories using microsimulation compared to the g-formula.

Using microsimulation, the future distributions of individuals' states must be generated by the data-generating processes specified in the model. Mis-specification of these processes therefore has more consequential effects on estimation of causal effects using microsimulation compared to the g-formula, though the magnitude of any biases depends on the true causal structure, the degree of mis-specification, and the true causal parameters. MSMs are also more likely to simplify data-generating processes, due to the difficulties involved in parameter specification. Few authors to date have considered microsimulation within a formal causal framework, despite widespread use of the method for making causal inferences; this research therefore provides a foundation for exploring and identifying additional causal considerations associated with this method.

Our simulations demonstrate the importance of using causal parameters in MSMs and of modelling plausible data-generating processes. DAGs are demonstrated to be a useful means

by which the causal assumptions of an MSM can be made explicit, which can additionally aid model interpretability and reproducibility.

7.3 Contributions to the literature

Several pieces of work related to this thesis have already been published or accepted for publication.

Chapter 3 contains a critical comparison of statistical versus individual-based simulation methods for causal inference (§3.5), which was published in its entirety in the *International Journal of Epidemiology* (3). This manuscript arose due to a lack of clarity in the literature surrounding the distinction between microsimulation and agent-based modelling. Moreover, because DAG-informed regression modelling and individual-based simulation modelling have largely been confined to separate research disciplines, there has existed little overlap in the knowledge about them and skills necessary for implementing them; the published manuscript fills this gap. Chapter 3 also contains three examples which illustrate the benefits of applying DAGs in new contexts; two of the three examples arise directly from research conducted by the author of this thesis. Section 3.3.1.2 discusses the use of DAGs to understand the analysis of compositional data for causal inference, and Section 3.3.2 discusses the distinction between models for prediction and models for causal inference. Both pieces of research have been published in the *International Journal of Epidemiology* (1, 2).

The work featured in Chapter 4 has been submitted to the *International Journal of Epidemiology*; it is currently being revised following a second round of peer review and is available as pre-print on *ArXiv* (4). This manuscript arose due to historical confusion in the literature surrounding the use of 'change scores' and their suitability for causal analysis. It is hoped that the published manuscript will demonstrate the problems with change-score analyses, and will inform alternate analytical strategies by encouraging researchers to think about the plausible causal structure that may arise in specific scenarios involved in the analysis of change.

The work featured in Chapter 5 has been published in *Statistical Methods for Medical Research* (5). This manuscript arose due to a lack of clarity in the literature concerning how to robustly extend UR models to accommodate both longitudinal exposures measured at more than two time points as well as confounding variables. The published manuscript therefore demonstrates how to implement UR models correctly in more complex longitudinal scenarios (including the requirement that standard errors be estimated via bootstrapping), which is useful for future researchers seeking to use this method.

It is anticipated that the work featured in Chapter 6 will form the basis of one to two manuscripts, and that the simulations developed here will provide a foundation for postdoctoral research to be conducted by the author of this thesis.

The author of this thesis has also contributed to a manuscript which examines the use of DAGs in applied health research and offers recommendations for improving their transparency and utility in future research. This manuscript has been submitted to the *International Journal of Epidemiology* and is currently available as a pre-print on *medRxiv* (57).

7.4 Limitations and future work

Across all contexts considered in this thesis, we did not consider additional issues which might threaten estimation of unbiased causal effects, such as measurement error or incorrect model specification. All simulated examples were purposely oversimplified to highlight the key issues involved from a causal perspective, though the work contained in this thesis provides a foundation for which to explore additional issues. Several of the most promising avenues for future research are outlined in the following subsections.

7.4.1 Understanding regression to the mean (RTM) using DAGs

The issues surrounding the use of change scores which were discussed in Chapter 4 have historically been bound up in discussions surround measurement error and regression to the mean (RTM) (158, 193). This is because in observational data, the results of change-score analyses are expected to agree with those obtained from follow-up adjusted for baseline analyses only under the specific condition in which the baseline and follow-up outcome are perfectly correlated (as demonstrated in §4.4). However, RTM is fundamentally distinct from the estimand confusion problem – namely, that the estimand targeted by change-score analyses is neither meaningful nor useful for causal analyses. Previous research focusing on the challenges of measurement error and biological variation abound, but these have often been confused by the misunderstanding around what constitutes change. It is anticipated that using DAGs to consider RTM and associated phenomena of regression dilution and Bayesian shrinkage can cast light on many of these issues. Additionally, the practice of considering deterministic relationships within DAGs – including the explicit depiction of error terms (as in Figure 4.1) – is likely to be useful for understanding phenomena associated with patterns of *variation* rather than mean structures.

7.4.2 Generalisability, transportability, and MSMs

There exists an emerging literature related to the issue effect generalisability and transportability, both of which represent threats to the external validity of a model (194-196). Generalisability relates to the issue of making inferences from a possibly biased sample of a target population back to the full target population (which includes the study sample); in contrast, transportability relates to the issue of making inferences for a target population where the study sample and target population are at least partially non-overlapping (194). Both of these issues are of particular relevance to the field of microsimulation, in which parameters are often combined from multiple datasets (and thus likely from multiple populations) into one single model in order to estimate causal effects; this is related to the issues discussed in Murray, E.J. et al. (99).

In the simulations presented in Chapter 6, we did not consider the added complexity of parameterising our MSM using estimates from different populations or datasets. However, this likely represents a fruitful area for future research, in which recent advances in graphical model theory may be applied in microsimulation contexts.

7.4.3 Integrating DAGs with ABMs

All methods considered in this thesis are typically applied in contexts in which there exists no interference, i.e. in which each individual's outcome is affected only by his/her exposure and not that of any other individual. However, there exists an emerging literature relating to causal inference in the presence of interference (197-200). A notable example is that of Ogburn, E.L. and T.J. VanderWeele (197), who distinguish amongst three causal mechanisms which give rise to interference using DAGs and provide criteria for the identification of causal effects in these scenarios.[28] These developments have obvious applications for vaccination programmes and social network data, but likely also have more general applications to ABMs. Moreover, ABMs provide a useful framework for simulating interference, since ABMs are defined by interactions between agents.

7.5 Summary

Longitudinal data constitute a large proportion of the new and emerging forms of data in the era of 'big data'. In this context, robust methods for identifying and estimating causal effects are necessary to inform clinical and public health interventions. However, longitudinal data present both theoretical and analytical challenges for causal inference, over and above those presented by cross-sectional data. The counterfactual framework provides a valuable paradigm for conceptualising and identifying causal effects in longitudinal data. This thesis has illustrated the utility of using DAGs to think through causal processes and inform causal analyses across a wide variety of longitudinal scenarios and statistical- and simulation-based methods. It is anticipated that the application of DAGs to other methods (e.g. ABMs) provide fruitful areas for research, and the analyses conducted in this thesis provide a basis and a starting point for doing so.

[28] These three mechanisms which give rise to interference are: (1) *direct interference*, in which one individual's exposure directly affects another individual's outcome; (2) *interference by contagion*, in which one individual's outcome affects the outcome of other individuals; and (3) *allocational interference*, in which an individual is allocated to a group and his outcome is affected by which individuals are allocated to the same group.

Appendix A
The analysis of change

A.1 Introduction

This appendix contains additional material relating to Chapter 4. In particular, it contains additional details relating to the simulated example described in Section 4.6.1.

A.2 Simulated example

A.2.1 DAGs

In this subsection, we depict the DAGs from which data were simulated in both the original simulation and the additional simulation. The 'original simulation' refers to the simulation whose results are presented in Section 4.6.1.2. The 'additional simulation' refers to the simulation in which additional baseline confounding was included, which was referred to in Section 4.6.1.3.

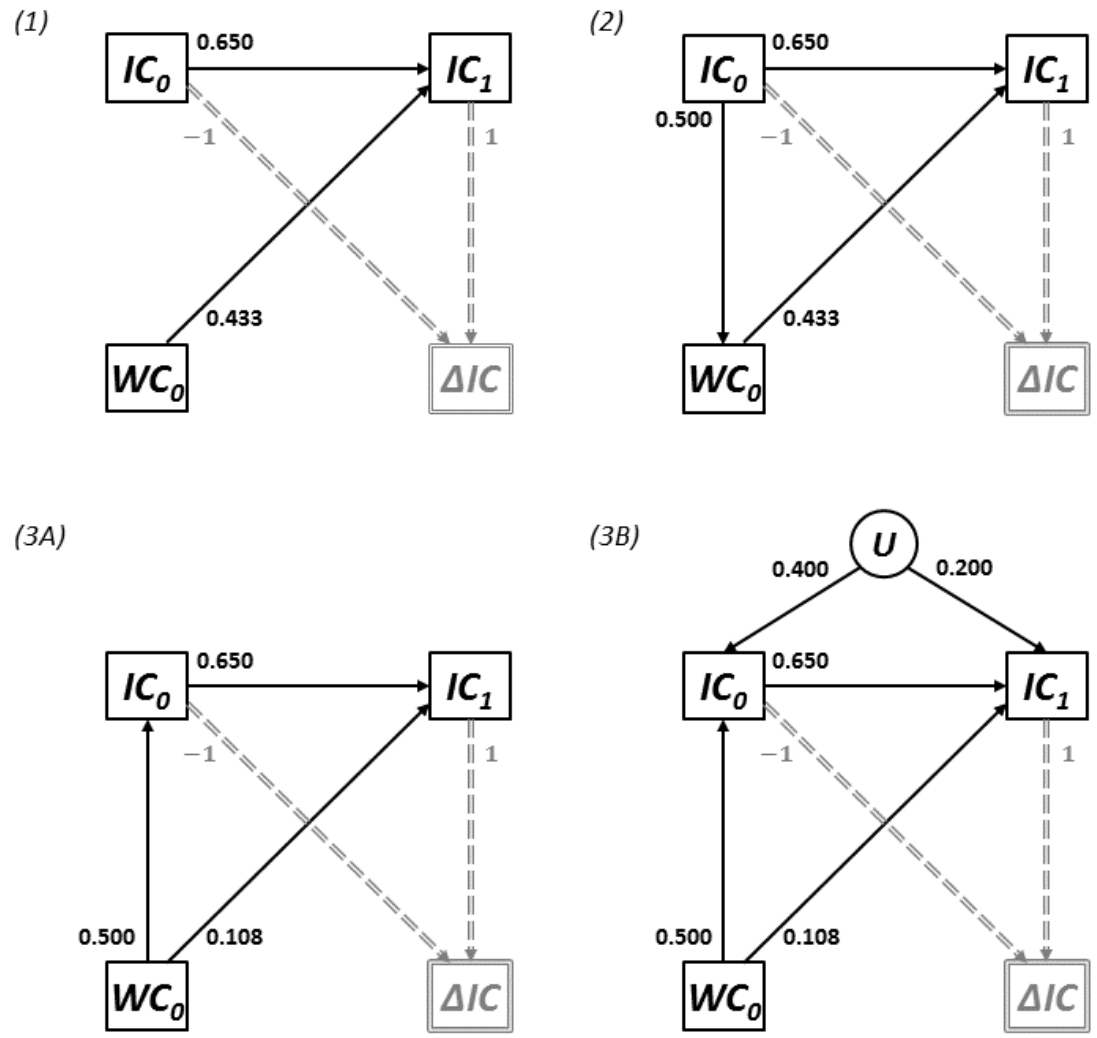
A.2.1.1 Original simulation

Figure A.1 depicts the DAGs from which multivariate normal data were simulated in order to demonstrate the degree of inferential bias that might be introduced by a change-score analysis.

Insulin concentration (IC) appears log-normally distributed (160), and so was simulated and analysed in its log-transformed form. For each of the scenarios depicted in Figure A.1, 10,000 non-overlapping random samples of 1,000 observations from a multivariate normal distribution were simulated, using the 'dagitty' package (v. 0.2-2)(46, 47) in R (v. 3.3.2)(161). Standardised path coefficients were selected to approximately match observed cross-sectional patterns. The path coefficient between WC_0 and IC_1 was simulated as 0.500 where applicable (i.e. Scenarios 2, 3A, and 3B); the path coefficient between IC_0 and IC_1 was simulated as 0.650, to represent a strong but imperfect correlation over time. In Scenario 3B, an unobserved variable U was simulated to introduce confounding a confounding correlation averaging 0.08.

The total causal effect of WC_0 on IC_1 was fixed at 0.200 Log(mmol/L)/dm; when mediated through IC_0, this was partitioned into an indirect causal effect of 0.150 Log(mmol/L)/dm and a direct causal effect of 0.050 Log(mmol/L)/dm.

Figure A.1 DAGs from which multivariate normal data were simulated to demonstrate the degree of inferential bias that might be introduced by a change-score analysis

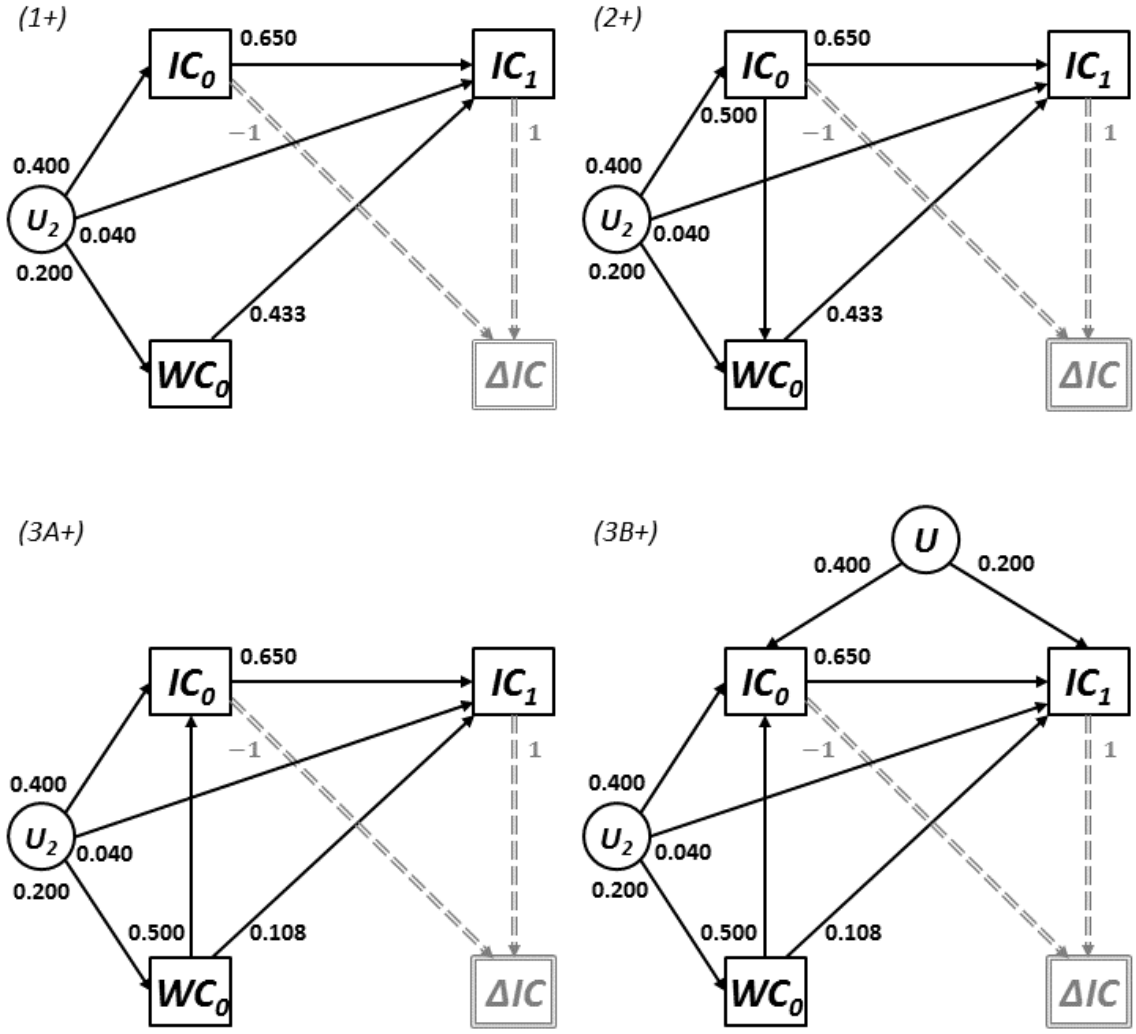


In Scenario (1), IC_0 is a competing exposure for the effect of WC_0 on IC_1. In Scenario (2), IC_0 is a confounder for the effect of WC_0 on IC_1. In Scenario (3), IC_0 is a mediator for the effect of WC_0 on IC_1; U represents an unobserved or unmeasured variable that confounds the relationship between IC_0 and IC_1 (i.e. a mediator-outcome confounder). Numbers represent standardised path coefficients. Deterministic relationships are indicated by double-lined arrows, and fully determined nodes are indicated by double-outlined rectangles.

A.2.1.2 Additional simulation with unmeasured baseline confounder U_2

Because our original simulation was deliberately simplified, we also considered the four causal scenarios depicted in Figure A.2, in which an unmeasured baseline confounder U_2 affecting each of WC_0, IC_0, and IC_1.

Figure A.2 DAGs from Figure A.1, with an additional unmeasured baseline confounder U_2



In Scenario (1), IC_0 is a competing exposure for the effect of WC_0 on IC_1. In Scenario (2), IC_0 is a confounder for the effect of WC_0 on IC_1. In Scenario (3), IC_0 is a mediator for the effect of WC_0 on IC_1; U represents an unobserved or unmeasured variable that confounds the relationship between IC_0 and IC_1 (i.e. a mediator-outcome confounder). Numbers represent standardised path coefficients. Deterministic relationships are indicated by double-lined arrows, and fully determined nodes are indicated by double-outlined rectangles.

Path coefficients for U_2 were chosen which induced a confounded correlation of approximately 0.08 between WC_0 and both IC_0 and IC_1. All other details from the original simulation (§A.2.1.1) remained unchanged.

Results of this additional simulation are presented in Section A.2.3.

A.2.2 Simulation parameters

The simulated mean and standard deviation (SD) specified in the simulation are provided in Table A.1.

Table A.1 Mean (SD) of waist circumference and insulin concentration, as reported in three separate waves of NHANES data and as simulated

	NHANES			Simulated
	2009-2010	**2011-2012**	**2013-2014**	
Waist circumference (dm)	9.50 (1.58)	9.42 (1.61)	9.52 (1.65)	9.50 (1.60)
Insulin concentration (Log(mmol/L))	4.20 (0.70)	4.08 (0.74)	3.98 (0.77)	4.00 (0.74) - *baseline* 4.20 (0.74) - *follow-up*
Pearson correlation	0.58[a]	0.58[a]	0.60[a]	0.50 – 0.60

[a]*Between waist circumference and log insulin concentration*

The mean values for insulin concentration (IC) were simulated to represent a notional five percent increase between baseline and follow-up.

A.2.3 Results of additional simulation with unmeasured baseline confounder U_2

The results of the additional simulation with unmeasured baseline confounder U_2 are summarised in Table A.2. As expected, all three methods provided biased estimates of the total causal effect of WC_0 on IC_1. However, a follow-up adjusted for baseline analysis appeared to be the least biased for Scenarios 1, 2, and 3A, whereas a follow-up unadjusted for baseline analysis was preferred for Scenario 3B. The change-score analysis performed poorly across all scenarios.

Table A.2 Median regression coefficient of WC_0 (and 95% simulation limits) for each method of analysis, for each causal scenario depicted in Figure A.2

	IC_0 is:	Competing exposure	Confounder	Mediator	
Method of analysis: ↓	**Scenario:**	**1+**	**2+**	**3A+**	**3B+**
Change-score		0.191	0.114	-0.040	-0.040
($\widehat{\Delta IC} = \hat{\alpha}_0 + \hat{\boldsymbol{\alpha}}_{\mathbf{1}} WC_0$)		(0.187, 0.220)	(0.104, 0.123)	(-0.061, -0.019)	(-0.058, -0.023)
Follow-up adjusted for baseline		0.203	0.205	0.048	0.015
($\widehat{IC_1} = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}_{\mathbf{1}} WC_0 + \hat{\beta}_1 IC_0$)		(0.187, 0.220)	(0.199, 0.211)	(0.023, 0.071)	(-0.006, 0.035)
Follow-up unadjusted for baseline		0.228	0.382	0.228	0.228
($\widehat{IC_1} = \hat{\gamma}_0 + \hat{\boldsymbol{\gamma}}_{\mathbf{1}} WC_0$)		(0.203, 0.253)	(0.366, 0.398)	(0.203, 0.252)	(0.203, 0.253)

The true total effect was simulated to be 0.200; where this was when mediated through IC_0 (i.e. Scenarios 3A and 3B), the true direct effect was simulated to be 0.050.

A.2.4 Annotated R code

```
################################################################################
## CHANGE SCORES: SIMULATED EXAMPLE ############################################
################################################################################

# This code demonstrates the degree of inferential bias that might be
# introduced by a change-score analysis (compared to a follow-up adjusted
# for baseline analysis & a follow-up UNadjusted for baseline analysis)

# The scenario considered is for the baseline exposure waist cicumference (WC0),
# two longitudinal measures of the time-varying outcome insulin concentration
# (IC0, IC1), one baseline confounder (U2), and one mediator-outcome
# confounder (U) [all variables continuous]

################################################################################
## (1) SET UP ------------------------------------------------------------------

# Load required packages
if(packageVersion('dagitty') < "0.2.3"){
  warning("Please install at least version 0.2.3 of the dagitty package!")
  stop("Use this command: devtools::install_github('jtextor/dagitty/r')")
}

require(dagitty)
require(MASS)
require(rpsychi)

# Set simulation parameters
N <- 1000
Nreps <- 10000

### (a) Functions --------------------------------------------------------------

#### (i) runSims function -----------------------------------------------------

# This function executes multiple simulations and summarise findings

runSims  <- function(Means, Sigma, N, Nreps, Seed) {
  start  <- Sys.time()
  Sum    <- NULL
  for (itn in 1:Nreps) {
    seed       <- Seed*N*itn
    dat        <- data.frame(mvrnorm(N, Means, Sigma, empirical = FALSE))
    names(dat) <- c("X", "Y0", "Y1", "U2")[VarOrd]
    dat$DY     <- dat$Y1 - dat$Y0
    mod1       <- lm(DY ~ X, data = dat)
    mod2       <- lm(Y1 ~ X + Y0, data = dat)
    mod3       <- lm(Y1 ~ X, data = dat)
    Coeffs     <- c(mod1$coefficients[2], mod2$coefficients[2], mod3$coefficients[2])
    names(Coeffs) <- c("Beta1", "Beta2", "Beta3")
    Sum        <- rbind(Sum, Coeffs) }
  end   <- Sys.time(); print(end - start)
  Sims  <- apply(Sum, 2, function(x){quantile(x, c(0.025, 0.5, 0.975))})
  return(Sims) }

runSims2  <- function(Mu, Sigma, Nobs, Nsims, Seed) {
  start   <- Sys.time()
  Sum     <- NULL
  for (itn in 1:Nsims) {
    seed       <- Seed*Nobs*itn
    dat        <- data.frame(mvrnorm(Nobs, Mu, Sigma, empirical = FALSE))
    names(dat) <- c("X", "Y0", "Y1", "U2", "U")[VarOrd2]
    dat$DY     <- dat$Y1 - dat$Y0
    mod1       <- lm(DY ~ X, data = dat)
    mod2       <- lm(Y1 ~ X + Y0,data = dat)
    mod3       <- lm(Y1 ~ X, data = dat)
    Coeffs     <- c(mod1$coefficients[2], mod2$coefficients[2], mod3$coefficients[2])
    names(Coeffs) <- c("Beta1", "Beta2", "Beta3")
    Sum        <- rbind(Sum,Coeffs) }
  end    <- Sys.time(); print(end - start)
  Sims   <- apply(Sum, 2, function(x){quantile(x,c(0.025, 0.5, 0.975))})
  return(Sims) }

#### (ii) DAG functions -------------------------------------------------------

# These functions generate DAGs according to the given scenarios

# Scenarios 1 and 2:
DAG_base_confounder <- function(pU2_WC0, pU2_IC0, pU2_IC1,
                                pWC0_IC0, pWC0_IC1, pIC0_IC1) {
  dag  <- dagitty(paste0("dag{ U2->WC0 [beta=",pU2_WC0,"]
                           U2->IC0 [beta=",pU2_IC0,"]
                           U2->IC1 [beta=",pU2_IC1,"]
                           IC0->WC0 [beta=",pWC0_IC0,"]
                           WC0->IC1 [beta=",pWC0_IC1,"]
                           IC0->IC1 [beta=",pIC0_IC1,"]}"))
  return(dag) }

# Scenario 3A:
DAG_base_mediator <- function(pU2_WC0, pU2_IC0, pU2_IC1, pWC0_IC0,
                              pWC0_IC1, pIC0_IC1) {
  dag  <- dagitty(paste0("dag{ U2->WC0 [beta=",pU2_WC0,"]
                               U2->IC0 [beta=",pU2_IC0,"]
                               U2->IC1 [beta=",pU2_IC1,"]
                               IC0<-WC0 [beta=",pWC0_IC0,"]
                               WC0->IC1 [beta=",pWC0_IC1,"]
                               IC0->IC1 [beta=",pIC0_IC1,"]}"))
```

```
  return(dag) }

# Scenario 3B:
DAG_base_mediator2 <- function(pU2_WC0, pU2_IC0, pU2_IC1, pU_IC0,
                               pU_IC1, pWC0_IC0, pWC0_IC1, pIC0_IC1) {
  dag  <- dagitty(paste0("dag{ U2->WC0 [beta=",pU2_WC0,"]
                               U2->IC0 [beta=",pU2_IC0,"]
                               U2->IC1 [beta=",pU2_IC1,"]
                               IC0<-WC0 [beta=",pWC0_IC0,"]
                               U->IC0 [beta=",pU_IC0,"]
                               U->IC1 [beta=",pU_IC1,"]
                               WC0->IC1 [beta=",pWC0_IC1,"]
                               IC0->IC1 [beta=",pIC0_IC1,"]}"))
  return(dag) }

################################################################################
## (2) DATA SIMULATION ---------------------------------------------------------

### (a) Set up ----------------------------------------------------------------

# Set means and variances
WCmu <- 9.5
WCvar <- 1.6^2
IC0mu <- 4.0
IC0var <- 0.74^2
IC1mu <- 4.2
IC1var <- 0.74^2

# Set vectors for simulations
VarNames <- c("WC0", "IC0", "IC1", "U2")
Nmu <- c(WCmu, IC0mu, IC1mu, 0)
Nvar <- c(WCvar, IC0var, IC1var, 1)

# Set consistent path coefficient
pIC0_IC1 <- 0.65
pWC0_IC1 <- EffSize <- 0.433

# Set final output file identifier
Name <- paste0("final-", (Nreps/1000), "k", "-", substr(Sys.time(), 1, 10), ".csv")

### (b) Scenario 1 ------------------------------------------------------------

pU2_IC0 <- 0.0; pU2_WC0 <- 0.0; pU2_IC1 <- 0.0; pWC0_IC0 <- 0.0;
dag <- DAG_base_confounder(pU2_WC0, pU2_IC0, pU2_IC1,
                           pWC0_IC0, pWC0_IC1, pIC0_IC1); # plot(graphLayout(dag))
Cor <- impliedCovarianceMatrix(dag)
VarOrd <- as.integer(sapply(colnames(Cor), function(x){which(VarNames == x)}))
Means <- Nmu[VarOrd]
Sigma <- r2cov(sqrt(Nvar[VarOrd]), Cor)
Sim1 <- data.frame(runSims(Means, Sigma, N, Nreps, 13))
Filename <- paste0("1-full-", Name)
write.csv(Sim1, file = Filename, row.names = FALSE)

### (c) Scenario 1+ -----------------------------------------------------------

# (Scenario 1 plus baseline confounder U2)

pU2_IC0 <- 0.4; pU2_IC1 <- 0.04; pU2_WC0 <- 0.2; pWC0_IC0 <- 0.0
dag <- DAG_base_confounder(pU2_WC0, pU2_IC0, pU2_IC1,
                           pWC0_IC0, pWC0_IC1, pIC0_IC1)
Cor <- impliedCovarianceMatrix(dag)
VarOrd <- as.integer(sapply(colnames(Cor), function(x){which(VarNames ==x )}))
Means <- Nmu[VarOrd]
Sigma <- r2cov(sqrt(Nvar[VarOrd]), Cor)
Sim1plus <- data.frame(runSims(Means, Sigma, N, Nreps, 17))
Filename <- paste0("1plus-full-", Name)
write.csv(Sim1plus, file = Filename, row.names = FALSE)

### (d) Scenario 2 ------------------------------------------------------------

# set consistent path coefficient
pWC0_IC0 <- 0.5

pU2_IC0 <- 0.0; pU2_IC1 <- 0.0; pU2_WC0 <- 0.0
dag <- DAG_base_confounder(pU2_WC0, pU2_IC0, pU2_IC1,
                           pWC0_IC0, pWC0_IC1, pIC0_IC1); # plot(graphLayout(dag))
Cor <- impliedCovarianceMatrix(dag)
VarOrd <- as.integer(sapply(colnames(Cor), function(x){which(VarNames == x)}))
Means <- Nmu[VarOrd]
Sigma <- r2cov(sqrt(Nvar[VarOrd]), Cor)
Sim2 <- data.frame(runSims(Means, Sigma, N, Nreps, 19))
Filename <- paste0("2-full-", Name)
write.csv(Sim2, file = Filename, row.names = FALSE)

### (e) Scenario 2+ -----------------------------------------------------------

# (Scenario 2 plus baseline confounder U2)

pU2_IC0 <- 0.4; pU2_IC1 <- 0.04; pU2_WC0 <- 0.2
dag <- DAG_base_confounder(pU2_WC0, pU2_IC0, pU2_IC1,
                           pWC0_IC0, pWC0_IC1, pIC0_IC1)
Cor <- impliedCovarianceMatrix(dag)
VarOrd <- as.integer(sapply(colnames(Cor), function(x){which(VarNames == x)}))
Means <- Nmu[VarOrd]
Sigma <- r2cov(sqrt(Nvar[VarOrd]), Cor)
Sim2plus <- data.frame(runSims(Means, Sigma, N, Nreps, 23))
Filename <- paste0("2plus-full-", Name)
write.csv(Sim2plus, file = Filename, row.names = FALSE)

### (f) Scenario 3A -----------------------------------------------------------

pU2_IC0 <- 0.0; pU2_IC1 <- 0.0; pU2_WC0 <- 0.0
pWC0_IC1 <- EffSize - (pIC0_IC1*pWC0_IC0)
```

```
dag <- DAG_base_mediator(pU2_WC0, pU2_IC0, pU2_IC1,
                         pWC0_IC0, pWC0_IC1, pIC0_IC1)
Cor <- impliedCovarianceMatrix(dag)
VarOrd <- as.integer(sapply(colnames(Cor), function(x){which(VarNames ==x )}))
Means <- Nmu[VarOrd]
Sigma <- r2cov(sqrt(Nvar[VarOrd]), Cor)
Sim3a <- data.frame(runSims(Means, Sigma, N, Nreps, 29))
Filename <- paste0("3A-full-", Name)
write.csv(Sim3a, file = Filename, row.names = FALSE)

### (g) Scenario 3A+ ------------------------------------------------------------

# (Scenario 3A plus baseline confounder U2)

pU2_IC0 <- 0.4; pU2_IC1 <- 0.04; pU2_WC0 <- 0.2
pWC0_IC1 <- EffSize - (pIC0_IC1*pWC0_IC0)
dag <- DAG_base_mediator(pU2_WC0, pU2_IC0, pU2_IC1,
                         pWC0_IC0, pWC0_IC1, pIC0_IC1)
Cor <- impliedCovarianceMatrix(dag)
VarOrd <- as.integer(sapply(colnames(Cor), function(x){which(VarNames == x)}))
Means <- Nmu[VarOrd]
Sigma <- r2cov(sqrt(Nvar[VarOrd]), Cor)
Sim3aplus <- data.frame(runSims(Means, Sigma, N, Nreps, 31))
Filename <- paste0("3Aplus-full-", Name)
write.csv(Sim3aplus, file = Filename, row.names = FALSE)

### (h) Scenario 3B ------------------------------------------------------------

# (mediator-outcome confounder U)

# Reset vetors for simulations
VarNames <- c("WC0", "IC0", "IC1", "U2", "U")
Nmu <- c(WCmu, IC0mu, IC1mu, 0, 0)
Nvar <- c(WCvar, IC0var, IC1var, 1, 1)

pU2_IC0 <- 0.0; pU2_IC1 <- 0.0; pU2_WC0  <- 0.0
pU_IC0 <- 0.4; pU_IC1 <- 0.2
pWC0_IC1 <- EffSize - (pIC0_IC1*pWC0_IC0)
dag <- DAG_base_mediator2(pU2_WC0, pU2_IC0, pU2_IC1, pU_IC0,
                          pU_IC1, pWC0_IC0, pWC0_IC1, pIC0_IC1)
Cor <- impliedCovarianceMatrix(dag)
VarOrd2 <- as.integer(sapply(colnames(Cor), function(x){which(VarNames == x)}))
Means <- Nmu[VarOrd2]
Sigma <- r2cov(sqrt(Nvar[VarOrd2]), Cor)
Sim3b <- data.frame(runSims2(Means, Sigma, N, Nreps, 37))
Filename <- paste0("3B-full-", Name)
write.csv(Sim3b, file = Filename, row.names = FALSE)

### (i) Scenario 3B+ ------------------------------------------------------------

# (Scenario 3B plus baseline confounder U2)

pU2_IC0 <- 0.4; pU2_IC1 <- 0.04; pU2_WC0  <- 0.2
pU_IC0 <- 0.4; pU_IC1 <- 0.2
pWC0_IC1 <- EffSize - (pIC0_IC1*pWC0_IC0)
dag <- DAG_base_mediator2(pU2_WC0, pU2_IC0, pU2_IC1, pU_IC0,
                          pU_IC1, pWC0_IC0, pWC0_IC1, pIC0_IC1)
Cor <- impliedCovarianceMatrix(dag)
VarOrd2 <- as.integer(sapply(colnames(Cor), function(x){which(VarNames == x)}))
Means <- Nmu[VarOrd2]
Sigma <- r2cov(sqrt(Nvar[VarOrd2]), Cor)
Sim3bplus <- data.frame(runSims2(Means, Sigma, N, Nreps, 41))
Filename <- paste0("3Bplus-full-", Name)
write.csv(Sim3bplus, file = Filename, row.names = FALSE)

### (h) Export data ------------------------------------------------------------

Summ      <- NULL
for (itn in c("1", "1plus", "2", "2plus", "3A", "3Aplus", "3B", "3Bplus")) {
  Label    <- paste0(itn, "-full")
  Filename <- paste0(Label, "-",Name)
  File     <- read.csv(Filename)[, c("Beta1", "Beta2", "Beta3")]
  rownames(File) <- paste(Label, c("2.5%", "50%", "97.5%"))
  Summ     <- rbind(Summ, File) }

write.table(signif(Summ), Name, sep = ",", row.names = TRUE, col.names = N
```

Appendix B
Regression with 'unexplained residuals'

B.1 Introduction

This appendix contains additional material relating to Chapter 5. In particular, it contains formal mathematical proofs relating to the key properties of UR models, and details relating to the simulation in Section 5.6 which demonstrates the artefactual standard error reduction.

We first define the key three properties of UR models for a longitudinal exposure measured at T time points (§B.2). We also define the key properties of ordinary least squares (OLS) regression estimators and introduce two lemmas (§B.3), which will be required in the subsequent sections. We then prove the three properties of UR models for a longitudinal exposure measured at T time points in the absence of any confounders (§B.4), where there exists a baseline confounder (§B.5), and where there exists a time-dependent confounder (§B.6). We finally provide additional details relating to the simulation which demonstrates the artefactual standard error reduction in UR models, including annotated R code (§B.7).

B.2 Key properties of UR models for a longitudinal exposure measured at k time points

The key properties of UR models for a longitudinal exposure X measured at T time points (i.e. $X_0, X_1, \ldots, X_{T-1}$) are summarised in Table B.1.

Note that the original scenario examined by Keijzer-Veen et al. (93) is equivalent to the scenario considered here where $T = 2$.

Table B.1 Description of key properties of UR models for a longitudinal exposure X measured at T time points (i.e. $X_0, X_1, \ldots, X_{T-1}$) and one outcome Y

Property	Description	Mathematical formulation
(i)	The outcome values predicted by the final standard regression model (i.e. for exposure X_{T-1}) are equal to those predicted by the UR model.	$\hat{Y}_S^{(T-1)} = \hat{Y}_{UR}^{(T-1)}$
(ii)	The estimated coefficient for X_0 in the initial standard regression model (i.e. for exposure X_0) is equal to the estimated coefficient for X_0 in the UR model.	$\hat{\alpha}_{X0}^{(0)} = \hat{\lambda}_{X0}^{(T-1)}$
(iii)	The estimated coefficient for each X_t in its individual standard regression model (i.e. for exposure X_t) is equal to the estimated coefficient for the corresponding UR term e_{Xt} in the UR model, for $1 \leq t \leq (T-1)$.	$\hat{\alpha}_{Xt}^{(i)} = \hat{\lambda}_{eXt}^{(i)}$

B.3 Lemmas

The proofs that follow (in §B.4, §B.5, and §B.6) rely upon the following key properties of ordinary least squares (OLS) regression estimators and two lemmas.

B.3.1 Key properties of ordinary least squares (OLS) estimators

We may represent the regression equation $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_T x_T + \varepsilon$ in summary notation as:

$$y = X\beta + \varepsilon \,,$$

where: y represents the vector of n continuous observations of the outcome; X represents the $n \times (T+1)$ matrix of n observations for T continuous covariates and 1 constant; β represents the $T+1$ vector of coefficients for each covariate and constant; and ε represents the vector of n residuals.

The OLS estimate of β is given by:

$$\hat{\beta} = (X'X)^{-1}X'y$$

On the assumption that the inverse matrix exists, this equation has a unique solution.

Further, for the given OLS equation $y = X\hat{\beta} + e$, it can be shown that the vector of residuals (e) is orthogonal (denoted $\perp$) to every column $(1, x_1, x_2, \ldots x_T)$ of X.[29]

[29] Note that detailed proofs have not been provided, but can be located in the referenced material (203).

B.3.2 Lemma 1

For two orthogonal components τ and δ (i.e. $\tau \perp \delta$), the estimated coefficients of the regression of y on τ and δ are equal to the estimated coefficients for the separate regressions of y on τ and y on δ.

Proof: The regression of y on τ and δ may be written as:

$$y = [\tau \quad \delta]\begin{bmatrix}\beta_\tau \\ \beta_\delta\end{bmatrix} + \epsilon = \tau\beta_\tau + \delta\beta_\delta + \varepsilon$$

From Definition 1, the OLS estimate of β_τ and β_δ is given by $\hat{\beta} = (X'X)^{-1}X'y$. In this scenario,

$$X'X = \begin{bmatrix}\tau' \\ \delta'\end{bmatrix}[\tau \quad \delta] = \begin{bmatrix}\tau'\tau & \tau'\delta \\ \delta'\tau & \delta'\delta\end{bmatrix} = \begin{bmatrix}\tau'\tau & 0 \\ 0 & \delta'\delta\end{bmatrix}$$

where the final equivalency follows from the condition of orthogonality. Then

$$(X'X)^{-1} = \begin{bmatrix}\tau'\tau & 0 \\ 0 & \delta'\delta\end{bmatrix}^{-1} = \begin{bmatrix}(\tau'\tau)^{-1} & 0 \\ 0 & (\delta'\delta)^{-1}\end{bmatrix}$$

and

$$X'y = \begin{bmatrix}\tau' \\ \delta'\end{bmatrix}y = \begin{bmatrix}\tau'y \\ \delta'y\end{bmatrix}$$

Combining these elements gives:

$$\begin{bmatrix}\hat{\beta}_\tau \\ \hat{\beta}_\delta\end{bmatrix} = \begin{bmatrix}(\tau'\tau)^{-1} & 0 \\ 0 & (\delta'\delta)^{-1}\end{bmatrix}\begin{bmatrix}\tau'y \\ \delta'y\end{bmatrix} = \begin{bmatrix}(\tau'\tau)^{-1}\tau'y \\ (\delta'\delta)^{-1}\delta'y\end{bmatrix}$$

From this, we see that the estimated coefficients are equivalent to those that would be produced for the separate regressions of y on τ and y on δ. ■

B.3.3 Lemma 2

If $\tau_i \perp \delta_j$ for $0 \leq i \leq h$ and $0 \leq j \leq k$, then $span(\tau_0, \tau_1, \dots, \tau_h) \perp span(\delta_0, \delta_1, \dots \delta_k)$ for any vectors $\tau_0, \tau_1, \dots, \tau_h, \delta_0, \delta_1, \dots \delta_k$. [30]

Proof: $\tau_i \perp \delta_j$ implies that $\tau_i \cdot \delta_j = 0$ for $0 \leq i \leq h$ and $0 \leq j \leq k$. Then

$$\begin{aligned}
&span(\tau_0, \tau_1, \dots, \tau_h) \cdot span(\delta_0, \delta_1, \delta_2, \dots \delta_k) \\
&\quad = (c_0\tau_0 + c_1\tau_1 + \cdots + c_h\tau_h) \cdot (d_0\delta_0 + d_1\delta_1 + \cdots + d_k\delta_k) \\
&\quad = c_0d_0(\tau_0 \cdot \delta_0) + c_0d_1(\tau_0 \cdot \delta_1) + \cdots + c_0d_k(\tau_0 \cdot \delta_k) + c_1d_0(\tau_1 \cdot \delta_0) + \\
&\qquad c_1d_1(\tau_1 \cdot \delta_1) + \cdots + c_1d_k(\tau_1 \cdot \delta_k) + \cdots + c_hd_0(\tau_h \cdot \delta_0) + c_hd_1(\tau_h \cdot \delta_1) + \\
&\qquad \cdots + c_hd_k(\tau_h \cdot \delta_k) \\
&\quad = c_0d_0(0) + c_0d_1(0) + \cdots + c_0d_k(0) + c_1d_0(0) + c_1d_1(0) + \cdots + c_1d_k(0) + \\
&\qquad \cdots + c_hd_0(0) + c_hd_1(0) + \cdots + c_hd_k(0)
\end{aligned}$$

---

[30] The span of a set of vectors $\delta_0, \delta_1, \delta_2, \dots \delta_k$ is the set of all possible linear combinations of $\delta_0, \delta_1, \delta_2, \dots \delta_k$, i.e.: $span(\delta_0, \delta_1, \delta_2, \dots \delta_k) = c_0\delta_0 + c_1\delta_1 + c_2\delta_2 + \cdots + c_k\delta_k$, where the coefficients $c_0, c_1, c_2, \dots, c_k$ are scalars.

$$= 0$$

Thus, $span(\tau_0, \tau_1, \dots, \tau_h) \perp span(\delta_0, \delta_1, \delta_2, \dots \delta_k)$. ■

B.4 UR models with no confounders (Figure 5.10)

B.4.1 Definitions

B.4.1.1 Definition 1

We define the ordinary least-squares (OLS) regression model $\hat{Y}_S^{(t)}$ for each measurement of the exposure variable X_t, for $0 \leq t \leq (T-1)$. Because the relationship between X_t and Y is confounded by all previous values of X (i.e. $X_0, X_1, \dots, X_{t-1}$), we represent Y as a function of $1, X_0, X_1, \dots, X_t$:

$$\hat{Y}_S^{(0)} = \hat{\alpha}_0^{(0)} + \hat{\alpha}_{X0}^{(0)} X_0$$

$$\hat{Y}_S^{(1)} = \hat{\alpha}_0^{(1)} + \hat{\alpha}_{X0}^{(1)} X_0 + \hat{\alpha}_{X1}^{(1)} X_1$$

$$\vdots$$

$$\hat{Y}_S^{(T-1)} = \hat{\alpha}_0^{(T-1)} + \hat{\alpha}_{X0}^{(T-1)} X_0 + \hat{\alpha}_{X1}^{(T-1)} X_1 + \cdots + \hat{\alpha}_{X(T-1)}^{(T-1)} X_{T-1} \quad \textbf{Equation B.1}$$

The coefficient of the last/most recent measurement of X (i.e. $\hat{\alpha}_{Xt}^{(t)}$) may be interpreted as the total causal effect of X_t on Y.

B.4.1.2 Definition 2

As established by Keijzer-Veen et al. (93), each UR term e_{Xt} is derived from the OLS regression of X_t on all previous measurements of X (i.e. $X_0, X_1, \dots, X_{t-1}$):

$$X_t = \hat{\gamma}_0^{(t)} + \hat{\gamma}_{X0}^{(t)} X_0 + \hat{\gamma}_{X1}^{(t)} X_1 + \cdots + \hat{\gamma}_{X(t-1)}^{(t)} X_{t-1} + e_{Xt} \quad \textbf{Equation B.2}$$

for $1 \leq t \leq (T-1)$. Thus,

$$e_{X1} = -\hat{\gamma}_0^{(1)} - \hat{\gamma}_{X0}^{(1)} X_0 + X_1$$

$$e_{X2} = -\hat{\gamma}_0^{(2)} - \hat{\gamma}_{X0}^{(2)} X_0 - \hat{\gamma}_{X1}^{(2)} X_1 + X_2$$

$$\vdots$$

$$e_{X(T-1)} = -\hat{\gamma}_0^{(T-1)} - \hat{\gamma}_{X0}^{(T-1)} X_0 - \hat{\gamma}_{X1}^{(T-1)} X_1 - \cdots - \hat{\gamma}_{X(T-2)}^{(T-1)} X_{T-2} + X_{T-1} \quad \textbf{Equation B.3}$$

By its formulation, e_{Xt} represents the difference between the actual value of X_t and the value of X_t as predicted by all previous measurements of X.

B.4.1.3 Definition 3

The UR model $\hat{Y}_{UR}^{(t)}$ is defined as an OLS regression model which represents Y as a function of $1, X_0, e_{X1}, \dots, e_{Xt}$, for $0 \leq t \leq (T-1)$:

$$\hat{Y}_{UR}^{(0)} = \hat{\lambda}_0^{(0)} + \hat{\lambda}_{X0}^{(0)} X_0$$

$$\hat{Y}_{UR}^{(1)} = \hat{\lambda}_0^{(1)} + \hat{\lambda}_{X0}^{(1)} X_0 + \hat{\lambda}_{eX1}^{(1)} e_{X1}$$

$$\vdots$$

$$\hat{Y}_{UR}^{(T-1)} = \hat{\lambda}_0^{(T-1)} + \hat{\lambda}_{X0}^{(T-1)} X_0 + \hat{\lambda}_{eX1}^{(T-1)} e_{X1} + \cdots + \hat{\lambda}_{eX(T-1)}^{(T-1)} e_{X(T-1)} \qquad \textbf{Equation B.4}$$

B.4.2 Mathematical proofs

B.4.2.1 Covariate orthogonality

We prove in Lemma 3 that all UR terms $e_{X1}, e_{X2}, \dots, e_{X(T-1)}$ are orthogonal to all preceding variables in the composite UR model (Equation B.4), and therefore orthogonal to their span in Theorem 1.

B.4.2.1.1 Lemma 3

$e_{Xt} \perp e_{X1}, e_{X2}, \dots, e_{X(t-1)}$, for $1 \leq t \leq (T-1)$.

Proof: By construction, e_i represents the residuals from the OLS regression of $X_t \sim 1, X_0, X_1, \dots, X_{t-1}$ (Equation B.2). Thus, $e_{Xt} \perp 1, X_0, X_1, \dots, X_{t-1}$, which implies that $e_{Xt} \perp span(1, X_0, X_1, \dots, X_{t-1})$ by Lemma 2.

It is clear that $e_{X1}, e_{X2}, \dots, e_{X(t-1)} \in span(1, X_0, X_1, \dots, X_{t-1})$ for $1 \leq t \leq (T-1)$ by construction; we are therefore able to conclude that $e_{Xt} \perp e_{X1}, e_{X2}, \dots, e_{X(t-1)}$. ■

B.4.2.1.2 Theorem 1

$e_{Xt} \perp span(1, X_0, e_{X1}, e_{X2}, \dots, e_{X(t-1)})$, for $1 \leq t \leq (T-1)$.

Proof: $e_{Xt} \perp 1, X_0$ because e_{Xt} represents the residuals from the OLS regression of $X_t \sim 1, X_0, X_1, \dots, X_{t-1}$. Further, $e_{Xt} \perp e_{X1}, e_{X2}, \dots, e_{X(t-1)}$ for $1 \leq t \leq (T-1)$ by Lemma 3.

Thus, $e_{Xt} \perp span(1, X_0, e_{X1}, e_{X2}, \dots, e_{X(t-1)})$ by Lemma 2. ■

B.4.2.2 Property (i)

$\hat{Y}_S^{(T-1)} = \hat{Y}_{UR}^{(T-1)}$.

Proof: This equality follows from the fact that each UR model $\hat{Y}_{UR}^{(t)}$ is a function of the same variables as the corresponding standard regression model $\hat{Y}_S^{(t)}$.

By Definition 3, $\hat{Y}_{UR}^{(t)} = f(1, X_0, e_{X1}, \dots, e_{Xt})$, where $e_{Xt} = f(1, X_0, X_1, \dots, X_t)$ by Definition 2. Thus, it also holds that

$$\hat{Y}_{UR}^{(t)} = f(1, X_0, X_1, \dots, X_t)$$

Moreover, by Definition 1,

$$\hat{Y}_S^{(t)} = f(1, X_0, X_1, \dots, X_t)$$

From this, it follows that $\hat{Y}_S^{(t)} = \hat{Y}_{UR}^{(t)}$ and, consequently, $\hat{Y}_S^{(T-1)} = \hat{Y}_{UR}^{(T-1)}$. ■

B.4.2.3 Property (ii)

$\hat{\alpha}_{X0}^{(0)} = \hat{\lambda}_{X0}^{(T-1)}$.

Proof: By definition, $\hat{Y}_S^{(0)} = \hat{Y}_{UR}^{(0)} = f(1, X_0)$, and so it is trivially true that $\hat{\alpha}_{X0}^{(0)} = \hat{\lambda}_{X0}^{(0)}$.

Because $e_{Xt} \perp span(1, X_0, e_{X1}, e_{X2}, \dots, e_{X(t-1)})$ for $1 \leq t \leq (T-1)$ by Theorem 1, we are able to apply Lemma 1 and conclude that $\hat{\lambda}_{X0}^{(0)} = \hat{\lambda}_{X0}^{(1)} = \cdots = \hat{\lambda}_{X0}^{(T-1)}$.

Therefore, $\hat{\alpha}_{X0}^{(0)} = \hat{\lambda}_{X0}^{(T-1)}$. ■ [31]

B.4.2.4 Property (iii)

$\hat{\alpha}_{Xt}^{(t)} = \hat{\lambda}_{eXt}^{(T-1)}$.

Proof: Consider the UR model:

$$\hat{Y}_{UR}^{(t)} = \hat{\lambda}_0^{(t)} + \hat{\lambda}_{X0}^{(t)} X_0 + \hat{\lambda}_{eX1}^{(t)} e_{X1} + \cdots + \boldsymbol{\hat{\lambda}_{eXt}^{(t)}} e_{Xt}$$

If we substitute the expansion for e_{Xt} (Equation B.3) into this equation and rearrange, we produce:

$$\begin{aligned}\hat{Y}_{UR}^{(t)} &= \hat{\lambda}_0^{(t)} + \hat{\lambda}_{X0}^{(t)} X_0 + \hat{\lambda}_{eX1}^{(t)}[-\hat{\gamma}_0^{(1)} - \hat{\gamma}_{X0}^{(1)} X_0 + X_1] + \cdots + \hat{\lambda}_{eXt}^{(t)}[-\hat{\gamma}_0^{(t)} - \hat{\gamma}_{X0}^{(t)} X_0 - \\ &\qquad \hat{\gamma}_{X1}^{(t)} X_1 - \cdots - \hat{\gamma}_{X(t-1)}^{(t)} X_{t-1} + X_t] \\ &= [\hat{\lambda}_0^{(t)} - \hat{\lambda}_{eX1}^{(t)} \gamma_0^{(1)} - \cdots - \hat{\lambda}_{eXi}^{(t)} \gamma_0^{(t)}] + \left[\hat{\lambda}_{x0}^{(t)} - \hat{\lambda}_{eX1}^{(t)} \gamma_{X0}^{(1)} - \cdots - \hat{\lambda}_{eXt}^{(t)} \gamma_{X0}^{(t)}\right] X_0 + \\ &\qquad \left[\hat{\lambda}_{eX1}^{(t)} - \hat{\lambda}_{eX2}^{(t)} \gamma_{X1}^{(2)} - \cdots - \hat{\lambda}_{eXt}^{(t)} \gamma_{X1}^{(t)}\right] X_1 + \cdots + \left[\boldsymbol{\hat{\lambda}_{eXt}^{(t)}}\right] X_t\end{aligned}$$

Since we have already established that $\hat{Y}_S^{(t)} = \hat{Y}_{UR}^{(t)}$ (i.e. Property (i)) because they are functions of the same covariates, it follows that the estimated coefficients for those covariates must themselves be equal. Specifically, we are able to see that the coefficient for X_t will always equal the coefficient for e_{Xt}, i.e. $\hat{\alpha}_{Xt}^{(t)} = \hat{\lambda}_{eXt}^{(t)}$.

Finally, because $e_{Xt} \perp span(1, X_0, e_{X1}, e_{X2}, \dots, e_{X(t-1)})$, we can again apply Lemma 1 and conclude that $\hat{\lambda}_{eXt}^{(0)} = \hat{\lambda}_{eXt}^{(1)} = \cdots = \hat{\lambda}_{eXt}^{(T-1)}$, from which it follows that $\hat{\alpha}_{Xt}^{(t)} = \hat{\lambda}_{eXt}^{(T-1)}$. ■

B.5 UR models with baseline confounding (Figure 5.11)

B.5.1 Definitions

B.5.1.1 Definition 4

Because the relationship between each measurement X_t and Y is confounded by M (for $0 \leq t \leq (T-1)$), adjustment for M is necessary to estimate the total effect of X_t on Y in the standard regression models:

$$\begin{aligned}\hat{Y}_S^{(0)} &= \hat{\alpha}_0^{(0)} + \hat{\alpha}_M^{(0)} M + \boldsymbol{\hat{\alpha}_{X0}^{(0)} X_0} \\ \hat{Y}_S^{(1)} &= \hat{\alpha}_0^{(1)} + \hat{\alpha}_M^{(1)} M + \hat{\alpha}_{X0}^{(1)} X_0 + \boldsymbol{\hat{\alpha}_{X1}^{(1)} X_1} \\ &\vdots \\ \hat{Y}_S^{(T-1)} &= \hat{\alpha}_0^{(T-1)} + \hat{\alpha}_M^{(T-1)} M + \hat{\alpha}_{X0}^{(T-1)} X_0 + \hat{\alpha}_{X1}^{(T-1)} X_1 + \cdots + \boldsymbol{\hat{\alpha}_{X(T-1)}^{(T-1)} X_{T-1}}\end{aligned}$$ **Equation B.5**

---

[31] Although no causal meaning/significance can be attributed to the intercept term, the logic applied in this proof may be easily extended to show that $\hat{\alpha}_0^{(0)} = \hat{\lambda}_0^{(T-1)}$.

B.5.1.2 Definition 5

The relationship between X_t and $X_0, X_1, \dots, X_{t-1}$ for $1 \leq t \leq (T-1)$ is confounded by M, and thus adjustment for M is necessary when regressing $X_t \sim X_0, X_1, \dots, X_{t-1}$ to generate each UR term e_{Xt}, i.e.:

$$X_t = \hat{\gamma}_0^{(t)} + \hat{\gamma}_M^{(t)} M + \hat{\gamma}_{X0}^{(t)} X_0 + \hat{\gamma}_{X1}^{(t)} X_1 + \cdots + \hat{\gamma}_{X(t-1)}^{(t)} X_{t-1} + e_{Xt} \qquad \textbf{Equation B.6}$$

and

$$e_{Xt} = -\hat{\gamma}_0^{(t)} - \hat{\gamma}_M^{(t)} M - \hat{\gamma}_{X0}^{(t)} X_0 - \hat{\gamma}_{X1}^{(t)} X_1 - \cdots - \hat{\gamma}_{X(t-1)}^{(t)} X_{t-1} + X_t \qquad \textbf{Equation B.7}$$

In this way, e_{Xt} represents the difference between the actual value of X_t and the value of X_t as predicted by all previous measurements $M, X_0, X_1, \dots, X_{t-1}$.

B.5.1.3 Definition 6

M also confounds the relationship between X_0 and Y, and so adjustment must be made in the composite UR model:

$$\hat{Y}_{UR}^{(T-1)} = \hat{\lambda}_0^{(T-1)} + \hat{\lambda}_M^{(T-1)} M + \hat{\lambda}_{X0}^{(T-1)} X_0 + \hat{\lambda}_{eX1}^{(T-1)} e_{X1} + \cdots + \hat{\lambda}_{eX(T-1)}^{(T-1)} e_{X(T-1)}$$

Equation B.8

B.5.2 Mathematical proofs

B.5.2.1 Covariate orthogonality

We prove in Lemma 4 that all UR terms $e_{X1}, e_{X2}, \dots, e_{X(T-1)}$ are orthogonal to all preceding variables in the composite UR model

(Equation B.8), and therefore orthogonal to their span in Theorem 2.

B.5.2.1.1 Lemma 4

$e_{Xt} \perp e_{X1}, e_{X2}, \dots, e_{X(t-1)}$, for $1 \leq t \leq (T-1)$.

Proof: By construction, e_{Xt} represents the residuals from the OLS regression of $X_t \sim 1, M, X_0, X_1, \dots, X_{t-1}$ (Equation B.6). Thus, $e_{Xt} \perp 1, M, X_0, X_1, \dots, X_{t-1}$, from which it follows that $e_{Xt} \perp span(1, M, X_0, X_1, \dots, X_{t-1})$ by Lemma 2.

Because $e_{X1}, e_{X2}, \dots, e_{X(t-1)} \in span(1, M, X_0, X_1, \dots, X_{t-1})$ for $1 \leq t \leq (T-1)$ by construction, we are able to conclude that $e_{Xt} \perp e_{X1}, e_{X2}, \dots, e_{X(t-1)}$. ■

B.5.2.1.2 Theorem 2

$e_{Xt} \perp span(1, M, X_0, e_{X1}, e_{X2}, \dots, e_{X(t-1)})$, for $1 \leq t \leq (T-1)$.

Proof: $e_{Xt} \perp 1, M, X_0$ because e_{Xt} represents the residuals from the OLS regression of $X_t \sim 1, M, X_0, X_1, \dots, X_{t-1}$. Further, $e_{Xt} \perp e_{X1}, e_{X2}, \dots, e_{X(t-1)}$ for $1 \leq t \leq (T-1)$ by Lemma 4 above.

Thus, $e_{Xt} \perp span(1, M, X_0, e_{X1}, e_{X2}, \dots, e_{X(t-1)})$ by Lemma 2. ■

B.5.2.2 Property (i)

$\hat{Y}_S^{(T-1)} = \hat{Y}_{UR}^{(T-1)}$

Proof: As before, this equality follows from the fact that $\hat{Y}_{UR}^{(t)}$ is a function of the same variables as $\hat{Y}_S^{(t)}$.

By Definition 6, $\hat{Y}_{UR}^{(t)} = f(1, M, X_0, e_{X1}, \dots, e_{Xt})$, where $e_{Xt} = f(1, M, X_0, X_1, \dots, X_t)$ by Definition 5. Thus, it also holds that

$$\hat{Y}_{UR}^{(t)} = f(1, M, X_0, X_1, \dots, X_t)$$

Moreover, by Definition 4,

$$\hat{Y}_S^{(t)} = f(1, M, X_0, X_1, \dots, X_t)$$

From this, it follows that $\hat{Y}_S^{(t)} = \hat{Y}_{UR}^{(t)}$ and, consequently, $\hat{Y}_S^{(T-1)} = \hat{Y}_{UR}^{(T-1)}$. ■

B.5.2.3 Property (ii)

$\hat{\alpha}_{X0}^{(0)} = \hat{\lambda}_{X0}^{(T-1)}$

Proof: By definition, $\hat{Y}_S^{(0)} = \hat{Y}_{UR}^{(0)} = f(1, M, X_0)$, and it is trivially true that $\hat{\alpha}_{X0}^{(0)} = \hat{\lambda}_{X0}^{(0)}$.

Because $e_{Xt} \perp span(1, M, X_0, e_{X1}, e_{X2}, \dots, e_{X(t-1)})$ for $1 \leq t \leq (T-1)$ by Theorem 2, we conclude that $\hat{\lambda}_{X0}^{(0)} = \hat{\lambda}_{X0}^{(1)} = \cdots = \hat{\lambda}_{X0}^{(T-1)}$ from Lemma 1.

Therefore, $\hat{\alpha}_{X0}^{(0)} = \hat{\lambda}_{X0}^{(T-1)}$. ■ [32]

B.5.2.4 Property (iii)

$\hat{\alpha}_{Xt}^{(t)} = \hat{\lambda}_{eXt}^{(T-1)}$

Proof: Consider the UR model:

$$\hat{Y}_{UR}^{(t)} = \hat{\lambda}_0^{(t)} + \hat{\lambda}_M^{(t)} M + \hat{\lambda}_{X0}^{(t)} X_0 + \hat{\lambda}_{eX1}^{(t)} e_{X1} + \cdots + \hat{\boldsymbol{\lambda}}_{\boldsymbol{eXt}}^{(\boldsymbol{t})} e_{Xt}$$

If we substitute the expansion for e_{Xt} (Equation B.7) into this equation and rearrange, we produce:

$$\begin{aligned}\hat{Y}_{UR}^{(t)} &= \hat{\lambda}_0^{(t)} + \hat{\lambda}_M^{(t)} M + \hat{\lambda}_{X0}^{(t)} X_0 + \hat{\lambda}_{eX1}^{(t)} \left[-\hat{\gamma}_0^{(1)} - \hat{\gamma}_{X0}^{(1)} X_0 + X_1 - \hat{\gamma}_M^{(2)} M\right] + \cdots + \\ &\quad \hat{\lambda}_{eXt}^{(t)} \left[-\hat{\gamma}_0^{(t)} - \hat{\gamma}_{X0}^{(t)} X_0 - \hat{\gamma}_{X1}^{(t)} X_1 - \cdots - \hat{\gamma}_{X(t-1)}^{(t)} X_{t-1} + X_t - \hat{\gamma}_M^{(t)} M\right] \\ &= \left[\hat{\lambda}_0^{(t)} - \hat{\lambda}_{eX1}^{(t)} \gamma_0^{(1)} - \cdots - \hat{\lambda}_{eXt}^{(t)} \gamma_0^{(t)}\right] + \left[\hat{\lambda}_{X0}^{(t)} - \hat{\lambda}_{eX1}^{(t)} \gamma_{X0}^{(1)} - \cdots - \hat{\lambda}_{eXt}^{(t)} \gamma_{X0}^{(t)}\right] X_0 + \\ &\quad \left[\hat{\lambda}_{eX1}^{(t)} - \hat{\lambda}_{eX2}^{(t)} \gamma_{X2}^{(2)} - \cdots - \hat{\lambda}_{eXt}^{(t)} \gamma_{X1}^{(t)}\right] X_1 + \cdots + \left[\hat{\boldsymbol{\lambda}}_{\boldsymbol{eXt}}^{(\boldsymbol{t})}\right] X_t + \left[\hat{\lambda}_M^{(t)} - \hat{\lambda}_{eX1}^{(t)} \gamma_M^{(1)} - \right. \\ &\quad \left. \hat{\lambda}_{eXt}^{(t)} \gamma_M^{(t)}\right] M\end{aligned}$$

We have already established that $\hat{Y}_S^{(t)} = \hat{Y}_{UR}^{(t)}$ (i.e. Property (i)) because they are functions of the same covariates, so it follows that the estimated coefficients for those covariates must

---

[32] Although no causal meaning/significance can be attributed to the coefficient of the confounder M, the logic applied in this proof may be easily extended to show that $\hat{\alpha}_M^{(0)} = \hat{\lambda}_M^{(T-1)}$.

themselves be equal. Specifically, we see that the coefficient for X_t will always equal the coefficient for e_{Xt}, i.e. $\hat{\alpha}_{Xt}^{(t)} = \hat{\lambda}_{eXt}^{(t)}$.

Because $e_{Xt} \perp span(1, M, X_0, e_{X1}, e_{X2}, \dots, e_{X(t-1)})$, we may apply Lemma 1 and conclude that $\hat{\lambda}_{eXt}^{(0)} = \hat{\lambda}_{eXt}^{(1)} = \cdots = \hat{\lambda}_{eXt}^{(T-1)}$, from which it follows that $\hat{\alpha}_{Xt}^{(t)} = \hat{\lambda}_{eXt}^{(T-1)}$. ■

B.6 UR models with time-dependent confounding (Figure 5.12)

B.6.1 Definitions

B.6.1.1 Definition 7

The relationship between each X_t and Y is confounded by all previous measurements of the exposure $X_0, X_1, \dots, X_{t-1}$, as well as all previous and current measurements of the confounder $M_0, M_1, \dots, M_t$ (for $0 \leq t \leq (T-1)$). These covariates must all be included in the standard regression models to obtain an unbiased estimate of the total causal effect of each measurement X_t on Y, i.e.:

$$\hat{Y}_S^{(0)} = \hat{\alpha}_0^{(0)} + \hat{\alpha}_{M0}^{(0)} M_0 + \boldsymbol{\hat{\alpha}_{X0}^{(0)} X_0}$$

$$\hat{Y}_S^{(1)} = \hat{\alpha}_0^{(1)} + \hat{\alpha}_{M0}^{(1)} M_1 + \hat{\alpha}_{X0}^{(1)} X_0 + \hat{\alpha}_{M1}^{(1)} M_1 + \boldsymbol{\hat{\alpha}_{X1}^{(1)} X_1}$$

$$\vdots$$

$$\hat{Y}_S^{(T-1)} = \hat{\alpha}_0^{(T-1)} + \hat{\alpha}_{M0}^{(T-1)} M_0 + \hat{\alpha}_{X0}^{(T-1)} X_0 + \cdots + \hat{\alpha}_{M(T-1)}^{(T-1)} M_{T-1} + \boldsymbol{\hat{\alpha}_{X(T-1)}^{(T-1)} X_{T-1}}$$

Equation B.9

B.6.1.2 Definition 8

The relationship between each measurement X_t and all previous measurements of the exposure $X_0, X_1, \dots, X_{t-1}$ is confounded by all previous and current measurements of the confounder $M_0, M_1, \dots, M_t$, for $1 \leq t \leq (T-1)$. Thus, we create UR terms e_{Xt} for each measurement of the exposure variable X_t by adjusting for $M_0, M_1, \dots, M_t$, i.e.:

$$X_t = \hat{\gamma}_0^{(t)} + \hat{\gamma}_{M0}^{(t)} M_0 + \hat{\gamma}_{X0}^{(t)} X_0 + \cdots + \hat{\gamma}_{M(t-1)}^{(t)} M_{t-1} + \hat{\gamma}_{X(t-1)}^{(t)} X_{t-1} + \hat{\gamma}_{Mt}^{(t)} M_t + e_{Xt}$$

Equation B.10

and

$$e_{Xt} = -\hat{\gamma}_0^{(t)} - \hat{\gamma}_{M0}^{(t)} M_0 - \hat{\gamma}_{X0}^{(t)} X_0 - \cdots - \hat{\gamma}_{M(t-1)}^{(t)} M_{t-1} - \hat{\gamma}_{X(t-1)}^{(t)} X_{t-1} - \hat{\gamma}_{Mt}^{(i)} M_t + X_t$$

Equation B.11

In this way, e_{Xt} represents the difference between the observed value of X_t and the value of X_t as predicted by all previous measurements $M_0, X_0, M_1, X_1, \dots, M_{t-1}, X_{t-1}, M_t$.

Previous proofs have relied upon the orthogonality of the terms in the composite UR model (i.e. Theorem 1 and Theorem 2 in §B.4.2.1.2 and §B.5.2.1.2, respectively). This necessitates the creation of UR terms e_{Mt} for each measurement of the time-dependent confounding variable M_t, for $1 \leq t \leq (T-1)$. Each e_{Mt} is derived from the OLS regression of M_t on all previous

values of the confounder $M_0, M_1, \dots, M_{t-1}$ and all previous values of the exposure $X_0, X_1, \dots, X_{t-1}$, i.e.:

$$M_t = \hat{\eta}_0^{(t)} + \hat{\eta}_{M0}^{(t)} M_0 + \hat{\eta}_{X0}^{(t)} X_0 + \cdots + \hat{\eta}_{M(t-1)}^{(t)} M_{t-1} + \hat{\eta}_{X(t-1)}^{(t)} X_{t-1} + e_{Mt}$$

Equation B.12

and

$$e_{Mt} = -\hat{\eta}_0^{(t)} - \hat{\eta}_{M0}^{(t)} M_0 - \hat{\eta}_{X0}^{(t)} X_0 - \cdots - \hat{\eta}_{M(t-1)}^{(t)} M_{t-1} - \hat{\eta}_{t-1}^{(t)} X_{t-1} + M_t$$

Equation B.13

These adjustments follow from the DAG in Figure xx, in which it is evident that $X_0, X_1, \dots, X_{t-1}$ confound the relationship between M_t and $M_0, M_1, \dots, M_{t-1}$. Thus, e_{Mt} has a similar interpretation to the original UR terms, in that it represents the part of M_t unexplained by all previous values $M_0, X_0, M_1, X_1, \dots, M_{t-1}, X_{t-1}$.

B.6.1.3 Definition 9

Finally, we represent the composite UR model as a function of the initial value of the exposure X_0 and all subsequent URs for the exposure $e_{X1}, e_{X2}, \dots, e_{Xt}$, and the initial value of the confounder M_0 and all subsequent URs for the confounder $e_{M1}, e_{M2}, \dots, e_{Mt}$:

$$\hat{Y}_{UR}^{(T-1)} = \hat{\lambda}_0^{(T-1)} + \hat{\lambda}_{M0}^{(T-1)} M_0 + \hat{\lambda}_{X0}^{(T-1)} X_0 + \hat{\lambda}_{eM1}^{(T-1)} e_{M1} + \hat{\lambda}_{eX1}^{(T-1)} e_{X1} + \cdots + \hat{\lambda}_{eM(T-1)}^{(T-1)} e_{M(T-1)} + \hat{\lambda}_{eX(T-1)}^{(T-1)} e_{X(T-1)}$$

Equation B.14

B.6.2 Mathematical proofs

B.6.2.1 Covariate orthogonality

Here, we show that: the UR terms for each measurement of the confounder (i.e. $e_{M1}, e_{M2}, \dots, e_{Mt}$) are mutually orthogonal (Lemma 6); the UR terms for each measurement of the exposure (i.e. $e_{X1}, e_{X2}, \dots, e_{Xt}$) are mutually orthogonal (Lemma 7); and, importantly, the UR terms $e_{M1}, e_{M2}, \dots, e_{Mt}$ are orthogonal to $e_{X1}, e_{X2}, \dots, e_{Xt}$ (Lemma 8).

B.6.2.1.1 Lemma 6

$e_{Mt} \perp e_{M1}, e_{M2}, \dots, e_{M(t-1)}$, for $1 \leq t \leq (T-1)$.

Proof: By construction, e_{Mt} represents the residuals from the OLS regression of $M_t \sim 1, M_0, X_0, \dots, M_{t-1}, X_{t-1}$ (Equation B.12). Thus, $e_{Mt} \perp 1, M_0, X_0, \dots, M_{t-1}, X_{t-1}$, which implies $e_{Mt} \cdot 1 = 0$, $e_{Mt} \cdot M_0 = 0$, $e_{Mt} \cdot X_0 = 0$, …, $e_{Mt} \cdot M_{t-1} = 0$, $e_{Mt} \cdot X_{t-1} = 0$.

From this, it follows that $e_{Mt} \perp span(1, M_0, X_0, \dots, M_{t-1}, X_{t-1})$ from Lemma 2.

Because $e_{M1}, e_{M2}, \dots, e_{M(t-1)} \in span(1, M_0, X_0, \dots, M_{t-1}, X_{t-1})$ for $1 \leq t \leq (T-1)$ by construction, we are able to conclude that $e_{Mt} \perp e_{M1}, e_{M2}, \dots, e_{M(t-1)}$. ■

B.6.2.1.2 Lemma 7

$e_{Xt} \perp e_{X1}, e_{X2}, \dots, e_{X(t-1)}$, for $1 \leq t \leq (T-1)$.

Proof: By construction, e_{xt} represents the residuals from the OLS regression of $X_t \sim 1, M_0, X_0, \dots, M_{t-1}, X_{t-1}, M_t$ (Equation B.10). Thus, $e_{Xt} \perp 1, M_0, X_0, \dots, M_{t-1}, X_{t-1}, M_t$, which implies $e_{Xt} \cdot 1 = 0$, $e_{Xt} \cdot M_0 = 0$, $e_{Xt} \cdot X_0 = 0$, …, $e_{Xt} \cdot M_{t-1} = 0$, $e_{Xt} \cdot X_{t-1} = 0$, $e_{Xt} \cdot M_t = 0$.

From this, it follows that $e_{Xt} \perp span(1, M_0, X_0, \dots, M_{t-1}, X_{t-1}, M_t)$ from Lemma 2.

Because $e_{X1}, e_{X2}, \dots, e_{X(t-1)} \in span(1, M_0, X_0, \dots, M_{t-1}, X_{t-1}, M_t)$ for $1 \leq t \leq (T-1)$ by construction, we are able to conclude that $e_{Xt} \perp e_{X1}, e_{X2}, \dots, e_{X(t-1)}$. ■

B.6.2.1.3 Lemma 8

$e_{Xt} \perp e_{Xi}$, for $1 \leq t \leq (T-1)$ and $1 \leq i \leq (T-1)$.

Proof: As established previously, $e_{Xt} \perp span(1, M_0, X_0, \dots, M_{t-1}, X_{t-1}, M_t)$ by Lemma 2, for $1 \leq t \leq (T-1)$. Because $e_{M1}, e_{M2}, \dots, e_{Mt} \in span(1, M_0, X_0, \dots, M_{t-1}, X_{t-1}, M_t)$ by construction, it is evident that $e_{Xt} \perp e_{M1}, e_{M2}, \dots, e_{Mt}$.

Further, $e_{Mi} \perp span(1, M_0, X_0, \dots, M_{i-1}, X_{i-1})$ by Lemma 2, for $1 \leq i \leq (T-1)$. Because $e_{X1}, e_{X2}, \dots, e_{X(i-1)} \in span(1, M_0, X_0, \dots, M_{i-1}, X_{i-1})$ by construction, it is evident that $e_{Mi} \perp e_{X1}, e_{X2}, \dots, e_{X(i-1)}$.

Combining these two results, it follows that $e_{Xt} \perp e_{Mi}$ for $1 \leq t \leq (T-1)$ and $1 \leq i \leq (T-1)$. ■

B.6.2.1.4 Theorem 3

$span(e_{Xt}, e_{Mt}) \perp span(1, M_0, X_0, \dots, e_{M(t-1)}, e_{X(t-1)})$, for $1 \leq t \leq (T-1)$.

Proof: By definition, $e_{Xt} \perp 1, M_0, X_0$. As established in Lemma 7 and Lemma 8, $e_{Xt} \perp e_{X1}, \dots, e_{X(t-1)}, e_{M0}, \dots, e_{M(t-1)}$.

Further, $e_{Mt} \perp 1, M_0, X_0$ by definition, and as established in Lemma 6 and Lemma 8, $e_{Mt} \perp e_{X1}, \dots, e_{X(t-1)}, e_{M0}, \dots, e_{M(t-1)}$.

Thus, by Lemma 2, it follows that $span(e_{Xt}, e_{Mt}) \perp span(1, M_0, X_0, \dots, e_{M(t-1)}, e_{X(t-1)})$. ■

B.6.2.2 Property (i)

$\hat{Y}_S^{(T-1)} = \hat{Y}_{UR}^{(T-1)}$.

Proof: As previously, Property (i) follows from the fact that $\hat{Y}_{UR}^{(t)}$ is a function of the same variables as $\hat{Y}_S^{(t)}$.

By Definition 9, $\hat{Y}_{UR}^{(t)} = f(1, M_0, X_0, e_{M1}, e_{X1}, \dots, e_{Mt}, e_{Xt})$, where $e_{Xt} = f(1, M_0, X_0, \dots, M_t, X_t)$ and $e_{Mt} = f(1, M_0, X_0, \dots, M_{t-1}, X_{t-1}, M_t)$ by Definition 8. Thus, it also holds that

$$\hat{Y}_{UR}^{(t)} = f(1, M_0, X_0, \dots, M_t, X_t)$$

Moreover, by Definition 7,

$$\hat{Y}_S^{(t)} = f(1, M_0, X_0, \dots, M_t, X_t)$$

From this, it follows that $\hat{Y}_S^{(t)} = \hat{Y}_{UR}^{(t)}$ and, consequently, $\hat{Y}_S^{(T-1)} = \hat{Y}_{UR}^{(T-1)}$. ■

B.6.2.3 Property (ii)

$\hat{\alpha}_{X0}^{(0)} = \hat{\lambda}_{X0}^{(T-1)}$.

Proof: By definition, $\hat{Y}_S^{(0)} = \hat{Y}_{UR}^{(0)} = f(1, M_0, X_0)$, and it is trivially true that $\hat{\alpha}_{X0}^{(0)} = \hat{\lambda}_{X0}^{(0)}$.

Because $span(e_{Xt}, e_{Mt}) \perp span(1, M_0, X_0, \dots, e_{M(t-1)}, e_{X(t-1)})$ for $1 \leq t \leq (T-1)$ by Theorem 3, we are able to conclude that $\hat{\lambda}_{X0}^{(0)} = \hat{\lambda}_{X0}^{(1)} = \cdots = \hat{\lambda}_{X0}^{(T-1)}$ by applying Lemma 1.

Therefore, $\hat{\alpha}_{X0}^{(0)} = \hat{\lambda}_{X0}^{(T-1)}$. ■ [33]

B.6.2.4 Property (iii)

$\hat{\alpha}_{Xt}^{(t)} = \hat{\lambda}_{eXt}^{(T-1)}$.

Proof: Consider the UR model:

$$\hat{Y}_{UR}^{(t)} = \hat{\lambda}_0^{(t)} + \hat{\lambda}_{M0}^{(t)} M_0 + \hat{\lambda}_{X0}^{(t)} X_0 + \hat{\lambda}_{eM1}^{(t)} e_{M1} + \hat{\lambda}_{eX1}^{(t)} e_{X1} + \cdots + \hat{\lambda}_{eMt}^{(t)} e_{Mt} + \boldsymbol{\hat{\lambda}_{eXt}^{(t)}} e_{Xt}$$

By substituting the expansions for e_{Xt} (Equation B.11) and e_{Mt} (Equation B.13) into this equation and rearranging, we produce:

$$\begin{aligned}
\hat{Y}_{UR}^{(t)} = {} & \hat{\lambda}_0^{(t)} + \hat{\lambda}_{M0}^{(t)} M_0 + \hat{\lambda}_{X0}^{(t)} X_0 + \hat{\lambda}_{eM1}^{(t)} \left[-\hat{\eta}_0^{(1)} - \hat{\eta}_{X0}^{(1)} X_0 - \hat{\eta}_{M0}^{(1)} M_0 + M_1 \right] + \\
& \hat{\lambda}_{eX1}^{(t)} \left[-\hat{\gamma}_0^{(1)} - \hat{\gamma}_{X0}^{(1)} X_0 + X_1 - \hat{\gamma}_{M0}^{(1)} M_0 - \hat{\gamma}_{M1}^{(1)} M_1 \right] + \cdots + \hat{\lambda}_{eMt}^{(t)} \left[-\hat{\eta}_0^{(t)} - \hat{\eta}_{X0}^{(t)} X_0 - \right. \\
& \left. \cdots - \hat{\eta}_{t-1}^{(t)} X_{t-1} - \hat{\eta}_{M0}^{(t)} M_0 - \cdots - \hat{\eta}_{M(t-1)}^{(t)} M_{t-1} + M_t \right] + \hat{\lambda}_{eXt}^{(t)} \left[-\hat{\gamma}_0^{(t)} - \hat{\gamma}_{X0}^{(t)} X_0 - \right. \\
& \left. \cdots - \hat{\gamma}_{t-1}^{(t)} X_{t-1} + X_t - \hat{\gamma}_{M0}^{(t)} M_0 - \cdots - \hat{\gamma}_{Mt}^{(t)} M_t \right] \\
= {} & \left[\hat{\lambda}_0^{(t)} - \hat{\lambda}_{eM1}^{(t)} \hat{\eta}_0^{(1)} - \hat{\lambda}_{eX1}^{(t)} \hat{\gamma}_0^{(1)} - \cdots - \hat{\lambda}_{eMt}^{(t)} \hat{\eta}_0^{(t)} - \hat{\lambda}_{eXt}^{(t)} \hat{\gamma}_0^{(t)} \right] + \left[\hat{\lambda}_{M0}^{(t)} - \hat{\lambda}_{eM1}^{(t)} \hat{\eta}_{M0}^{(1)} - \right. \\
& \left. \hat{\lambda}_{eX1}^{(t)} \hat{\gamma}_{M0}^{(1)} - \cdots - \hat{\lambda}_{eMt}^{(t)} \hat{\eta}_{M0}^{(t)} - \hat{\lambda}_{eXt}^{(t)} \hat{\gamma}_{M0}^{(t)} \right] M_0 + \left[\hat{\lambda}_{X0}^{(t)} - \hat{\lambda}_{eM1}^{(t)} \hat{\eta}_{X0}^{(1)} - \hat{\lambda}_{eX1}^{(t)} \hat{\gamma}_{X0}^{(1)} - \cdots - \right. \\
& \left. \hat{\lambda}_{eMt}^{(t)} \hat{\eta}_{X0}^{(t)} - \hat{\lambda}_{eXt}^{(t)} \hat{\gamma}_{X0}^{(t)} \right] X_0 + \cdots + \left[\hat{\lambda}_{eMt}^{(t)} - \hat{\lambda}_{eXt}^{(t)} \hat{\gamma}_{Mt}^{(t)} \right] M_t + \left[\boldsymbol{\hat{\lambda}_{eXt}^{(t)}} \right] X_t
\end{aligned}$$

Having established that $\hat{Y}_S^{(t)} = \hat{Y}_{UR}^{(t)}$ (i.e. Property (i)) because they are functions of the same covariates, it follows that the estimated coefficients for those covariates must themselves be equal. Specifically, we see that the coefficient for X_t will always equal the coefficient for e_{Xt}, i.e. $\hat{\alpha}_{Xt}^{(t)} = \hat{\lambda}_{eXt}^{(t)}$.

Finally, using the fact that $e_{Xt} \perp span(1, M_0, X_0, e_{M1}, e_{X1}, \dots, e_{M(t-1)}, e_{X(t-1)}, e_{Mt})$, we apply Lemma 1 and conclude that $\hat{\lambda}_{eXt}^{(0)} = \hat{\lambda}_{eXt}^{(1)} = \cdots = \hat{\lambda}_{eXt}^{(T-1)}$, from which it follows that $\hat{\alpha}_{Xt}^{(t)} = \hat{\lambda}_{eXt}^{(T-1)}$. ■

---

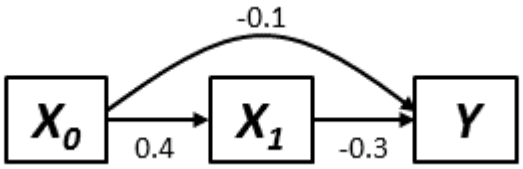
[33] Although no causal meaning/significance can be attributed to the intercept term or the coefficients of the UR terms for the confounder $e_{M1}, \dots, e_{M(T-1)}$, the logic applied in this proof may be easily extended to show that $\hat{\alpha}_0^{(0)} = \hat{\lambda}_0^{(T-1)}$ and $\hat{\alpha}_{M1}^{(0)} = \hat{\lambda}_{eM1}^{(T-1)}$, …, $\hat{\alpha}_{M(T-1)}^{(0)} = \hat{\lambda}_{eM(T-1)}^{(T-1)}$, respectively.

B.7 Artefactual standard error reduction using UR models: Simulation details and code

B.7.1 Directed acyclic graph (DAG)

Figure B.1 depicts the DAG from which multivariate normal data were simulated in order to demonstrate the artefactual reduction in standard errors that results from using UR models.

Figure B.1 Directed acyclic graph from which multivariate normal data were simulated to demonstrate standard error reduction in UR models



Numbers represent standardised path coefficients.

The DAG in Figure B.1 implies the correlation matrix given in Table B.2.

Table B.2 Correlation matrix implied by the DAG in Figure B.1

	X_0	X_1	Y
X_0	1.00	-	-
X_1	0.40	1.00	-
Y	-0.22	-0.34	1.00

B.7.2 Population parameters

The population mean and standard deviation (SD) specified in the simulation are provided in Table B.3.

Table B.3 Population mean and standard deviation (SD) used in the data simulation based on the DAG in Figure B.1

	Mean	SD
X_0	10.00	2.50
X_1	15.00	3.75
Y	20.00	5.00

B.7.3 Annotated R code

```
################################################################################
# UR MODELS: BOOTSTRAPPED STANDARD ERRORS ######################################
################################################################################

# This code demonstrates the artificial reduction in standard errors (SEs)
# in UR models compared to standard regression models

# The scenario considered is for two longitudinal measures of a time-varying
# exposure (X0, X1) and one outcome (Y) [all variables continuous]

################################################################################
## (1) SET UP ----------------------------------------------------------------
```

```

# Load required packages - data simulation
require(Matrix); require(matrixcalc); require(MASS)
#require(devtools); devtools::install_github("jtextor/dagitty/r") ## update regularly
require(dagitty)

# Load required packages - plots
require(ggplot2); require(gridExtra); require(extrafont); require(Hmisc); library(tidyr)
font_import(pattern="[C/c]alibri"); loadfonts(device="win") ## use fonttable() to see options

### (a) Functions ------------------------------------------------------------

#### (i) Covar function ------------------------------------------------------

# This function converts SDs and pairwise correlations to a covariace matrix

Covar <- function(n=2,SD=data.frame(1,1),c.vec=data.frame(0.5)) {
  check   <- n-length(SD)
  if (check !=0) stop("Incorrect SD specifications!")
  check   <- (n*(n-1)/2)-length(c.vec)
  if (check !=0) stop("Incorrect correlation specifications!")
  Cor     <- NULL
  for (i in 1:(n+1)) {
    Row <- NULL
    for (j in 1:(n+1)) {
      if (i==j) Element <- 1
      else if (i<j) Element <- c.vec[((i-1)*(2*n-i)/2)+(j-i)]
      else if (i>j) Element <- c.vec[((j-1)*(2*n-j)/2)+(i-j)]
      Row <- c(Row,Element)
    }
    Cor <- rbind(Cor,Row)
  } # cov(i,j) = cor(i,j)*sd(i)*sd(j)
  Cov <- matrix(nrow=n,ncol=n)
  for (i in 1:n) { for (j in 1:n) { Cov[i,j] <- Cor[i,j]*SD[i]*SD[j] }}
  Cov <- as.matrix(forceSymmetric(Cov))
  if (!is.positive.definite(Cov)) {
    print("Warning: covariance matrix made Positive Definite")
    Cov <- as.matrix(nearPD(Cov)$mat) }
  return(Cov)
}

#### (ii) Present function ---------------------------------------------------

# This function presents model summary (point estimates and 95% CIs)

Present <- function(mod) {
  Est     <- summary(mod)$coefficients[,1]
  CI95    <- confint(mod)
  coeffs  <- cbind(Est,CI95); rownames(coeffs)[1] <- "Intercept"
  Trim    <- round(coeffs,3)
  Tidy    <- data.frame(apply(Trim,1,function(x) {paste0(x[1]," (",x[2],", ",x[3],")")}))
  names(Tidy)[1] <- "Model Summary"
  return(Tidy) }

#### (iii) Data summary ------------------------------------------------------

# This function produces summary statistics (mean and +/- sd)

data_summary <- function(x) {
  m <- mean(x)
  ymin <- m - sd(x)
  ymax <- m + sd(x)
  return(c(y=m, ymin=ymin, ymax=ymax))
}

################################################################################
## (2) DATA SIMULATION ---------------------------------------------------------

### (a) Define DAG from which data will be simulated ---------------------------

dag1 <- dagitty('dag{
                X0 [pos = "0.2, 0.2"]
                X1 [pos = "0.6, 0.2"]
                Y [pos = "1, 1"]
                X0 -> X1 [beta = 0.4]
                X0 -> Y [beta = -0.1]
                X1 -> Y [beta = -0.3]
                }')
#plot(dag1)
mod <- lm(Y ~ X0 + X1, data = simulateSEM(dag1, empirical = TRUE))
#Present(mod)

### (b) Calculate covariance matrix based on DAG -------------------------------

MyData  <- simulateSEM(dag1, empirical = TRUE)  ## (standardised data)
Names   <- c("X0","X1","Y")
SetCor  <- cor(MyData); Corr <- SetCor[lower.tri(SetCor)]
N       <- 1000
X0.mu   <- 10
X1.mu   <- 15
Y.mu    <- 20
Mu      <- c(X0.mu, X1.mu, Y.mu)
X0.sd   <- X0.mu/4
X1.sd   <- X1.mu/4
Y.sd    <- Y.mu/4
SD      <- c(X0.sd, X1.sd, Y.sd)
MyCov   <- Covar(3, SD, Corr)

### (c) Simulation -------------------------------------------------------------

# Set storage for SEs for X0
seX0.reg <- NULL  # standard regression models
seX0.UR <- NULL  # UR models (as reported)
```

```
seX0.UR.boot <- NULL  # UR models (bootstrapped)

# Set storage for SEs for X1/e1
seX1.reg <- NULL  # standard regression models
see1.UR <- NULL  # UR models (as reported)
see1.UR.boot <- NULL  # UR models (bootstrapped)

# Set seed
set.seed(23)

for (i in 1:1000) {

  # Simulate N observations
  MyData  <- data.frame(mvrnorm(N, Mu, MyCov, empirical = FALSE))
  names(MyData) <- Names

  # Create standard regression model for X0 and save SE
  modX0 <- lm(Y ~ X0, data = MyData)
  seX0.reg <- c(seX0.reg, summary(modX0)$coefficients[2, 2])

  # Create standard regression model for X1 and save SE
  modX1 <- lm(Y ~ X0 + X1, data = MyData)
  seX1.reg <- c(seX1.reg, summary(modX1)$coefficients[3, 2])

  # Create UR term
  modX1.resid <- lm(X1 ~ X0, data = MyData)
  MyData$e1 <- modX1.resid$residuals

  # Create UR model and save SEs for coeffs
  modUR <- lm(Y ~ X0 + e1, data = MyData)
  seX0.UR <- c(seX0.UR, summary(modUR)$coefficients[2, 2])
  see1.UR <- c(see1.UR, summary(modUR)$coefficients[3, 2])

  # Use bootstrapping to create distribution of coefficients for UR model
  coeffX0.UR.boot <- NULL  # set storage for coeffs for X0 from UR model
  coeffe1.UR.boot <- NULL  # set storage for coeffs for e1 from UR model

  for (j in 1:1000) {

    # Select random sample with replacement from MyData
    select <- sample(c(1:1000), 1000, replace = TRUE)
    MyData.boot <- MyData[select, ]

    # Create UR term
    modX1.resid.boot <- lm(X1 ~ X0, data = MyData.boot)
    MyData.boot$e1 <- modX1.resid.boot$residuals

    # create UR models and save coeffs
    modUR.boot <- lm(Y ~ X0 + e1, data = MyData.boot)
    coeffX0.UR.boot <- c(coeffX0.UR.boot, summary(modUR.boot)$coefficients[2, 1])
    coeffe1.UR.boot <- c(coeffe1.UR.boot, summary(modUR.boot)$coefficients[3, 1])


  }

  # calculate SES for UR model as standard deviation of distribution of coefficients
  seX0.UR.boot <- c(seX0.UR.boot, sd(coeffX0.UR.boot))
  see1.UR.boot <- c(see1.UR.boot, sd(coeffe1.UR.boot))

  # Display progress of simulation
  cat('\r', paste(round((i / 1000 * 100), 2),
                  "% done of simulation    ", sep = " "))


}

# Bind and export datasets
SimDataX0 <- data.frame(seX0.reg, seX0.UR, seX0.UR.boot)
SimDataX1 <- data.frame(seX1.reg, see1.UR, see1.UR.boot)
write.csv(SimDataX0, file = "SE bootstrap - X0.csv", row.names = FALSE)
write.csv(SimDataX1, file = "SE bootstrap - X1.csv", row.names = FALSE)

################################################################################
# (3) PLOTS COMPARING SEs ------------------------------------------------------

# Import datasets
SimDataX0 <- read.csv("./SE bootstrap - X0.csv", header = TRUE)
SimDataX1 <- read.csv("./SE bootstrap - X1.csv", header = TRUE)

# Label data with exposure variable
SimDataX0$Exp <- "X0"
SimDataX1$Exp <- "X1"

# Rename variable names
names(SimDataX0) <- c("se.reg", "se.UR", "se.UR.boot", "Exp")
names(SimDataX1) <- c("se.reg", "se.UR", "se.UR.boot", "Exp")

# Create combined long format data frame
DataFrame <- rbind(SimDataX0, SimDataX1)
DataFrame.long <- gather(data = DataFrame, key = Model, value = SE, 1:3)
#str(DataFrame.long)
DataFrame.long[, c("Exp", "Model")] <- data.frame(apply(DataFrame.long[c("Exp", "Model")], 2,
as.factor))

# Violin plot
plot <- ggplot(DataFrame.long, aes(x = Model, y = SE,
                                   group = interaction(Model, Exp))) +
  theme_bw() +
  theme(axis.line = element_line(size = 1, colour = "black"),
        panel.border = element_blank(),
        panel.grid.minor = element_blank(),
        text = element_text(size = 13, family = "Calibri Light"),
        axis.text.x = element_text(size = 13),
```

```
        axis.text.y = element_text(size = 11)) +
  geom_violin(size = 1.2, trim = TRUE,
              position = position_dodge(0),
              aes(fill = Exp, colour = Exp)) +
  stat_summary(fun.data = data_summary, color = "grey20", size = 0.7,
               position = position_dodge(0)) +
  scale_fill_manual(name = "Exposure",
                    breaks = c("X0", "X1"),
                    labels = c("X0", "X1"),
                    values = c("orchid", "slateblue1")) +
  scale_colour_manual(name = "Exposure",
                      breaks = c("X0", "X1"),
                      labels = c("X0", "X1"),
                      values = c("orchid4", "slateblue4")) +
  scale_x_discrete(name = "",
                   limits = c("se.UR.boot", "se.UR", "se.reg"),
                   labels = c("UR models \n(bootstrapped)",
                              "UR models \n(reported)",
                              "Standard \nregression \nmodels")) +
  scale_y_continuous(name = "Standard error",
                     limits = c(0.03, 0.08),
                     breaks = seq(from = 0.03, to = 0.08, by = 0.01),
                     expand = expand_scale(mult = c(0.02, 0.02))) +
  coord_flip()
#plot

# Export as png
ggsave("UR models - SE bootstrap - combined.png", plot = plot,
       width = 8, height = 5, units = "in", dpi = 800)
```

Appendix C
Microsimulation modelling

C.1 Introduction

This appendix contains additional material relating to Chapter 6. In particular, it contains additional details, methods, results, and annotated R relating to the simulation described in Section 6.4.

C.2 Simulated example

This section contains all details relating to the simulation described in Chapter 6, Section 6.4; the simulation is based on the example scenario described in Figure 6.1.

C.2.1 Simulation of a population according to the true data-generating process

In this subsection, we provide details relating to the simulation of a population according to the true data-generating process, which is described in Section 6.4.1. This includes simulation of both 'natural' (§C.2.1.1) and 'counterfactual' histories (§C.2.1.2).

C.2.1.1 Natural history

For the 'natural history' simulation, we provide the simulation parameters (§C.2.1.1.1), characteristics of the resulting simulated population (§C.2.1.1.2), a comparison of the simulated population with Health Survey for England (HSE) statistics (§C.2.1.1.3), and all annotated R code relating to these simulations (§C.2.1.1.4).

C.2.1.1.1 Simulation parameters

Parameters describing the distribution of sex, obesity, and diabetes at baseline (i.e. $t = 0$) are given in Table C.1.

Parameters describing the evolution of the baseline population (i.e. the transition parameters) for all subsequent time points (i.e. time t, for $1 \leq t \leq 10$) are given in Table C.2.

Table C.1 Parameters describing the joint distribution of sex, obesity, and diabetes in the baseline population (i.e. time $t = 0$)

Status	Covariate(s)	Probability
Male	n/a	0.521
Obese	Female	0.490
	Male	0.580
Diabetic	Female, non-obese	0.010
	Female, obese	0.030
	Male, non-obese	0.017
	Male, obese	0.037

The probability of the 'complement' states (i.e. Female, Non-obese, and Non-diabetic, respectively) are equal to 1 minus the stated probability. For example, the probability of being female is equal to 1 – 0.521 = 0.479.

Table C.2 Transition parameters describing the evolution of the baseline population (i.e. time t, for $1 \leq t \leq 10$)

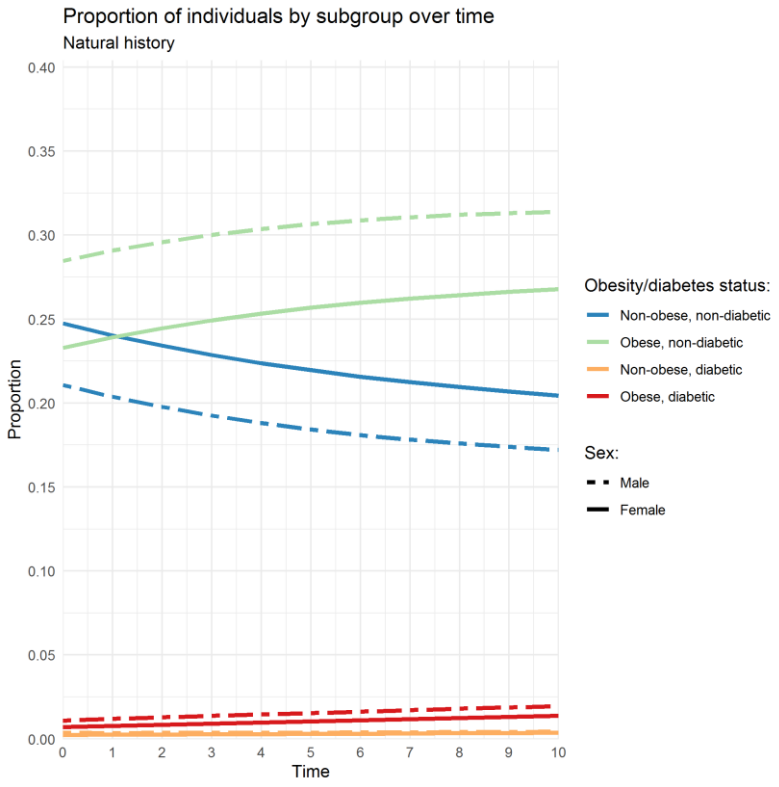
Updated status	Current status	Current covariates	Probability
Obese	Non-obese	Female, non-diabetic	0.07500
		Female, diabetic	0.10500
		Male, non-diabetic	0.10000
		Male, diabetic	0.13000
	Obese	Female, non-diabetic	0.95000
		Female, diabetic	0.97000
		Male, non-diabetic	0.95000
		Male, diabetic	0.97000
Diabetic	Non-diabetic	Female, non-obese	0.00060
		Female, obese	0.00260
		Male, non-obese	0.00065
		Male, obese	0.00265
	Diabetic	Female, non-obese	1.00000
		Female, obese	1.00000
		Male, non-obese	1.00000
		Male, obese	1.00000

The probability of the 'complement' updated states (i.e. Non-obese, and Non-diabetic, respectively) are equal to 1 minus the stated probability. For example, the probability of having the updated state of obesity at time t, given an individual is currently non-obese, female, and non-diabetic (i.e. line 1 of the table) is equal to 1 – 0.075 = 0.925.

C.2.1.1.2 Characteristics of the simulated population

The proportion of individuals with each combination of characteristics for each time point in the simulated population are displayed in Figure C.1.

Figure C.1 Proportion of individuals in the simulated population with each combination of sex, obesity status, and diabetes status at every time point



The proportion of obese individuals steadily increases over the course of the simulation, while the proportion of non-obese individuals decreases. Because diabetes has such a low overall prevalence, the proportion of obese individuals without diabetes represents a much larger subgroup than those with diabetes, though both increase throughout the simulation. As simulated, males represent a higher proportion of obese individuals – both diabetic and non-diabetic – compared to females.

The probabilities of becoming and remaining obese for every time point in the simulated population are displayed in Figure C.2; the probabilities of becoming and remaining diabetic for every time point in the simulated population are displayed in Figure C.3. These probabilities are consistent with the parameters specified in Table C.2, confirming that the simulation performed as expected.

Figure C.2 Probabilities of becoming and remaining obese in the simulated population at every time point

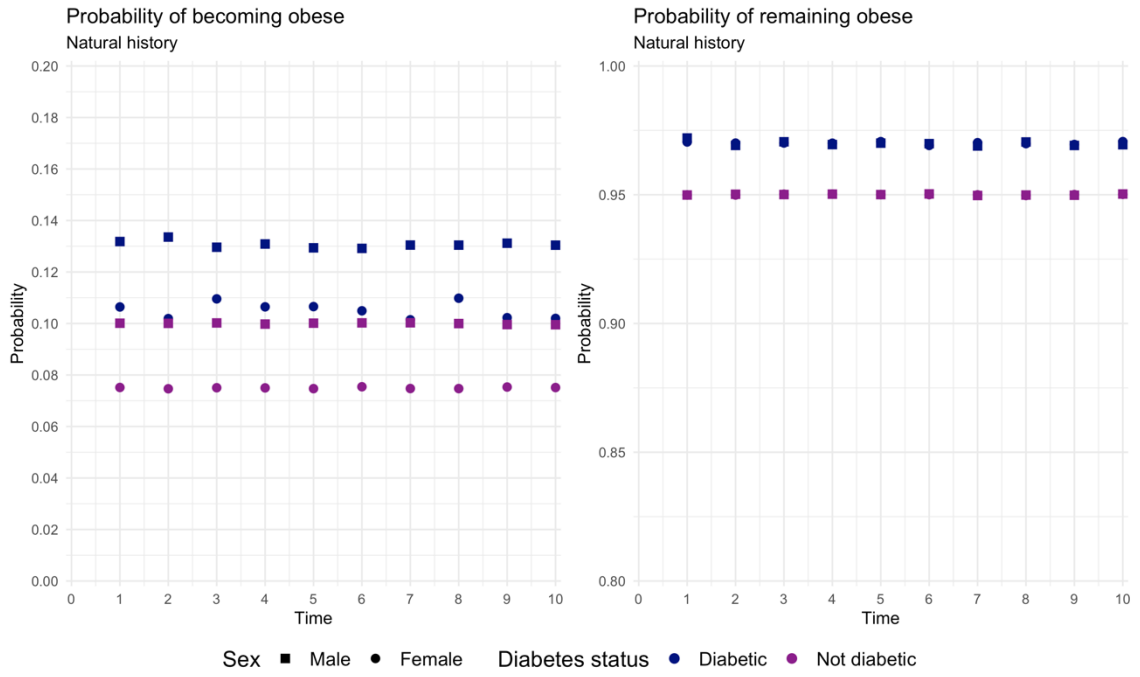
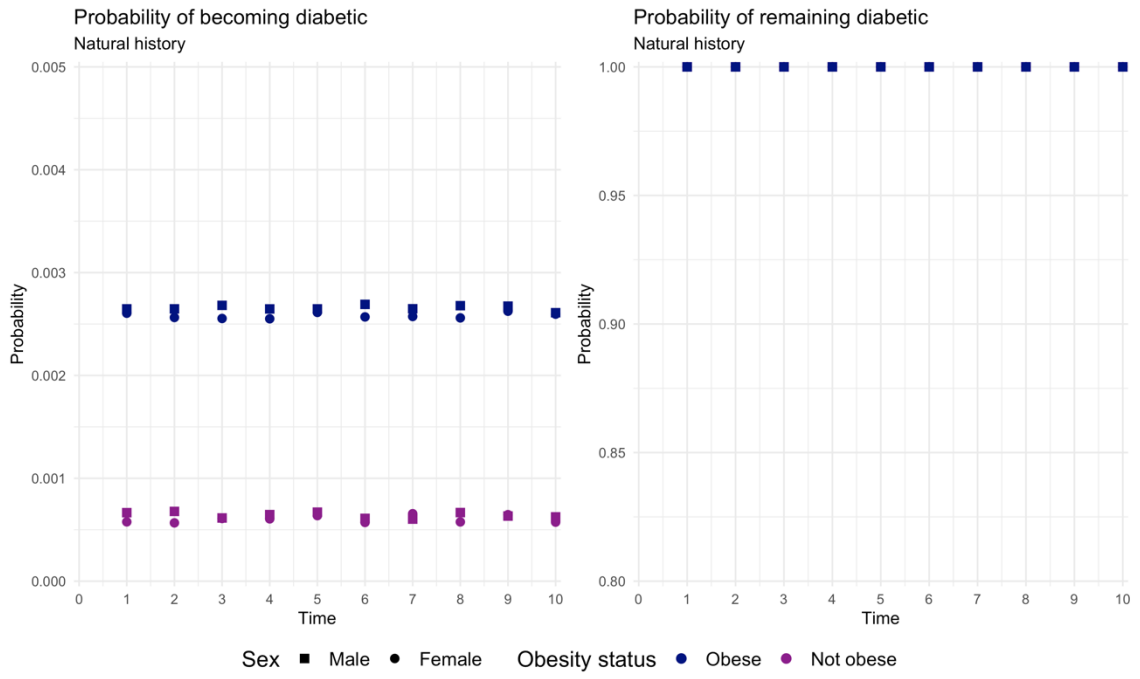


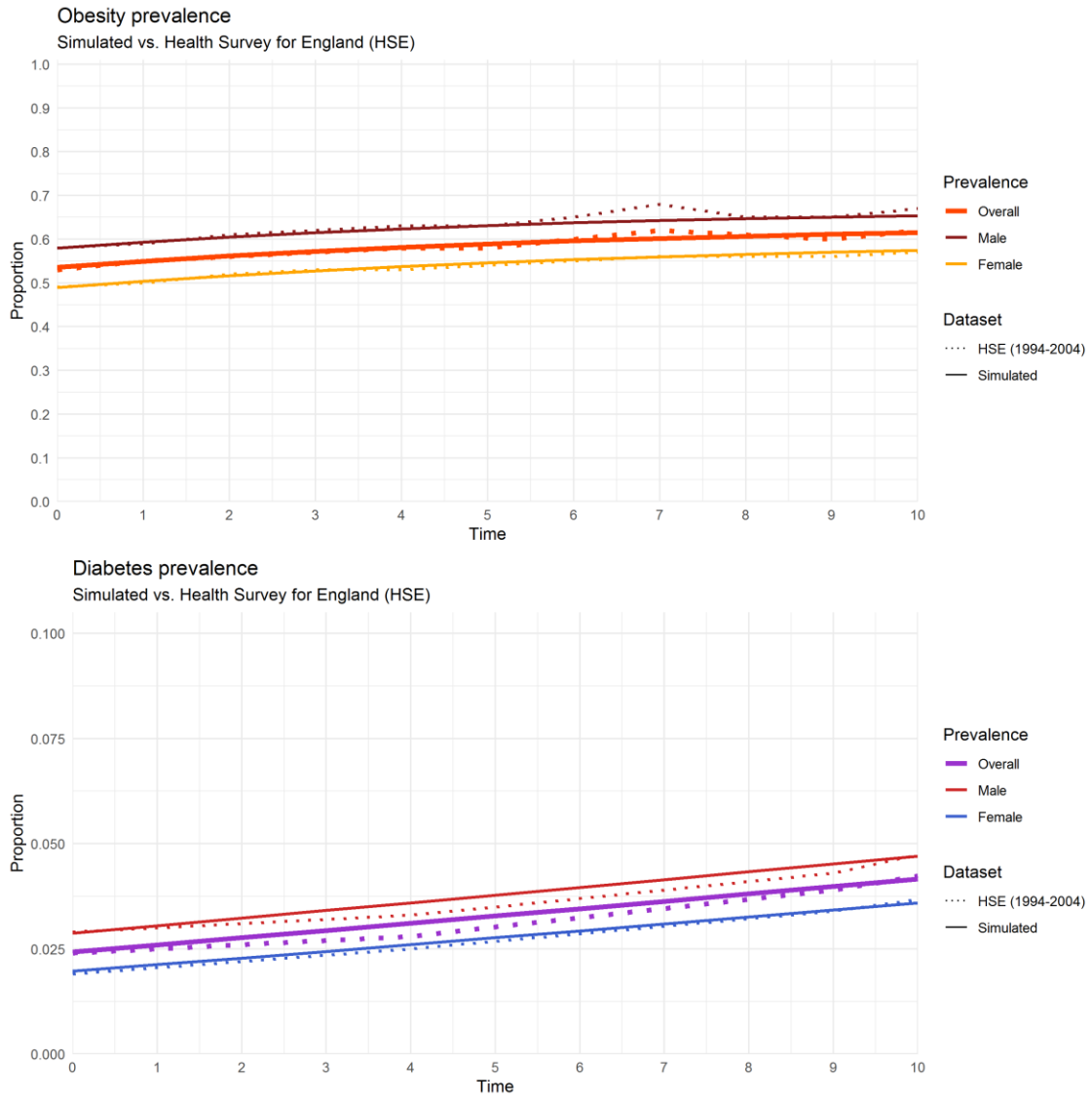
Figure C.3 Probabilities of becoming and remaining diabetic in the simulated population at every time point



C.2.1.1.3 Comparison of the simulated population with Health Survey for England (HSE) statistics

Comparison of simulated population data with Health Survey for England (HSE) statistics (191, 192) are depicted in Figure C.4.

Figure C.4 Obesity and diabetes prevalence in the simulated population compared to the Health Survey for England (HSE, years 1994-2004)



Obesity prevalence reported by the HSE includes both overweight and obese individuals.

Between 1994 and 2004, overall obesity prevalence increased from 53% (49% for females and 58% for males) to 62% (57% for females and 67% for males), as reported by the HSE (191, 192). This was approximated by the simulated population, in which overall obesity increased from 53.57% at baseline (49.00% for females and 57.97% for males) to 61.53% at time 10 (57.49% for females and 65.41% for males).

Similarly, between 1994 and 2004, overall diabetes prevalence increased from 2.40% (1.90% for females and 2.90% for males) to 4.23% (3.67% for females and 4.73% for males) (191, 192). In the simulated population, overall diabetes prevalence increased from 2.43% at baseline (1.97% for females and 2.87% for males) to 4.16% at time 10 (3.59% for females and 4.70% for males).

C.2.1.1.4 Annotated R code

```
1  ##############################################################################
```

```
# POPULATION SIMULATION ###########################################################
################################################################################

# This code generates an artificial longitudinal population using
# a time-based, discrete time microsimulation model

# Simulated individuals have the following 3 attributes:
# Sex (time-fixed): 0 = female, 1 = male
# Obesity (time-varying): 0 = nonobese, 1 = obese
# Diabetes (time-varying): 0 = nondiabetic, 1 = diabetic

################################################################################
## (1) SET UP ------------------------------------------------------------------

# Clear workspace
rm(list = ls())

# Load all required packages
library(readxl); library(stringr); library(Hmisc); library(plyr); library(scales)
library(ggplot2); library(gridExtra); library(HydeNet); library(data.table)

### (a) Population/simulation parameters ---------------------------------------

# Define population parameters
N.i.pop <- 5000000  # number of individuals
N.t.pop <- 11  # number of time points (including baseline)
Time.pop <- as.vector(seq(from = 0, to = (N.t.pop - 1), by = 1),
                      mode = "integer")  # time vector

### (b) Tables to store population data ----------------------------------------

# Create empty matrices to store individual-level population data
# Each row represents 1 individual (N.i.pop rows)
# Each column represents 1 time point (N.t.pop columns)
Sex.pop <- matrix(nrow = N.i.pop, ncol = 1,
                  dimnames = list(paste0("ind", 1:N.i.pop), "Sex"))
Obes.pop <- matrix(nrow = N.i.pop, ncol = N.t.pop,
                   dimnames = list(paste0("ind", 1:N.i.pop),
                                   paste0("O.t", Time.pop)))
Diab.pop <- matrix(nrow = N.i.pop, ncol = N.t.pop,
                   dimnames = list(paste0("ind", 1:N.i.pop),
                                   paste0("D.t", Time.pop)))

### (c) Tables to store summary data -------------------------------------------

# Create empty cross-sectional frequency table
Frequency.cs.pop <- data.frame(Time = numeric(), Sex = numeric(),
                               O.t = numeric(), D.t = numeric(),
                               freq = numeric())

# Create empty tables to record obesity & diabetes prevalence from population
# (overall and disaggregated by sex)
Obes.prev.pop <- data.frame(Time = numeric(), Subgroup = factor(),
                            prev = numeric())
Diab.prev.pop <- data.frame(Time = numeric(), Subgroup = factor(),
                            prev = numeric())

# Create empty tables to record cross-sectional conditional probabilities of obesity & diabetes
CProbability.Obes.cs <- data.frame(Time = numeric(), Sex = factor(),
                                   O.t = factor(), prob = numeric())
CProbability.Diab.cs <- data.frame(Time = numeric(), Sex = factor(),
                                   O.t = factor(), D.t = factor(),
                                   prob = numeric())

# Create empty tables to record cross-time conditional probabilities of obesity & diabetes
CProbability.Obes.ct <- data.frame(Time = numeric(), Sex = factor(),
                                   O.tminus1 = factor(), D.tminus1 = factor(),
                                   O.t = factor(), prob = numeric())
CProbability.Diab.ct <- data.frame(Time = numeric(), Sex = factor(),
                                   D.tminus1 = factor(), O.t = factor(),
                                   D.t = factor(), prob = numeric())

### (d) Functions --------------------------------------------------------------

#### (i) SampleV function ------------------------------------------------------

# samplev() function
# efficient implementation of the rMultinom() function of the Hmisc package
# from Krijkamp et al (2018) (185)
samplev <- function(probs, m) {
  d <- dim(probs)  # (dimensions of probability matrix)
  n <- d[1]  # (number of rows, i.e. individuals)
  k <- d[2]  # (number of columns, i.e. states)
  lev <- dimnames(probs)[[2]]  # (names of columns, i.e. state values)
  if (!length(lev))
    lev <- 1:k
  ran <- matrix(lev[1], ncol = m, nrow = n)
  U <- t(probs)
  for(i in 2:k) {
    U[i, ] <- U[i, ] + U[i - 1, ]
  }
  if (any((U[k, ] - 1) > 1e-05))
    stop("error in multinom: probabilities do not sum to 1")

  for (j in 1:m) {
    un <- rep(runif(n), rep(k, n))
    ran[, j] <- lev[1 + colSums(un > U)]
  }
  ran
}

#### (ii) Calculate prevalence proportions -------------------------------------
```

```
# Calculate Obesity prevalence
# args: group = subgroup, freqtable = cross-sectional frequency table (numeric)
# returns single number (prevalence)
CalculatePrevObesityT <- function (group, freqtable) {

  if (group == "overall") {

    prevalence <- sum(subset(freqtable, Time == (t-1) & O.t == 1)$freq) /
      N.i.pop
    return(prevalence)

  } else if (group == "female") {

    prevalence <- sum(subset(freqtable, Time == (t-1) & Sex == 0 & O.t == 1)$freq) /
      sum(subset(freqtable, Time == (t-1) & Sex == 0)$freq)
    return(prevalence)

  } else if (group == "male") {

    prevalence <- sum(subset(freqtable, Time == (t-1) & Sex == 1 & O.t == 1)$freq) /
      sum(subset(freqtable, Time == (t-1) & Sex == 1)$freq)
    return(prevalence)

  }

}  # (close funciton loop)

# Calculate Diabetes prevalence
# args: group = subgroup, freqtable = cross-sectional frequency table (numeric)
# returns single number (prevalence)
CalculatePrevDiabetesT <- function (group, freqtable) {

  if (group == "overall") {

    prevalence <- sum(subset(freqtable, Time == (t-1) & D.t == 1)$freq) /
      N.i.pop
    return(prevalence)

  } else if (group == "female") {

    prevalence <- sum(subset(freqtable, Time == (t-1) & Sex == 0 & D.t == 1)$freq) /
      sum(subset(freqtable, Time == (t-1) & Sex == 0)$freq)
    return(prevalence)

  } else if (group == "male") {

    prevalence <- sum(subset(freqtable, Time == (t-1) & Sex == 1 & D.t == 1)$freq) /
      sum(subset(freqtable, Time == (t-1) & Sex == 1)$freq)
    return(prevalence)

  }

}  # (close function loop)

#### (iii) Calculate conditional probabilities ---------------------------------

# Calculate conditional probability table at time t
# args: dv = dependent variable, iv = independent variable(s), dataset = data frame (factorised)
# returns conditional probability table (cprob.t)
CalculateCPT <- function(dv, iv, dataset) {

  # Define formula for use in cpt function (from HydeNet package)
  formula <- as.formula(paste(dv, paste(iv, collapse = " + "), sep = " ~ "))

  # Create conditional probability table
  cprob.t <- cbind(Time = (t-1), am_adt(cpt(formula, data = dataset)))

  return(cprob.t)

}

# Function for converting multidimensional arrays to tables
# (from https://github.com/Rdatatable/data.table/issues/1418)
am_adt <- function(inarray) {
  if (!is.array(inarray)) stop("input must be an array")
  dims <- dim(inarray)
  if (is.null(dimnames(inarray))) {
    inarray <- provideDimnames(inarray, base = list(as.character(seq_len(max(dims)))))
  }
  FT <- if (any(class(inarray) %in% "ftable")) inarray else ftable(inarray)
  out <- data.table(as.table(ftable(FT)))
  nam <- names(out)[seq_along(dims)]
  setorderv(out[, (nam) := lapply(.SD, type.convert), .SDcols = nam], nam)[]
}

################################################################################
## (2) SIMULATION --------------------------------------------------------------

# Set seed
set.seed(23)

### (a) Define (conditional) probabilities at baseline -------------------------

#### (i) Sex -------------------------------------------------------------------

p.male <- 0.51  # baseline P(Sex = 1)

#### (ii) Obesity --------------------------------------------------------------

# Function to calculate baseline P(Obesity = 1 | Sex)

CalculateProbObesity0 <- function(Sex) {

```

```r
  p.obes.0 <- 0.49 + 0.09*Sex
  return(p.obes.0)

}

#### (iii) Diabetes ------------------------------------------------------------

# Function to calculate baseline P(Diabetes = 1 | Sex, Obesity)

CalculateProbDiabetes0 <- function(Sex, Obes) {

  p.diab.0 <- 0.01 + 0.007*Sex + 0.02*Obes
  return(p.diab.0)

}

### (a) Define conditional probabilities at time t -----------------------------

#### (i) Obesity ---------------------------------------------------------------

# Function to calculate P(Obesity = 1 | Sex, Prev obesity, Prev diabetes) at time t

CalculateProbObesityT <- function(Sex, PrevObes, PrevDiab) {

  # Incident probability (PrevObes = 0): 0.075 + 0.025*Sex + 0.03*PrevDiab
  # Prevalent probability (PrevObes = 1): 0.95 + 0.02*PrevDiab

  p.obes.t <- 0.075 + 0.025*Sex + 0.03*PrevDiab +
    PrevObes*(0.875 - 0.025*Sex - 0.01*PrevDiab)
  return(p.obes.t)

}

#### (ii) Diabetes -------------------------------------------------------------

# Function to calculate P(Diabetes = 1 | Sex, Obesity, Prev diabetes) at time t

CalculateProbDiabetesT <- function(Sex, PrevDiab, Obes) {

  # Incident probability: 0.0006 + 0.00005*Sex + 0.002*Obes
  # Prevalent probability: 1

  p.diab.t <- 0.0006 + 0.00005*Sex + 0.002*Obes +
    PrevDiab*(0.9994 - 0.00005*Sex - 0.002*Obes)
  return(p.diab.t)

}

### (b) Simulation -------------------------------------------------------------

v <- Sys.time()  # record start time of simulation

# (1) Loop through time points
for (t in 1:N.t.pop) {

  ## Assign baseline characteristics & record summary data
  if (t == 1) {

    # Assign baseline characteristics ----------------------------------------

    # (1) Sex
    p.sex <- cbind(rep(1 - p.male, N.i.pop), rep(p.male, N.i.pop))
    Sex.pop[, 1] <- samplev(probs = p.sex, m = 1)
    Sex.pop[, 1] <- Sex.pop[, 1] - 1  # (factor levels should be 0 and 1)

    # (2) Obesity
    p.obes.0 <- cbind(1 - CalculateProbObesity0(Sex = Sex.pop[, 1]),
                      CalculateProbObesity0(Sex = Sex.pop[, 1]))
    Obes.pop[, 1] <- samplev(probs = p.obes.0, m = 1)
    Obes.pop[, 1] <- Obes.pop[, 1] - 1  # (factor levels should be 0 and 1)

    # (3) Diabetes
    p.diab.0 <- cbind(1 - CalculateProbDiabetes0(Sex = Sex.pop[, 1], Obes = Obes.pop[, 1]),
                      CalculateProbDiabetes0(Sex = Sex.pop[, 1], Obes = Obes.pop[, 1]))
    Diab.pop[, 1] <- samplev(probs = p.diab.0, m = 1)
    Diab.pop[, 1] <- Diab.pop[, 1] - 1  # (factor levels should be 0 and 1)

    # Record summary data ----------------------------------------------------

    # Bind variables from time t and baseline together
    Population.t <- data.frame(cbind(Sex.pop[, 1], Obes.pop[, t], Diab.pop[, t]))
    vars.cs <- c("Sex", paste0(c("O.t", "D.t"), (t-1)))  # define variables
    names(Population.t) <- vars.cs

    # (a) Cross-sectional frequency table ------------------------------------

    freq.t <- cbind(Time = (t-1), count(Population.t[, vars.cs]))  # create freq table for time t
    names(freq.t) <- names(Frequency.cs.pop)  # rename columns to match Frequency table
    Frequency.cs.pop <- rbind(Frequency.cs.pop, freq.t)

    # (b) Prevalence ---------------------------------------------------------

    ## Obesity
    prev.O <- cbind.data.frame(Time = (t-1), Subgroup = "Obes.prev",
                               prev = CalculatePrevObesityT("overall", Frequency.cs.pop))  # overall
    prev.O.f <- cbind.data.frame(Time = (t-1), Subgroup = "Obes.prev.f",
                                 prev = CalculatePrevObesityT("female", Frequency.cs.pop))  # f
    prev.O.m <- cbind.data.frame(Time = (t-1), Subgroup = "Obes.prev.m",
                                 prev = CalculatePrevObesityT("male", Frequency.cs.pop))  # m
    Obes.prev.pop <- rbind.data.frame(Obes.prev.pop, prev.O, prev.O.f, prev.O.m)

    ## Diabetes
    prev.D <- cbind.data.frame(Time = (t-1), Subgroup = "Diab.prev",
```

```
                                prev = CalculatePrevDiabetesT("overall", Frequency.cs.pop)) # overall
    prev.D.f <- cbind.data.frame(Time = (t-1), Subgroup = "Diab.prev.f",
                                 prev = CalculatePrevDiabetesT("female", Frequency.cs.pop))  # f
    prev.D.m <- cbind.data.frame(Time = (t-1), Subgroup = "Diab.prev.m",
                                 prev = CalculatePrevDiabetesT("male", Frequency.cs.pop))  # m
    Diab.prev.pop <- rbind.data.frame(Diab.prev.pop, prev.D, prev.D.f, prev.D.m)

    # (c) Conditional probabilities --------------------------------------------

    # Convert variables in Population.t dataset to factors
    # (requred for calculating conditonal probabilties)
    Population.t <- data.frame(lapply(Population.t, factor))

    ## (i) Cross-sectional -----------------------------------------------------

    ## Obesity
    var.d <- paste0("O.t", (t-1))  # (define dependent variable)
    var.i <- "Sex"  # (define independent variable)
    cprob.t <- CalculateCPT(dv = var.d, iv = var.i, dataset = Population.t)
    names(cprob.t) <- names(CProbability.Obes.cs)  # rename columns to match CP table
    CProbability.Obes.cs <- rbind.data.frame(CProbability.Obes.cs, cprob.t)

    ## Diabetes
    var.d <- paste0("D.t", (t-1))
    var.i <- c("Sex", paste0("O.t", (t-1)))
    cprob.t <- CalculateCPT(dv = var.d, iv = var.i, dataset = Population.t)
    names(cprob.t) <- names(CProbability.Diab.cs)
    CProbability.Diab.cs <- rbind(CProbability.Diab.cs, cprob.t)

    ## Update time-varying characteristics & record summary data
  } else {

    # Update time-varying characteristics --------------------------------------

    # (a) Obesity --------------------------------------------------------------

    p.obes.t <- cbind(1 - CalculateProbObesityT(Sex = Sex.pop[, 1],
                                                PrevObes = Obes.pop[, (t-1)],
                                                PrevDiab = Diab.pop[, (t-1)]),
                      CalculateProbObesityT(Sex = Sex.pop[, 1],
                                            PrevObes = Obes.pop[, (t-1)],
                                            PrevDiab = Diab.pop[, (t-1)]))
    Obes.pop[, t] <- samplev(probs = p.obes.t, m = 1)
    Obes.pop[, t] <- Obes.pop[, t] - 1  # (factor levels should be 0 and 1)

    # (b) Diabetes -------------------------------------------------------------

    p.diab.t <- cbind(1 - CalculateProbDiabetesT(Sex = Sex.pop[, 1],
                                                 PrevDiab = Diab.pop[, (t-1)],
                                                 Obes = Obes.pop[, t]),
                      CalculateProbDiabetesT(Sex = Sex.pop[, 1],
                                             PrevDiab = Diab.pop[, (t-1)],
                                             Obes = Obes.pop[, t]))
    Diab.pop[, t] <- samplev(probs = p.diab.t, m = 1)
    Diab.pop[, t] <- Diab.pop[, t] - 1  # (factor levels should be 0 and 1)

    # Record summary data ------------------------------------------------------

    # Bind variables from time t, time t-1, and baseline together
    Population.t <- data.frame(cbind(Sex.pop[, 1], Obes.pop[, (t-1)], Diab.pop[, (t-1)],
                                     Obes.pop[, t], Diab.pop[, t])
    vars.cs <- c("Sex", paste0(c("O.t", "D.t"), (t-1)))  # define cross-sectional variables
    vars.ct <- c("Sex", paste0(c("O.t", "D.t"), (t-2)), paste0(c("O.t", "D.t"), (t-1)))  # define
cross-time variables
    names(Population.t) <- vars.ct

    # (a) Cross-sectional frequency table --------------------------------------

    freq.t <- cbind(Time = (t-1), count(Population.t[, vars.cs]))  # create freq table for time t
    names(freq.t) <- names(Frequency.cs.pop)  # rename columns to match Frequency table
    Frequency.cs.pop <- rbind(Frequency.cs.pop, freq.t)

    # (b) Prevalence -----------------------------------------------------------

    ## Obesity
    prev.O <- cbind.data.frame(Time = (t-1), Subgroup = "Obes.prev",
                               prev = CalculatePrevObesityT("overall", Frequency.cs.pop))  # overall
    prev.O.f <- cbind.data.frame(Time = (t-1), Subgroup = "Obes.prev.f",
                                 prev = CalculatePrevObesityT("female", Frequency.cs.pop))  # f
    prev.O.m <- cbind.data.frame(Time = (t-1), Subgroup = "Obes.prev.m",
                                 prev = CalculatePrevObesityT("male", Frequency.cs.pop))  # m
    Obes.prev.pop <- rbind.data.frame(Obes.prev.pop, prev.O, prev.O.f, prev.O.m)

    ## Diabetes
    prev.D <- cbind.data.frame(Time = (t-1), Subgroup = "Diab.prev",
                               prev = CalculatePrevDiabetesT("overall", Frequency.cs.pop)) # overall
    prev.D.f <- cbind.data.frame(Time = (t-1), Subgroup = "Diab.prev.f",
                                 prev = CalculatePrevDiabetesT("female", Frequency.cs.pop))  # f
    prev.D.m <- cbind.data.frame(Time = (t-1), Subgroup = "Diab.prev.m",
                                 prev = CalculatePrevDiabetesT("male", Frequency.cs.pop))  # m
    Diab.prev.pop <- rbind.data.frame(Diab.prev.pop, prev.D, prev.D.f, prev.D.m)

    # (c) Conditional probabilities --------------------------------------------

    # Convert variables in Population.t dataset to factors
    # (requred for calculating conditonal probabilties)
    Population.t <- data.frame(lapply(Population.t, factor, levels = c("0", "1")))

    ## (i) Cross-sectional -----------------------------------------------------

    ## Obesity
    var.d <- paste0("O.t", (t-1))  # (define dependent variable)
    var.i <- "Sex"  # (define independent variable)
```

```
    cprob.t <- CalculateCPT(dv = var.d, iv = var.i, dataset = Population.t)
    names(cprob.t) <- names(CProbability.Obes.cs)  # rename columns to match CP table
    CProbability.Obes.cs <- rbind.data.frame(CProbability.Obes.cs, cprob.t)

    ## Diabetes
    var.d <- paste0("D.t", (t-1))
    var.i <- c("Sex", paste0("O.t", (t-1)))
    cprob.t <- CalculateCPT(dv = var.d, iv = var.i, dataset = Population.t)
    names(cprob.t) <- names(CProbability.Diab.cs)
    CProbability.Diab.cs <- rbind.data.frame(CProbability.Diab.cs, cprob.t)

    ## (ii) Cross-time ------------------------------------------------------

    ## Obesity
    var.d <- paste0("O.t", (t-1))
    var.i <- c("Sex", paste0(c("O.t", "D.t"), (t-2)))
    cprob.t <- CalculateCPT(dv = var.d, iv = var.i, dataset = Population.t)
    names(cprob.t) <- names(CProbability.Obes.ct)
    CProbability.Obes.ct <- rbind.data.frame(CProbability.Obes.ct, cprob.t)

    ## Diabetes
    var.d <- paste0("D.t", (t-1))
    var.i <- c("Sex", paste0("D.t", (t-2)), paste0("O.t", (t-1)))
    cprob.t <- CalculateCPT(dv = var.d, iv = var.i, dataset = Population.t)
    names(cprob.t) <- names(CProbability.Diab.ct)
    CProbability.Diab.ct <- rbind.data.frame(CProbability.Diab.ct, cprob.t)

  }

  # Display progress of simulation
  cat('\r', paste(round((t / N.t.pop * 100), 0),
                  "% done of simulation    ", sep = " "))

}  # (close time loop)

comp.time <- Sys.time() - v; comp.time  # print total simulation time

### (c) Export simulated population data & aggregate tables --------------------

# Export univariable datasets
write.csv(Sex.pop, file = "./Population simulation - vectorised/PopSexData.csv",
          row.names = TRUE)
write.csv(Obes.pop, file = "./Population simulation - vectorised/PopObesData.csv",
          row.names = TRUE)
write.csv(Diab.pop, file = "./Population simulation - vectorised/PopDiabData.csv",
          row.names = TRUE)

# Export complete individual-level dataset
# (1) Create list of all variables to be in whole dataset
allvars <- c("Sex", apply(expand.grid(c("O.t", "D.t"), Time.pop), 1, paste0,
                          collapse = ""))
allvars <- str_replace_all(allvars, fixed(" "), "")  # remove blank spaces from variable names
# (2) Column bind univariable datasets
Population <- data.frame(cbind(Sex.pop, Obes.pop, Diab.pop))
# (3) Reorder variables
Population <- Population[, allvars]
# (4) Export complete dataset
write.csv(Population, file = "./Population simulation - vectorised/PopData.csv",
          row.names = TRUE)
# (5) Export baseline datset
Population.t0 <- Population[, 1:3]
write.csv(Population.t0, file = "./Population simulation - vectorised/PopDataBaseline.csv",
          row.names = TRUE)

# Export frequency table
write.csv(Frequency.cs.pop, file = "./Population simulation - vectorised/PopFreq.csv",
          row.names = FALSE)

# Export prevalence tables
write.csv(Obes.prev.pop, file = "./Population simulation - vectorised/PopObesPrev.csv",
          row.names = FALSE)
write.csv(Diab.prev.pop, file = "./Population simulation - vectorised/PopDiabPrev.csv",
          row.names = FALSE)

# Export conditional probability tables
write.csv(CProbability.Obes.cs, file = "./Population simulation - vectorised/PopObesCPcs.csv",
          row.names = FALSE)
write.csv(CProbability.Obes.ct, file = "./Population simulation - vectorised/PopObesCPct.csv",
          row.names = FALSE)
write.csv(CProbability.Diab.cs, file = "./Population simulation - vectorised/PopDiabCPcs.csv",
          row.names = FALSE)
write.csv(CProbability.Diab.ct, file = "./Population simulation - vectorised/PopDiabCPct.csv",
          row.names = FALSE)
```

C.2.1.2 Counterfactual histories under hypothetical interventions

For the ‘counterfactual history’ simulations, we provide the simulation parameters for each intervention (§C.2.1.2.1), characteristics of the simulated population under each of the six interventions (§C.2.1.2.2), and all annotated R code relating to these simulations (§C.2.1.2.3).

C.2.1.2.1 Simulation parameters

Table C.3 describes the transition parameters governing obesity status at time t for each intervention, compared to the original parameters governing the natural history of the population. All interventions were applied to the population at each time point post-baseline (i.e. for $1 \leq t \leq 10$).

Table C.3 Transition parameters governing obesity status at time *t* for Interventions 1 through 6, compared to those of the natural history

			Probability						
Updated status	**Current status**	**Current covariates**	**Natural history**	**Intervention 1**	**Intervention 2**	**Intervention 3**	**Intervention 4**	**Intervention 5**	**Intervention 6**
Obese	Non-obese	Female, non-diabetic	0.07500	0.00000	1.00000	0.00000	0.06375	0.07500	0.06375
		Female, diabetic	0.10500	0.00000	1.00000	0.00000	0.08925	0.10500	0.08925
		Male, non-diabetic	0.10000	0.00000	1.00000	0.00000	0.08500	0.10000	0.08500
		Male, diabetic	0.13000	0.00000	1.00000	0.00000	0.11050	0.13000	0.11050
	Obese	Female, non-diabetic	0.95000	0.00000	1.00000	0.95000	0.95000	0.85500	0.85500
		Female, diabetic	0.97000	0.00000	1.00000	0.97000	0.97000	0.87300	0.87300
		Male, non-diabetic	0.95000	0.00000	1.00000	0.95000	0.95000	0.85500	0.85500
		Male, diabetic	0.97000	0.00000	1.00000	0.97000	0.97000	0.87300	0.87300

Changes to the transition parameters for the natural history of the population are highlighted in red for easy identification. Transitions from non-obese to obese represent incident cases of obesity, whereas maintaining obesity represents prevalent cases.

Figure C.5 Probability of becoming obese at time *t* for Interventions 1 through 6, compared to those of the natural history

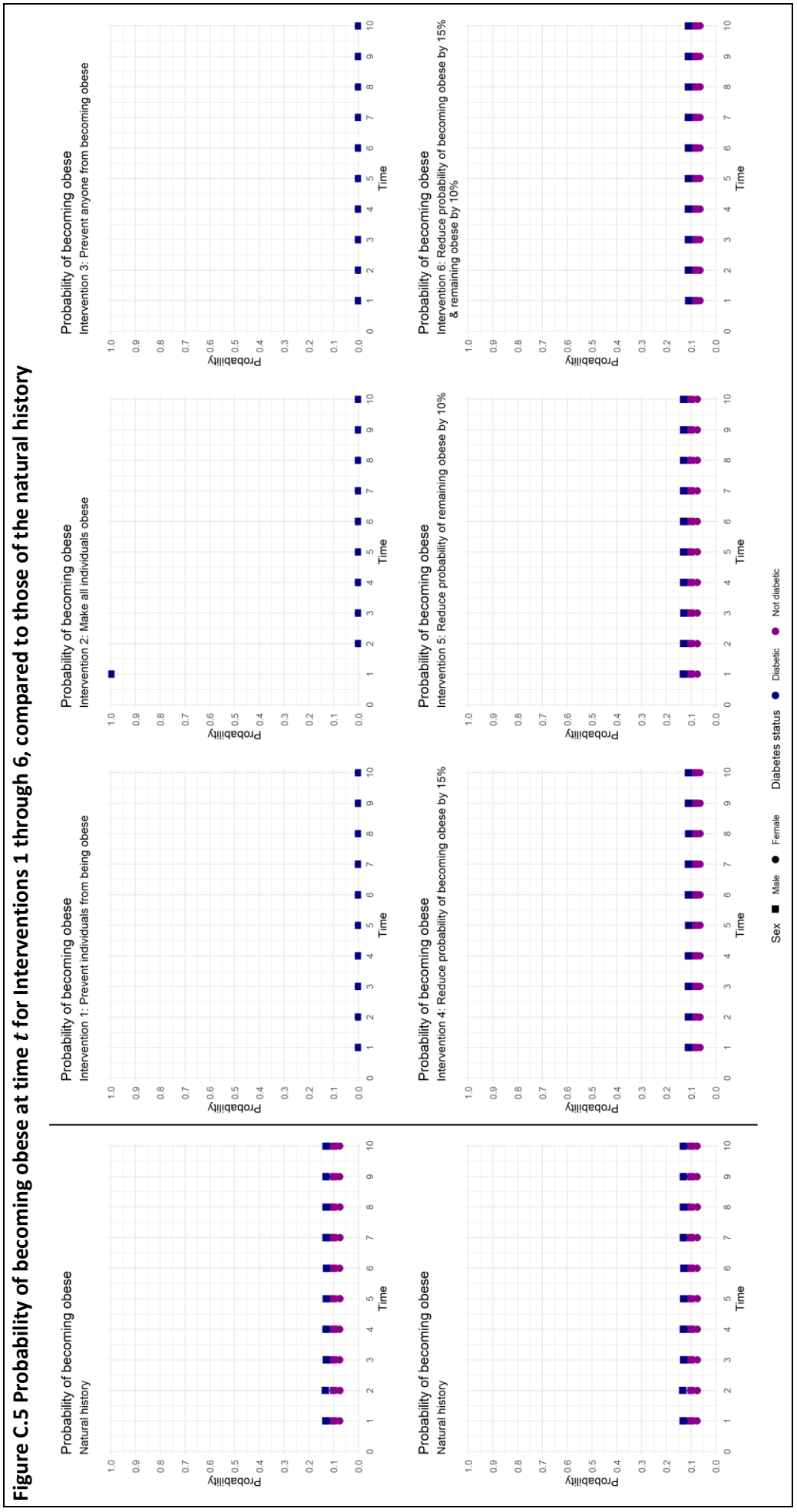


Figure C.6 Probability of remaining obese at time *t* for Interventions 1 through 6, compared to those of the natural history

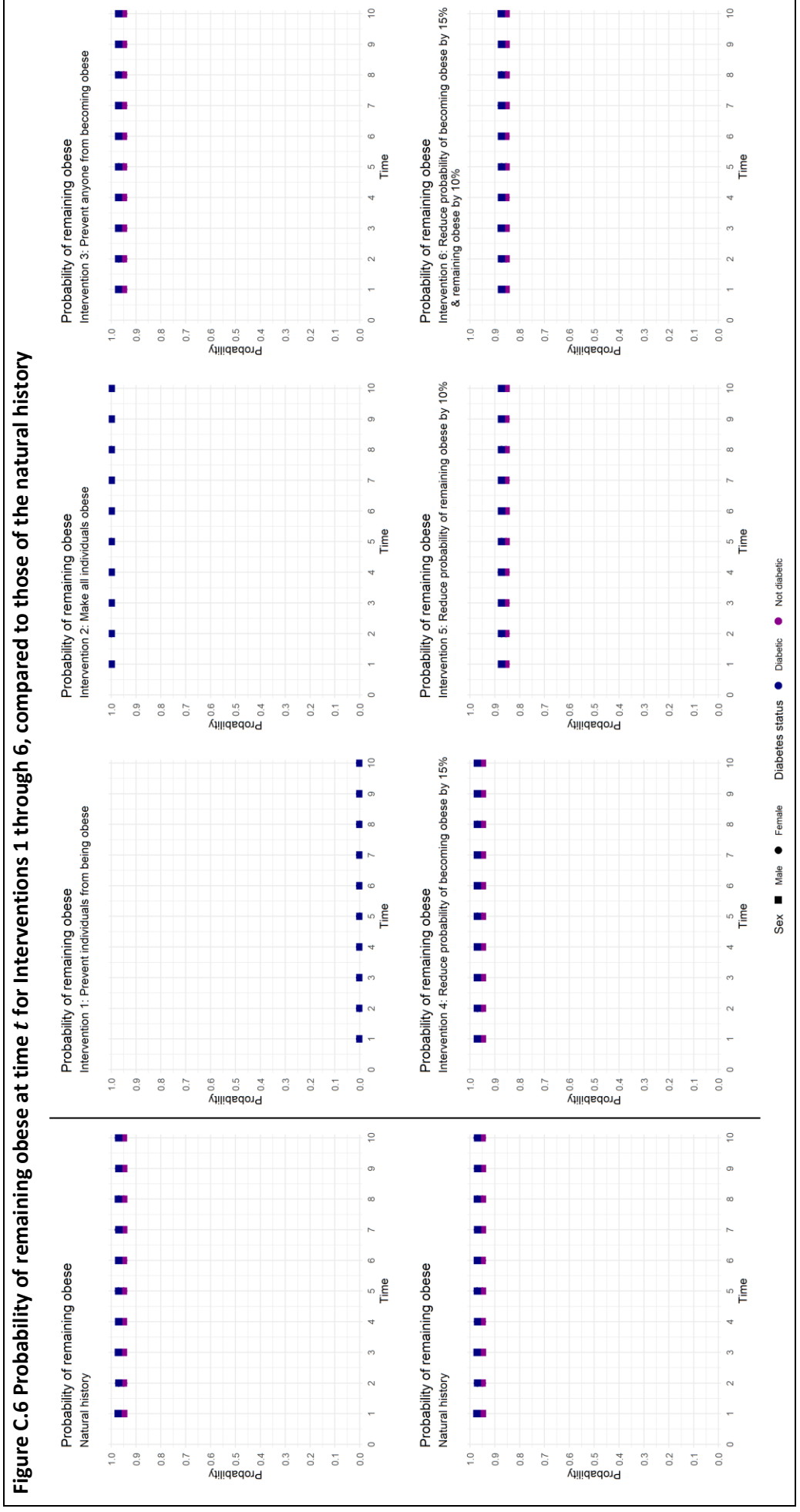


Figure C.7 Probability of becoming diabetic at time *t* for Interventions 1 through 6, compared to those of the natural history

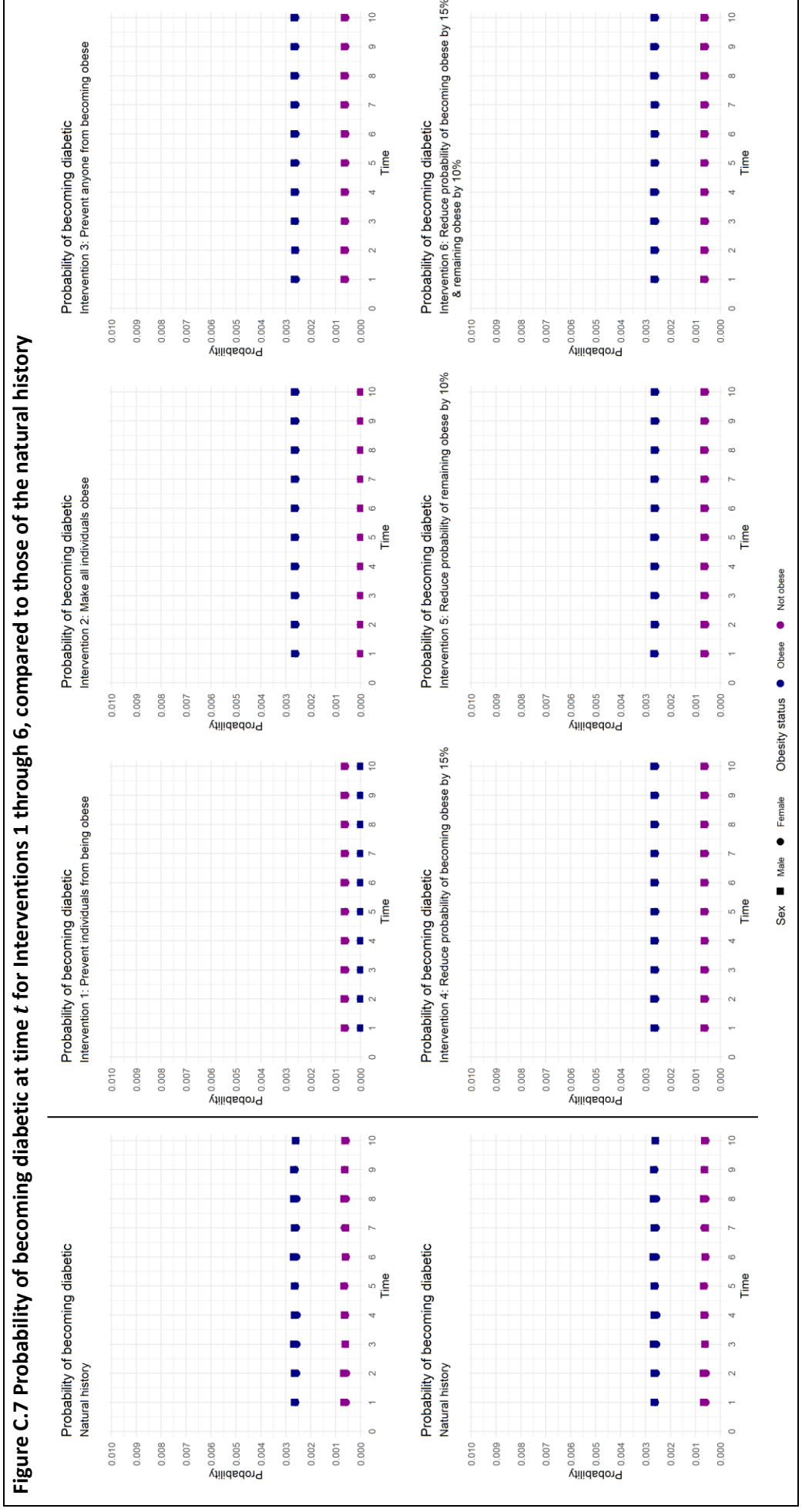
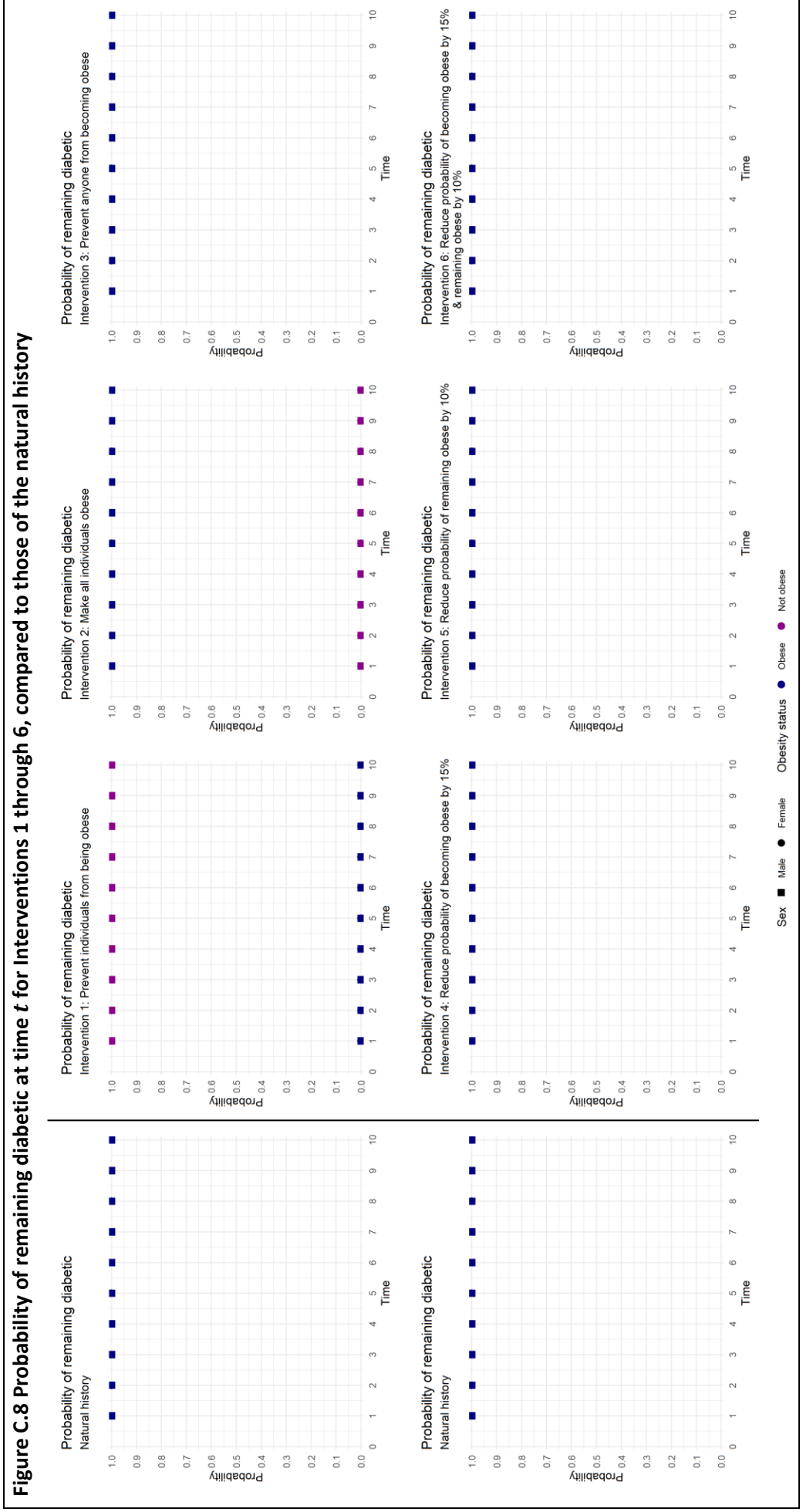


Figure C.8 Probability of remaining diabetic at time *t* for Interventions 1 through 6, compared to those of the natural history



C.2.1.2.2 Characteristics of the simulated population under hypothetical interventions

Figure C.5 and Figure C.6 display the probability of becoming and remaining obese, respectively, for each intervention in the simulated population, compared to that of the natural history. These probabilities are consistent with the parameters specified in Table C.3, confirming that the simulations performed as expected.

Figure C.7 and Figure C.8 display the probability of becoming and remaining diabetic, respectively, for each intervention in the simulated population, compared to that of the natural history. These probabilities are consistent with the original parameters specified in Table C.2, confirming that the simulations performed as expected.

G.5.1.1.1.1 Intervention 1

Prevent anyone from being obese (i.e. reduce the incident and prevalent probabilities of obesity to zero).

Under Intervention 1, all obese individuals at baseline transition to being non-obese at time 1 and maintain this status for the duration of the simulation. Similarly, all non-obese individuals are prevented from developing obesity for the duration of the simulation. This is apparent in Figure 6.2, in which obesity prevalence decreases from 53.57% at baseline to 0.00% at time 1, where it remains until time 10.

The effect of Intervention 1 on diabetes prevalence can be visualised in Figure 6.3, in which overall diabetes prevalence increases from 2.43% at time 0 to 3.04% at time 10. This increase in diabetes prevalence is substantially lower than that which is observed in the natural history of the population, where overall prevalence increases from 2.43% to 4.16% over the same period. Thus, Intervention 1 decreases overall diabetes prevalence by 1.12% compared to no intervention (Table 6.1).

G.5.1.1.1.2 Intervention 2

Make everyone obese (i.e. increase the incident and prevalent probabilities of obesity to one).

Under Intervention 2, all individuals who are not obese at baseline transition to being obese at time 1 and maintain this status for the duration of the simulation. Moreover, all obese individuals are prevented from transitioning out of obesity for the duration of the simulation. Figure 6.2 depicts this intervention, in which obesity prevalence increases from 43.57% at baseline to 100.00% at time 1, where it remains until time 10.

The effect of Intervention 2 on diabetes prevalence can be visualised in Figure 6.2, in which overall diabetes prevalence increases from 2.43% at time 0 to 4.96% at time 10. Compared to the natural history of the population, Intervention 2 thus increases overall diabetes prevalence by 0.80% (Table 6.1).

G.5.1.1.1.3 Intervention 3

Prevent any new individuals from becoming obese (i.e. reduce the incident probability to zero).

Under Intervention 3, non-obese individuals are prevented from becoming obese, but obese individuals may transition out of obesity as normal. This has the effect of dramatically decreasing obesity prevalence over the duration of the simulation – from 53.57% at baseline to 32.41% at time 10 (Figure 6.2). This effect is also evident in (Figure 6.4), in which the proportion of non-obese, nondiabetic individuals (in blue) increases at the expense of obese, nondiabetic individuals (in green).

Unsurprisingly, the effect of Intervention 3 on diabetes prevalence is less than that of Intervention 1 (in which both the incident and prevalent probabilities are reduced to zero), as shown in Figure 6.3. Under Intervention 3, overall diabetes increases from 2.43% to 3.82% between time 0 and time 10, which represents a modest 0.34% decrease in the overall prevalence of diabetes at time 10 compared to the natural history of the population (Table 6.1).

G.5.1.1.1.4 Intervention 4

Reduce the probability of becoming obese by 15% (i.e. reduce the incident probability by 0.15).

Under Intervention 4, obese individuals maintain their obesity status at each time point with the same probability as under the natural history. However, non-obese individuals have a 15% smaller chance of developing obesity at each time point. Because the incidence probability of obesity is relatively modest compared to the prevalent probability in the population under the natural history, Intervention 4 slows the rate of increase of obesity (from 52.57% at baseline to 58.37% at time 10) compared to that of the natural history, but does not reverse the upward trend in obesity prevalence (Figure 6.2).

Therefore, the effect of Intervention 4 is very modest, as evident in Figure 6.3. Diabetes prevalence increases from 2.43% at time 0 to 4.12% at time 10, representing just a 0.04% reduction in prevalence compared to the natural history of the population (Table 6.1).

G.5.1.1.1.5 Intervention 5

Reduce the probability of remaining obese by 10% (i.e. reduce the prevalent probability by 0.10).

Under Intervention 5, non-obese individuals have the same probability of becoming obese at each time point as under the natural history, but obese individuals are 10% less likely to maintain their obesity status at each time point. Because the prevalent probability of obesity is far greater than the incident probability under the natural history, Intervention 5 has the effect of reversing the upward trend in obesity prevalence (Figure 6.2). From baseline to time 10, overall obesity prevalence decrease from 53.57% to 38.97%. This effect can be further

visualised in Figure 6.4, in which the proportion on obese, non-diabetic individuals (in green) decreases at the expense of non-obese, non-diabetic individuals (in blue).

The effect of Intervention 5 on diabetes prevalence is depicted in Figure 6.3. Overall diabetes prevalence increases from 2.43% to 3.85% between baseline and time 10, and this represents a 0.31% decrease in prevalence compared to the natural history of the population (Table 6.1).

G.5.1.1.1.6 Intervention 6

Reduce the probability of becoming obese by 15% and reduce the probability of remaining obese by 10% (i.e. reduce the incident probability by 0.15 and reduce the prevalent probability by 0.10).

Intervention 6 represents a combination of Interventions 4 and 5 – non-obese individuals are 15% less likely to develop obesity at each time point, and obese individuals are 10% less likely to maintain their obesity status at each time point, compared to the natural history. Intervention 6 therefore reverses the upward trend in obese prevalence to an even greater degree than Intervention 5 alone; between baseline and time 10, obesity prevalence decreases from 53.57% to 35.76% (Figure 6.2).

Compared to the natural history of the population, Intervention 6 thus decreases overall diabetes prevalence by 0.35% (Figure 6.3 and Table 6.1).

C.2.1.2.3 Annotated R code

```
################################################################################
# POPULATION INTERVENTION 1 ####################################################
################################################################################

# This code simulates the effects on diabetes prevalence at time 10
# in the artificial longitudinal population
# (generated using the code 'Population simulation - vectorised.R')
# of the following intervention:

# (1) Preventing anyone from being obese

# Simulated individuals have the following 3 attributes:
# Sex (time-fixed): 0 = female, 1 = male
# Obesity (time-varying): 0 = nonobese, 1 = obese
# Diabetes (time-varying): 0 = nondiabetic, 1 = diabetic

################################################################################
## (1) SET UP ------------------------------------------------------------------

# Clear workspace
rm(list = ls())

# Load all required packages
library(readxl); library(stringr); library(Hmisc); library(plyr); library(scales)
library(ggplot2); library(gridExtra); library(HydeNet); library(data.table)

### (a) Population/simulation parameters & baseline data -----------------------

# Import baseline individual-level population dataset
Population.t0 <- read.csv("./Population simulation - vectorised/PopDataBaseline.csv",
                          header = TRUE, row.names = 1)

# Define population parameters
N.i.pop <- nrow(Population.t0)  # number of individuals
N.t.pop <- 11  # number of time points (including baseline)
Time.pop <- as.vector(seq(from = 0, to = (N.t.pop - 1), by = 1),
                      mode = "integer")  # time vector

# Define simulation parameters
N.sim <- 50  # number of simulation runs per intervention

### (b) Tables to store population data ----------------------------------------

# Create empty matrices to store individual-level population data
# Each row represents 1 individual (N.i.pop rows)
# Each column represents 1 time point (N.t.pop columns)
Sex.pop <- matrix(nrow = N.i.pop, ncol = 1,
                  dimnames = list(paste0("ind", 1:N.i.pop), "Sex"))
Obes.pop <- matrix(nrow = N.i.pop, ncol = N.t.pop,
```

```
                    dimnames = list(paste0("ind", 1:N.i.pop),
                                    paste0("O.t", Time.pop)))
Diab.pop <- matrix(nrow = N.i.pop, ncol = N.t.pop,
                    dimnames = list(paste0("ind", 1:N.i.pop),
                                    paste0("D.t", Time.pop)))

# Populate empty matrices with baseline data
# Population is same at time 0 for every simulation
Sex.pop[, 1] <- Population.t0[, 1]
Obes.pop[, 1] <- Population.t0[, 2]
Diab.pop[, 1] <- Population.t0[, 3]

# Remove baseline dataset
rm(Population.t0)

### (c) Tables to store summary data ------------------------------------------

# Create empty cross-sectional frequency table
Frequency.cs.pop.int1 <- data.frame(Sim = numeric(), Time = numeric(),
                                    Sex = numeric(), O.t = numeric(), D.t = numeric(),
                                    freq = numeric())

# Create empty tables to record obesity & diabetes prevalence from population
# (overall and disaggregated by sex)
Obes.prev.pop.int1 <- data.frame(Sim = numeric(), Time = numeric(),
                                 Subgroup = factor(), prev = numeric())
Diab.prev.pop.int1 <- data.frame(Sim = numeric(), Time = numeric(),
                                 Subgroup = factor(), prev = numeric())

# Create empty tables to record cross-sectional conditional probabilities of obesity & diabetes
CProbability.Obes.cs.int1 <- data.frame(Sim = numeric(), Time = numeric(),
                                        Sex = factor(), O.t = factor(),
                                        prob = numeric())
CProbability.Diab.cs.int1 <- data.frame(Sim = numeric(), Time = numeric(),
                                        Sex = factor(), O.t = factor(),
                                        D.t = factor(),
                                        prob = numeric())

# Create empty tables to record cross-time conditional probabilities of obesity & diabetes
CProbability.Obes.ct.int1 <- data.frame(Sim = numeric(), Time = numeric(),
                                        Sex = factor(), O.tminus1 = factor(),
                                        D.tminus1 = factor(), O.t = factor(),
                                        prob = numeric())
CProbability.Diab.ct.int1 <- data.frame(Sim = numeric(), Time = numeric(),
                                        Sex = factor(), D.tminus1 = factor(),
                                        O.t = factor(), D.t = factor(),
                                        prob = numeric())

### (d) Functions -------------------------------------------------------------

#### (i) SampleV function -----------------------------------------------------

# samplev() function
# efficient implementation of the rMultinom() function of the Hmisc package
# from Krijkamp et al (2018)
samplev <- function(probs, m) {
  d <- dim(probs)  # (dimensions of probability matrix)
  n <- d[1]  # (number of rows, i.e. individuals)
  k <- d[2]  # (number of columns, i.e. states)
  lev <- dimnames(probs)[[2]]  # (names of columns, i.e. state values)
  if (!length(lev))
    lev <- 1:k
  ran <- matrix(lev[1], ncol = m, nrow = n)
  U <- t(probs)
  for(i in 2:k) {
    U[i, ] <- U[i, ] + U[i - 1, ]
  }
  if (any((U[k, ] - 1) > 1e-05))
    stop("error in multinom: probabilities do not sum to 1")

  for (j in 1:m) {
    un <- rep(runif(n), rep(k, n))
    ran[, j] <- lev[1 + colSums(un > U)]
  }
  ran
}

#### (ii) Calculate prevalence proportions ------------------------------------

# Calculate Obesity prevalence
# args: group = subgroup, freqtable = frequency table (numeric)
# returns single number (prevalence)
CalculatePrevObesityT <- function (group, freqtable) {

  if (group == "overall") {

    prevalence <- sum(subset(freqtable, Sim == s & Time == (t-1) & O.t == 1)$freq) /
      N.i.pop
    return(prevalence)

  } else if (group == "female") {

    prevalence <- sum(subset(freqtable, Sim == s & Time == (t-1) & Sex == 0 & O.t == 1)$freq) /
      sum(subset(freqtable, Sim == s & Time == (t-1) & Sex == 0)$freq)
    return(prevalence)

  } else if (group == "male") {

    prevalence <- sum(subset(freqtable, Sim == s & Time == (t-1) & Sex == 1 & O.t == 1)$freq) /
      sum(subset(freqtable, Sim == s & Time == (t-1) & Sex == 1)$freq)
    return(prevalence)

  }
```

```
}  # (close funciton loop)


# Calculate Diabetes prevalence
# args: group = subgroup, freqtable = frequency table (numeric)
# returns single number (prevalence)
CalculatePrevDiabetesT <- function (group, freqtable) {

  if (group == "overall") {

    prevalence <- sum(subset(freqtable, Sim == s & Time == (t-1) & D.t == 1)$freq) /
      N.i.pop
    return(prevalence)

  } else if (group == "female") {

    prevalence <- sum(subset(freqtable, Sim == s & Time == (t-1) & Sex == 0 & D.t == 1)$freq) /
      sum(subset(freqtable, Sim == s & Time == (t-1) & Sex == 0)$freq)
    return(prevalence)

  } else if (group == "male") {

    prevalence <- sum(subset(freqtable, Sim == s & Time == (t-1) & Sex == 1 & D.t == 1)$freq) /
      sum(subset(freqtable, Sim == s & Time == (t-1) & Sex == 1)$freq)
    return(prevalence)

  }

}  # (close function loop)

#### (iii) Calculate conditional probabilities ----------------------------------

# Calculate conditional probability table at time t
# args: dv = dependent variable, iv = independent variable(s), dataset = data frame (factorised)
# returns conditional probability table (cprob.t)
CalculateCPT <- function(dv, iv, dataset) {

  # Define formula for use in cpt function (from HydeNet package)
  formula <- as.formula(paste(dv, paste(iv, collapse = " + "), sep = " ~ "))

  # Create conditional probability table
  cprob.t <- cbind(Time = (t-1), am_adt(cpt(formula, data = dataset)))

  return(cprob.t)

}

# Function for converting multidimensional arrays to tables
# (from https://github.com/Rdatatable/data.table/issues/1418)
am_adt <- function(inarray) {
  if (!is.array(inarray)) stop("input must be an array")
  dims <- dim(inarray)
  if (is.null(dimnames(inarray))) {
    inarray <- provideDimnames(inarray, base = list(as.character(seq_len(max(dims)))))
  }
  FT <- if (any(class(inarray) %in% "ftable")) inarray else ftable(inarray)
  out <- data.table(as.table(ftable(FT)))
  nam <- names(out)[seq_along(dims)]
  setorderv(out[, (nam) := lapply(.SD, type.convert), .SDcols = nam], nam)[]
}

################################################################################
## (2) SIMULATION: INTERVENTION 1 ----------------------------------------------

# Prevent anyone from becoming obese

# Set seed
set.seed(1)

### (a) Define conditional probabilities at time t -----------------------------

#### (i) Obesity ---------------------------------------------------------------

# Function to calculate P(Obesity = 1 | Sex, Prev obesity, Prev diabetes) at time t

CalculateProbObesityT <- function(Sex, PrevObes, PrevDiab) {

  p.obes.t <- 0  # under Intervention 1, no individuals may be obese
  return(p.obes.t)

}

#### (ii) Diabetes -------------------------------------------------------------

# Function to calculate P(Diabetes = 1 | Sex, Obesity, Prev diabetes) at time t

CalculateProbDiabetesT <- function(Sex, PrevDiab, Obes) {

  # Incident probability: 0.0006 + 0.00005*Sex + 0.002*Obes
  # Prevalent probability: 1

  p.diab.t <- 0.0006 + 0.00005*Sex + 0.002*Obes +
    PrevDiab*(0.9994 - 0.00005*Sex - 0.002*Obes)
  return(p.diab.t)

}

### (b) Simulation -------------------------------------------------------------

v <- Sys.time()  # record start time of simulation

# (1) Loop through simulation runs
```

```r
for (s in 1:N.sim) {

  # (2) Loop through time points
  for (t in 1:N.t.pop) {

    ## Record summary data
    if (t == 1) {

      # Record summary data -------------------------------------------------

      # Bind variables from time t and baseline together
      Population.t <- data.frame(cbind(Sex.pop[, 1], Obes.pop[, t], Diab.pop[, t]))
      vars.cs <- c("Sex", paste0(c("O.t", "D.t"), (t-1)))  # define variables
      names(Population.t) <- vars.cs

      # (a) Cross-sectional frequency table ---------------------------------

      freq.t <- cbind(Sim = s, Time = (t-1), count(Population.t[, vars.cs]))
      names(freq.t) <- names(Frequency.cs.pop.int1)  # rename columns to match Frequency table
      Frequency.cs.pop.int1 <- rbind(Frequency.cs.pop.int1, freq.t)

      # (b) Prevalence ------------------------------------------------------

      ## Obesity
      prev.O <- cbind.data.frame(Sim = s, Time = (t-1), Subgroup = "Obes.prev",
                                 prev = CalculatePrevObesityT("overall", Frequency.cs.pop.int1))
      prev.O.f <- cbind.data.frame(Sim = s, Time = (t-1), Subgroup = "Obes.prev.f",
                                   prev = CalculatePrevObesityT("female", Frequency.cs.pop.int1))
      prev.O.m <- cbind.data.frame(Sim = s, Time = (t-1), Subgroup = "Obes.prev.m",
                                   prev = CalculatePrevObesityT("male", Frequency.cs.pop.int1))
      Obes.prev.pop.int1 <- rbind.data.frame(Obes.prev.pop.int1, prev.O, prev.O.f, prev.O.m)

      ## Diabetes
      prev.D <- cbind.data.frame(Sim = s, Time = (t-1), Subgroup = "Diab.prev",
                                 prev = CalculatePrevDiabetesT("overall", Frequency.cs.pop.int1))
      prev.D.f <- cbind.data.frame(Sim = s, Time = (t-1), Subgroup = "Diab.prev.f",
                                   prev = CalculatePrevDiabetesT("female", Frequency.cs.pop.int1))
      prev.D.m <- cbind.data.frame(Sim = s, Time = (t-1), Subgroup = "Diab.prev.m",
                                   prev = CalculatePrevDiabetesT("male", Frequency.cs.pop.int1))
      Diab.prev.pop.int1 <- rbind.data.frame(Diab.prev.pop.int1, prev.D, prev.D.f, prev.D.m)

      # (c) Conditional probabilities ---------------------------------------

      # Convert variables in Population.t dataset to factors
      # (requred for calculating conditonal probabilties)
      Population.t <- data.frame(lapply(Population.t, factor))

      ## (i) Cross-sectional -----------------------------------------------

      ## Obesity
      var.d <- paste0("O.t", (t-1))  # (define dependent variable)
      var.i <- "Sex"  # (define independent variable)
      cprob.t <- cbind(Sim = s, CalculateCPT(dv = var.d, iv = var.i, dataset = Population.t))
      names(cprob.t) <- names(CProbability.Obes.cs.int1)  # rename columns to match CP table
      CProbability.Obes.cs.int1 <- rbind.data.frame(CProbability.Obes.cs.int1, cprob.t)

      ## Diabetes
      var.d <- paste0("D.t", (t-1))
      var.i <- c("Sex", paste0("O.t", (t-1)))
      cprob.t <- cbind(Sim = s, CalculateCPT(dv = var.d, iv = var.i, dataset = Population.t))
      names(cprob.t) <- names(CProbability.Diab.cs.int1)
      CProbability.Diab.cs.int1 <- rbind(CProbability.Diab.cs.int1, cprob.t)

      ## Update time-varying characteristics & record summary data
    } else {

      # Update time-varying characteristics ---------------------------------

      # (a) Obesity ---------------------------------------------------------

      p.obes.t <- cbind(1 - CalculateProbObesityT(Sex = Sex.pop[, 1],
                                                  PrevObes = Obes.pop[, (t-1)],
                                                  PrevDiab = Diab.pop[, (t-1)]),
                        CalculateProbObesityT(Sex = Sex.pop[, 1],
                                              PrevObes = Obes.pop[, (t-1)],
                                              PrevDiab = Diab.pop[, (t-1)]))
      Obes.pop[, t] <- samplev(probs = p.obes.t, m = 1)
      Obes.pop[, t] <- Obes.pop[, t] - 1  # (factor levels should be 0 and 1)

      # (b) Diabetes --------------------------------------------------------

      p.diab.t <- cbind(1 - CalculateProbDiabetesT(Sex = Sex.pop[, 1],
                                                   PrevDiab = Diab.pop[, (t-1)],
                                                   Obes = Obes.pop[, t]),
                        CalculateProbDiabetesT(Sex = Sex.pop[, 1],
                                               PrevDiab = Diab.pop[, (t-1)],
                                               Obes = Obes.pop[, t]))
      Diab.pop[, t] <- samplev(probs = p.diab.t, m = 1)
      Diab.pop[, t] <- Diab.pop[, t] - 1  # (factor levels should be 0 and 1)

      # Record summary data -------------------------------------------------

      # Bind variables from time t, time t-1, and baseline together ------------
      Population.t <- data.frame(cbind(Sex.pop[, 1], Obes.pop[, (t-1)], Diab.pop[, (t-1)]),
                                 Obes.pop[, t], Diab.pop[, t])
      vars.cs <- c("Sex", paste0(c("O.t", "D.t"), (t-1)))  # define cross-sectional variables
      vars.ct <- c("Sex", paste0(c("O.t", "D.t"), (t-2)), paste0(c("O.t", "D.t"), (t-1)))  # define
cross-time variables
      names(Population.t) <- vars.ct

      # (a) Cross-sectional frequency table ---------------------------------

      freq.t <- cbind(Sim = s, Time = (t-1), count(Population.t[, vars.cs]))
```

```
      names(freq.t) <- names(Frequency.cs.pop.int1)
      Frequency.cs.pop.int1 <- rbind(Frequency.cs.pop.int1, freq.t)

      # (b) Prevalence ------------------------------------------------------

      ## Obesity
      prev.O <- cbind.data.frame(Sim = s, Time = (t-1), Subgroup = "Obes.prev",
                                 prev = CalculatePrevObesityT("overall", Frequency.cs.pop.int1))
      prev.O.f <- cbind.data.frame(Sim = s, Time = (t-1), Subgroup = "Obes.prev.f",
                                   prev = CalculatePrevObesityT("female", Frequency.cs.pop.int1))
      prev.O.m <- cbind.data.frame(Sim = s, Time = (t-1), Subgroup = "Obes.prev.m",
                                   prev = CalculatePrevObesityT("male", Frequency.cs.pop.int1))
      Obes.prev.pop.int1 <- rbind.data.frame(Obes.prev.pop.int1, prev.O, prev.O.f, prev.O.m)

      ## Diabetes
      prev.D <- cbind.data.frame(Sim = s, Time = (t-1), Subgroup = "Diab.prev",
                                 prev = CalculatePrevDiabetesT("overall", Frequency.cs.pop.int1))
      prev.D.f <- cbind.data.frame(Sim = s, Time = (t-1), Subgroup = "Diab.prev.f",
                                   prev = CalculatePrevDiabetesT("female", Frequency.cs.pop.int1))
      prev.D.m <- cbind.data.frame(Sim = s, Time = (t-1), Subgroup = "Diab.prev.m",
                                   prev = CalculatePrevDiabetesT("male", Frequency.cs.pop.int1))
      Diab.prev.pop.int1 <- rbind.data.frame(Diab.prev.pop.int1, prev.D, prev.D.f, prev.D.m)

      # (c) Conditional probabilities ---------------------------------------

      # Convert variables in Population.t dataset to factors
      # (requred for calculating conditonal probabilties)
      Population.t <- data.frame(lapply(Population.t, factor, levels = c("0", "1")))

      ## (i) Cross-sectional ------------------------------------------------

      ## Obesity
      var.d <- paste0("O.t", (t-1))  # (define dependent variable)
      var.i <- "Sex"  # (define independent variable)
      cprob.t <- cbind(Sim = s, CalculateCPT(dv = var.d, iv = var.i, dataset = Population.t))
      names(cprob.t) <- names(CProbability.Obes.cs.int1)  # rename columns to match CP table
      CProbability.Obes.cs.int1 <- rbind.data.frame(CProbability.Obes.cs.int1, cprob.t)

      ## Diabetes
      var.d <- paste0("D.t", (t-1))
      var.i <- c("Sex", paste0("O.t", (t-1)))
      cprob.t <- cbind(Sim = s, CalculateCPT(dv = var.d, iv = var.i, dataset = Population.t))
      names(cprob.t) <- names(CProbability.Diab.cs.int1)
      CProbability.Diab.cs.int1 <- rbind(CProbability.Diab.cs.int1, cprob.t)

      ## (ii) Cross-time ----------------------------------------------------

      ## Obesity
      var.d <- paste0("O.t", (t-1))
      var.i <- c("Sex", paste0(c("O.t", "D.t"), (t-2)))
      cprob.t <- cbind(Sim = s, CalculateCPT(dv = var.d, iv = var.i, dataset = Population.t))
      names(cprob.t) <- names(CProbability.Obes.ct.int1)
      CProbability.Obes.ct.int1 <- rbind.data.frame(CProbability.Obes.ct.int1, cprob.t)

      ## Diabetes
      var.d <- paste0("D.t", (t-1))
      var.i <- c("Sex", paste0("D.t", (t-2)), paste0("O.t", (t-1)))
      cprob.t <- cbind(Sim = s, CalculateCPT(dv = var.d, iv = var.i, dataset = Population.t))
      names(cprob.t) <- names(CProbability.Diab.ct.int1)
      CProbability.Diab.ct.int1 <- rbind.data.frame(CProbability.Diab.ct.int1, cprob.t)

    }

    # Display progress of simulation
    cat('\r', paste(round((t / N.t.pop * 100), 0),
                    "% done of simulation", s, "of", N.sim, "    ", sep = " "))

  }  # (close time loop)


}  # (close simulation loop)

comp.time <- Sys.time() - v; comp.time  # print total simulation time
# (6.2 min / simulation run)
# (~xx hrs / 100 simulation runs)

### (c) Calculate mean trends ----------------------------------------------

#### (i) Cross-sectional frequencies ----------------------------------------

# Frequency table doesn't show combinations with empty cells
# Use expand.grid function to create full frequency table
f <- expand.grid(D.t = c(0, 1), O.t = c(0, 1), Sex = c(0, 1), Time = Time.pop,
                 Sim = seq(from = 1, to = N.sim, by = 1))
f <- cbind(f[, c("Sim", "Time", "Sex", "O.t", "D.t")], freq = 0)  # initialise frequencies with 0
# Fill f with data from (incomplete) Frequency table
for (i in 1:nrow(Frequency.cs.pop.int1)) {

  sim <- Frequency.cs.pop.int1[i, "Sim"]
  time <- Frequency.cs.pop.int1[i, "Time"]
  sex <- Frequency.cs.pop.int1[i, "Sex"]
  o.t <- Frequency.cs.pop.int1[i, "O.t"]
  d.t <- Frequency.cs.pop.int1[i, "D.t"]
  freq <- Frequency.cs.pop.int1[i, "freq"]

  f[f[, "Sim"] == sim &
      f[, "Time"] == time &
      f[, "Sex"] == sex &
      f[, "O.t"] == o.t &
      f[, "D.t"] == d.t,
    "freq"] <- freq

}
```

```
462
463  # Overwrite incomplete Frequency table
464  Frequency.cs.pop.int1 <- f; rm(f)
465  #write.csv(Frequency.cs.pop.int1, file = "./Population intervention/Intervention 1/PopFreqInt1.csv",
466  row.names = FALSE)
467
468  # Calculate mean frequency at each time
469  Mean.frequency.int1 <- expand.grid(D.t = c(0, 1), O.t = c(0, 1), Sex = c(0, 1),
470                                     Time = Time.pop)
471  Mean.frequency.int1 <- cbind(Sim = "mean", Mean.frequency.int1[, c("Time", "Sex", "O.t", "D.t")],
472                               freq = 0)  # initialise frequencies with 0
473  #str(Mean.frequency.int1)
474  for (i in 1:nrow(Mean.frequency.int1)) {
475
476    time <- Mean.frequency.int1[i, "Time"]
477    sex <- Mean.frequency.int1[i, "Sex"]
478    o.t <- Mean.frequency.int1[i, "O.t"]
479    d.t <- Mean.frequency.int1[i, "D.t"]
480
481    avg <- mean(subset(Frequency.cs.pop.int1, Time == time &
482                         Sex == sex &
483                         O.t == o.t &
484                         D.t == d.t)$freq)
485
486    Mean.frequency.int1[Mean.frequency.int1[, "Time"] == time &
487                          Mean.frequency.int1[, "Sex"] == sex &
488                          Mean.frequency.int1[, "O.t"] == o.t &
489                          Mean.frequency.int1[, "D.t"] == d.t,
490                        "freq"] <- avg
491
492  }
493
494  ### (ii) Prevalence ------------------------------------------------------------
495
496  # Calculate mean prevalence at each time
497  ## Obesity:
498  Mean.obes.prev.int1 <- expand.grid(Subgroup = c("Obes.prev", "Obes.prev.f", "Obes.prev.m"),
499                                     Time = Time.pop)
500  Mean.obes.prev.int1 <- cbind(Sim = "mean",
501                               Mean.obes.prev.int1[, c("Time", "Subgroup")],
502                               prev = 0)  # initialise prevalence with 0
503  #str(Mean.obes.prev.int1)
504  for (i in 1:nrow(Mean.obes.prev.int1)) {
505
506    time <- Mean.obes.prev.int1[i, "Time"]
507    sub <- Mean.obes.prev.int1[i, "Subgroup"]
508
509    avg <- mean(subset(Obes.prev.pop.int1, Time == time & Subgroup == sub)$prev)
510
511    Mean.obes.prev.int1[Mean.obes.prev.int1[, "Time"] == time &
512                          Mean.obes.prev.int1[, "Subgroup"] == sub,
513                        "prev"] <- avg
514
515  }
516  ## Diabetes:
517  Mean.diab.prev.int1 <- expand.grid(Subgroup = c("Diab.prev", "Diab.prev.f", "Diab.prev.m"),
518                                     Time = Time.pop)
519  Mean.diab.prev.int1 <- cbind(Sim = "mean",
520                               Mean.diab.prev.int1[, c("Time", "Subgroup")],
521                               prev = 0)  # initialise prevalence with 0
522  #str(Mean.diab.prev.int1)
523  for (i in 1:nrow(Mean.diab.prev.int1)) {
524
525    time <- Mean.diab.prev.int1[i, "Time"]
526    sub <- Mean.diab.prev.int1[i, "Subgroup"]
527
528    avg <- mean(subset(Diab.prev.pop.int1, Time == time & Subgroup == sub)$prev)
529
530    Mean.diab.prev.int1[Mean.diab.prev.int1[, "Time"] == time &
531                          Mean.diab.prev.int1[, "Subgroup"] == sub,
532                        "prev"] <- avg
533
534  }
535
536  ### (iii) Conditional probabilities --------------------------------------------
537
538  # Calculate mean CP at each time point
539  ## Obesity (cross-sectional):
540  Mean.CP.Obes.cs.int1 <- expand.grid(O.t = c(0, 1), Sex = c(0, 1), Time = Time.pop)
541  Mean.CP.Obes.cs.int1 <- cbind(Sim = "mean",
542                                Mean.CP.Obes.cs.int1[, c("Time", "Sex", "O.t")],
543                                prob = 0)
544  for (i in 1:nrow(Mean.CP.Obes.cs.int1)) {
545
546    time <- Mean.CP.Obes.cs.int1[i, "Time"]
547    sex <- Mean.CP.Obes.cs.int1[i, "Sex"]
548    o.t <- Mean.CP.Obes.cs.int1[i, "O.t"]
549
550    avg <- mean(subset(CProbability.Obes.cs.int1, Time == time &
551                         Sex == sex &
552                         O.t == o.t)$prob)
553
554    Mean.CP.Obes.cs.int1[Mean.CP.Obes.cs.int1[, "Time"] == time &
555                           Mean.CP.Obes.cs.int1[, "Sex"] == sex &
556                           Mean.CP.Obes.cs.int1[, "O.t"] == o.t,
557                         "prob"] <- avg
558
559  }
560  ## Obesity (cross-time):
561  Mean.CP.Obes.ct.int1 <- expand.grid(O.t = c(0, 1), D.tminus1 = c(0, 1),
562                                      O.tminus1 = c(0, 1), Sex = c(0, 1),
563                                      Time = Time.pop[-1])
564  Mean.CP.Obes.ct.int1 <- cbind(Sim = "mean",
```

```
                                Mean.CP.Obes.ct.int1[, c("Time", "Sex", "O.tminus1", "D.tminus1",
"O.t")],
                                prob = 0)
for (i in 1:nrow(Mean.CP.Obes.ct.int1)) {

  time <- Mean.CP.Obes.ct.int1[i, "Time"]
  sex <- Mean.CP.Obes.ct.int1[i, "Sex"]
  o.tminus1 <- Mean.CP.Obes.ct.int1[i, "O.tminus1"]
  d.tminus1 <- Mean.CP.Obes.ct.int1[i, "D.tminus1"]
  o.t <- Mean.CP.Obes.ct.int1[i, "O.t"]

  avg <- mean(subset(CProbability.Obes.ct.int1, Time == time &
                       Sex == sex &
                       O.tminus1 == o.tminus1 &
                       D.tminus1 == d.tminus1 &
                       O.t == o.t)$prob)

  Mean.CP.Obes.ct.int1[Mean.CP.Obes.ct.int1[, "Time"] == time &
                         Mean.CP.Obes.ct.int1[, "Sex"] == sex &
                         Mean.CP.Obes.ct.int1[, "O.tminus1"] == o.tminus1 &
                         Mean.CP.Obes.ct.int1[, "D.tminus1"] == d.tminus1 &
                         Mean.CP.Obes.ct.int1[, "O.t"] == o.t,
                       "prob"] <- avg

}
## Diabetes (cross-sectional):
Mean.CP.Diab.cs.int1 <- expand.grid(D.t = c(0, 1), O.t = c(0, 1),
                                    Sex = c(0, 1), Time = Time.pop)
Mean.CP.Diab.cs.int1 <- cbind(Sim = "mean",
                                Mean.CP.Diab.cs.int1[, c("Time", "Sex", "O.t", "D.t")],
                                prob = 0)
for (i in 1:nrow(Mean.CP.Diab.cs.int1)) {

  time <- Mean.CP.Diab.cs.int1[i, "Time"]
  sex <- Mean.CP.Diab.cs.int1[i, "Sex"]
  o.t <- Mean.CP.Diab.cs.int1[i, "O.t"]
  d.t <- Mean.CP.Diab.cs.int1[i, "D.t"]

  avg <- mean(subset(CProbability.Diab.cs.int1, Time == time &
                       Sex == sex &
                       O.t == o.t &
                       D.t == d.t)$prob)

  Mean.CP.Diab.cs.int1[Mean.CP.Diab.cs.int1[, "Time"] == time &
                         Mean.CP.Diab.cs.int1[, "Sex"] == sex &
                         Mean.CP.Diab.cs.int1[, "O.t"] == o.t &
                         Mean.CP.Diab.cs.int1[, "D.t"]== d.t,
                       "prob"] <- avg

}
## Diabetes (cross-time):
Mean.CP.Diab.ct.int1 <- expand.grid(D.t = c(0, 1), O.t = c(0, 1),
                                    D.tminus1 = c(0, 1), Sex = c(0, 1),
                                    Time = Time.pop[-1])
Mean.CP.Diab.ct.int1 <- cbind(Sim = "mean",
                                Mean.CP.Diab.ct.int1[, c("Time", "Sex", "D.tminus1", "O.t", "D.t")],
                                prob = 0)
for (i in 1:nrow(Mean.CP.Diab.ct.int1)) {

  time <- Mean.CP.Diab.ct.int1[i, "Time"]
  sex <- Mean.CP.Diab.ct.int1[i, "Sex"]
  d.tminus1 <- Mean.CP.Diab.ct.int1[i, "D.tminus1"]
  o.t <- Mean.CP.Diab.ct.int1[i, "O.t"]
  d.t <- Mean.CP.Diab.ct.int1[i, "D.t"]

  avg <- mean(subset(CProbability.Diab.ct.int1, Time == time &
                       Sex == sex &
                       D.tminus1 == d.tminus1 &
                       O.t == o.t &
                       D.t == d.t)$prob)

  Mean.CP.Diab.ct.int1[Mean.CP.Diab.ct.int1[, "Time"] == time &
                         Mean.CP.Diab.ct.int1[, "Sex"] == sex &
                         Mean.CP.Diab.ct.int1[, "D.tminus1"] == d.tminus1 &
                         Mean.CP.Diab.ct.int1[, "O.t"] == o.t &
                         Mean.CP.Diab.ct.int1[, "D.t"] == d.t,
                       "prob"] <- avg

}


### (d) Export aggregate tables & mean trends tables ----------------------------

# Export frequency table
write.csv(Frequency.cs.pop.int1, file = "./Population intervention/Intervention 1/PopFreqInt1.csv",
          row.names = FALSE)

# Export prevalence tables
write.csv(Obes.prev.pop.int1, file = "./Population intervention/Intervention 1/PopObesPrevInt1.csv",
          row.names = FALSE)
write.csv(Diab.prev.pop.int1, file = "./Population intervention/Intervention 1/PopDiabPrevInt1.csv",
          row.names = FALSE)

# Export conditional probability tables
write.csv(CProbability.Obes.cs.int1, file = "./Population intervention/Intervention
1/PopObesCPcsInt1.csv",
          row.names = FALSE)
write.csv(CProbability.Obes.ct.int1, file = "./Population intervention/Intervention
1/PopObesCPctInt1.csv",
          row.names = FALSE)
write.csv(CProbability.Diab.cs.int1, file = "./Population intervention/Intervention
1/PopDiabCPcsInt1.csv",
          row.names = FALSE)
```

```
668 write.csv(CProbability.Diab.ct.int1, file = "./Population intervention/Intervention
669 1/PopDiabCPctInt1.csv",
670           row.names = FALSE)
671
672 # Export mean cross-sectional frequency tables
673 write.csv(Mean.frequency.int1, file = "./Population intervention/Intervention
674 1/PopFreqMeanInt1.csv",
675           row.names = FALSE)
676
677 # Export mean prevalence tables
678 write.csv(Mean.obes.prev.int1, file = "./Population intervention/Intervention
679 1/PopObesPrevMeanInt1.csv",
680           row.names = FALSE)
681 write.csv(Mean.diab.prev.int1, file = "./Population intervention/Intervention
682 1/PopDiabPrevMeanInt1.csv",
683           row.names = FALSE)
684
685 # Export mean conditional probability tables
686 write.csv(Mean.CP.Obes.cs.int1, file = "./Population intervention/Intervention
687 1/PopObesCPcsMeanInt1.csv",
688           row.names = FALSE)
689 write.csv(Mean.CP.Obes.ct.int1, file = "./Population intervention/Intervention
690 1/PopObesCPctMeanInt1.csv",
691           row.names = FALSE)
692 write.csv(Mean.CP.Diab.cs.int1, file = "./Population intervention/Intervention
693 1/PopDiabCPcsMeanInt1.csv",
694           row.names = FALSE)
695 write.csv(Mean.CP.Diab.ct.int1, file = "./Population intervention/Intervention
696 1/PopDiabCPctMeanInt1.csv",
697           row.names = FALSE)
```

Note that the above code relates to Intervention 1; for all other interventions, the probability of obesity at time t (line 231) varies according to the specific intervention. For Interventions 2 through 6, respectively:

```
231 p.obes.t <- 1
```

```
231 p.obes.t <- PrevObes*(0.95 + 0.02*PrevDiab)
```

```
231 p.obes.t <- 0.06375 + 0.02125*Sex + 0.0255*PrevDiab +
        PrevObes*(0.88625 - 0.02125*Sex - 0.0055*PrevDiab)
```

```
231 p.obes.t <- 0.075 + 0.025*Sex + 0.03*PrevDiab +
        PrevObes*(0.78 - 0.025*Sex - 0.012*PrevDiab)
```

```
231 p.obes.t <- 0.06375 + 0.02125*Sex + 0.0255*PrevDiab +
        PrevObes*(0.79125 - 0.02125*Sex - 0.0075*PrevDiab)
```

The output from each simulated intervention is then saved to its respective subfolder ('Intervention 2' through 'Intervention 6', respectively).

C.2.2 Comparison of the g-formula versus microsimulation for estimating true causal effects in the population

In this subsection, we provide details relating to the simulations in which the g-formula and microsimulation were used to estimate the true natural and counterfactual histories in the population, which is described in Section 6.4.2. For each method (the g-formula in §C.2.2.1 and microsimulation in §C.2.2.2), we provide additional details relating to the simulations performed; this includes a fuller evaluation of the autocorrelation structures, the results of modelling the counterfactual histories under each of the six interventions, and all annotated R code.

C.2.2.1 The g-formula

Because it represents the true autocorrelation structure of the population, AS1 is expected to replicate the true natural history of the population, thereby producing unbiased estimates of obesity and diabetes prevalence in the population at all time points. Moreover, it is expected to replicate the true counterfactual histories under Interventions 1 through 6. Modelling AS1 is therefore expected to produce unbiased estimates of all intervention effects.

AS2 is not expected to faithfully replicate the true natural history of the population because it does not correctly model the dependence between time points. Neither is modelling AS2 expected to faithfully replicate the true counterfactual scenarios under interventions on obesity. Modelling AS2 is therefore expected to produce biased estimates of all intervention effects.

AS3 is expected to produce a population whose characteristics under the natural history are consistent with the true cross-sectional characteristics of the population. However, under Interventions 1 through 6, AS3 cannot be expected to replicate the true counterfactual histories because the effects of the interventions on obesity which are implemented at each time point are not carried forward. Modelling AS3 is thus expected to produce biased estimates of all intervention effects.

C.2.2.1.1 Counterfactual histories under hypothetical intervention

Here, we present the results of using the g-formula to model the counterfactual histories for Interventions 2 through 6 (the results of Intervention 1 are presented in Chapter 6, Section 6.4.2.2.3), according to each of the three autocorrelation structures (AS1 through AS3).

Figures which compare the true effect of each intervention on obesity and diabetes prevalence in the population with those modelled by the g-formula are shown in Figure C.9 (Intervention 2), Figure C.10 (Intervention 3), Figure C.11 (Intervention 4), Figure C.12 (Intervention 5), and Figure C.13 (Intervention 6).

Figure C.9 Counterfactual histories of obesity and diabetes prevalence under Intervention 2 for each of AS1 through AS3 modelled using the g-formula, compared to the true counterfactual history

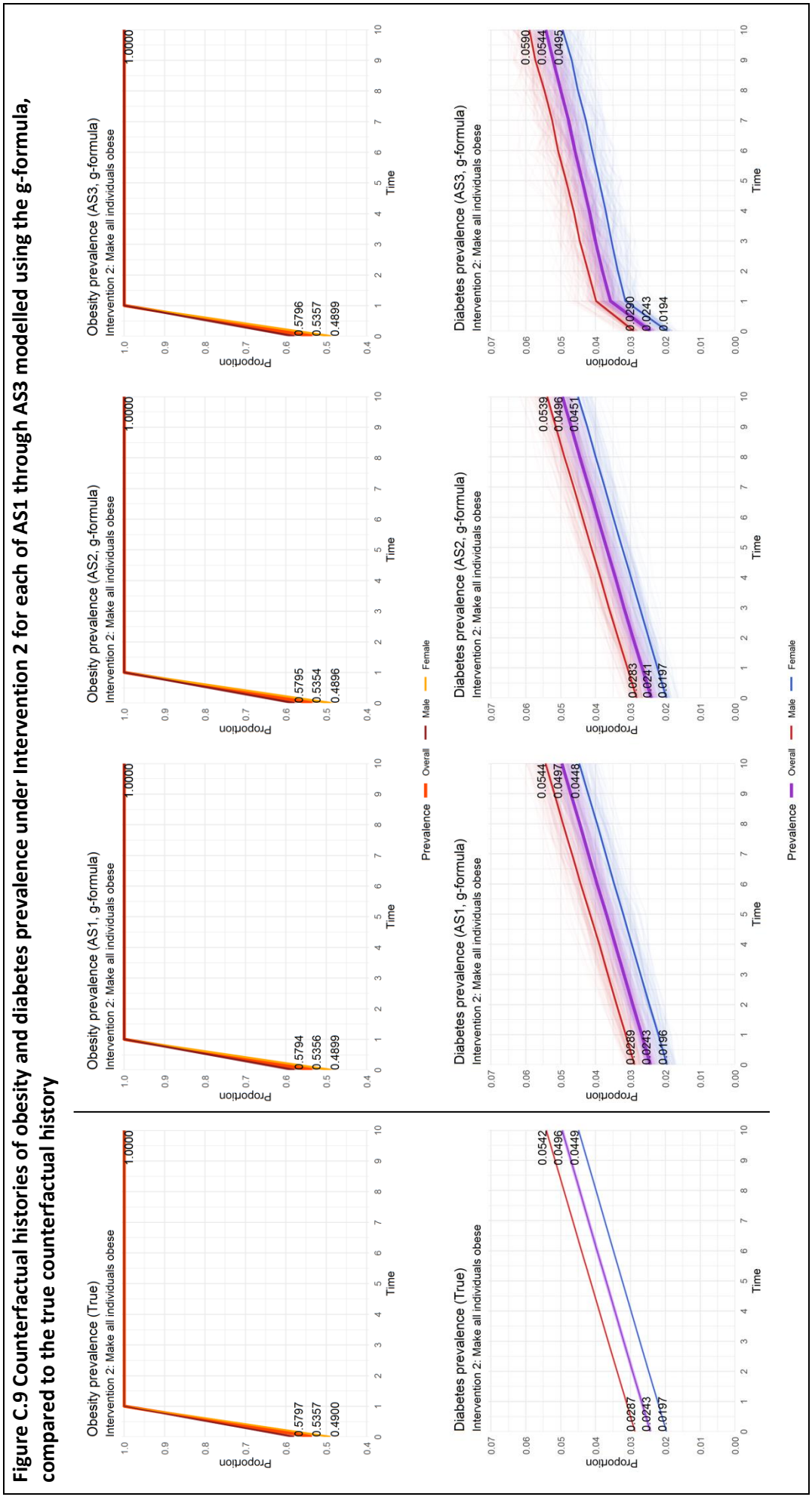


Figure C.10 Counterfactual histories of obesity and diabetes prevalence under Intervention 3 for each of AS1 through AS3 modelled using the g-formula, compared to the true counterfactual history

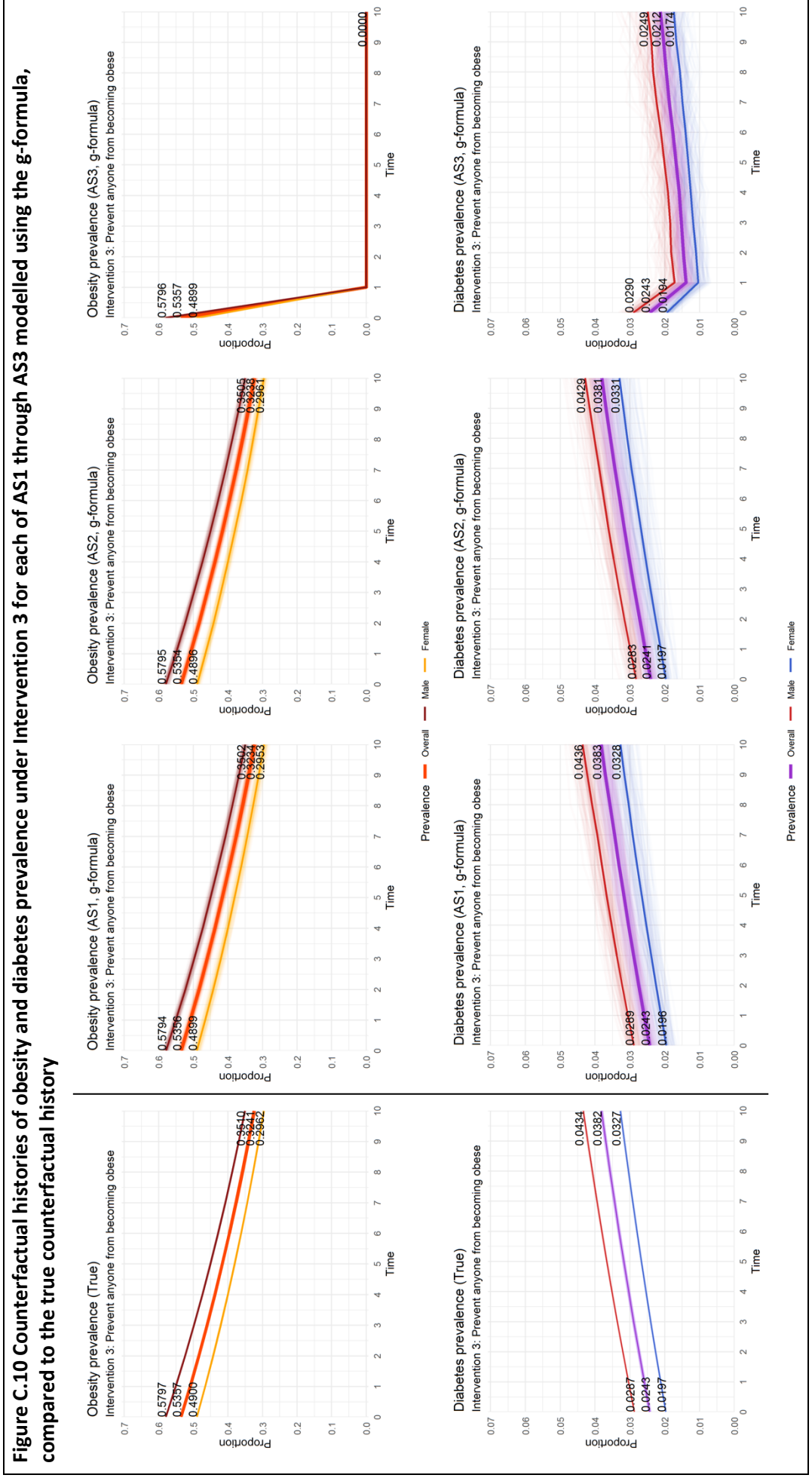


Figure C.11 Counterfactual histories of obesity and diabetes prevalence under Intervention 4 for each of AS1 through AS3 modelled using the g-formula, compared to the true counterfactual history

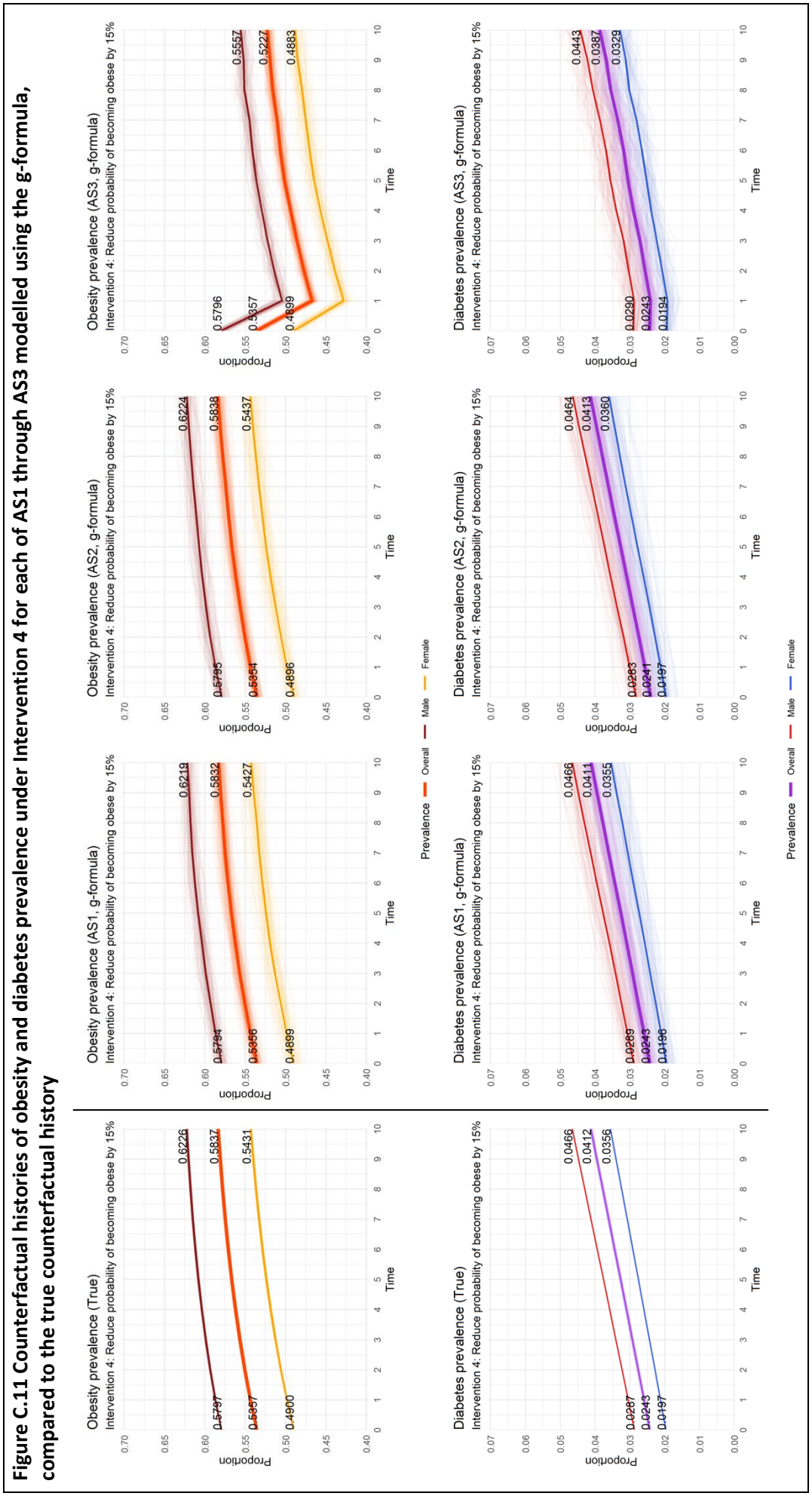


Figure C.12 Counterfactual histories of obesity and diabetes prevalence under Intervention 5 for each of AS1 through AS3 modelled using the g-formula, compared to the true counterfactual history

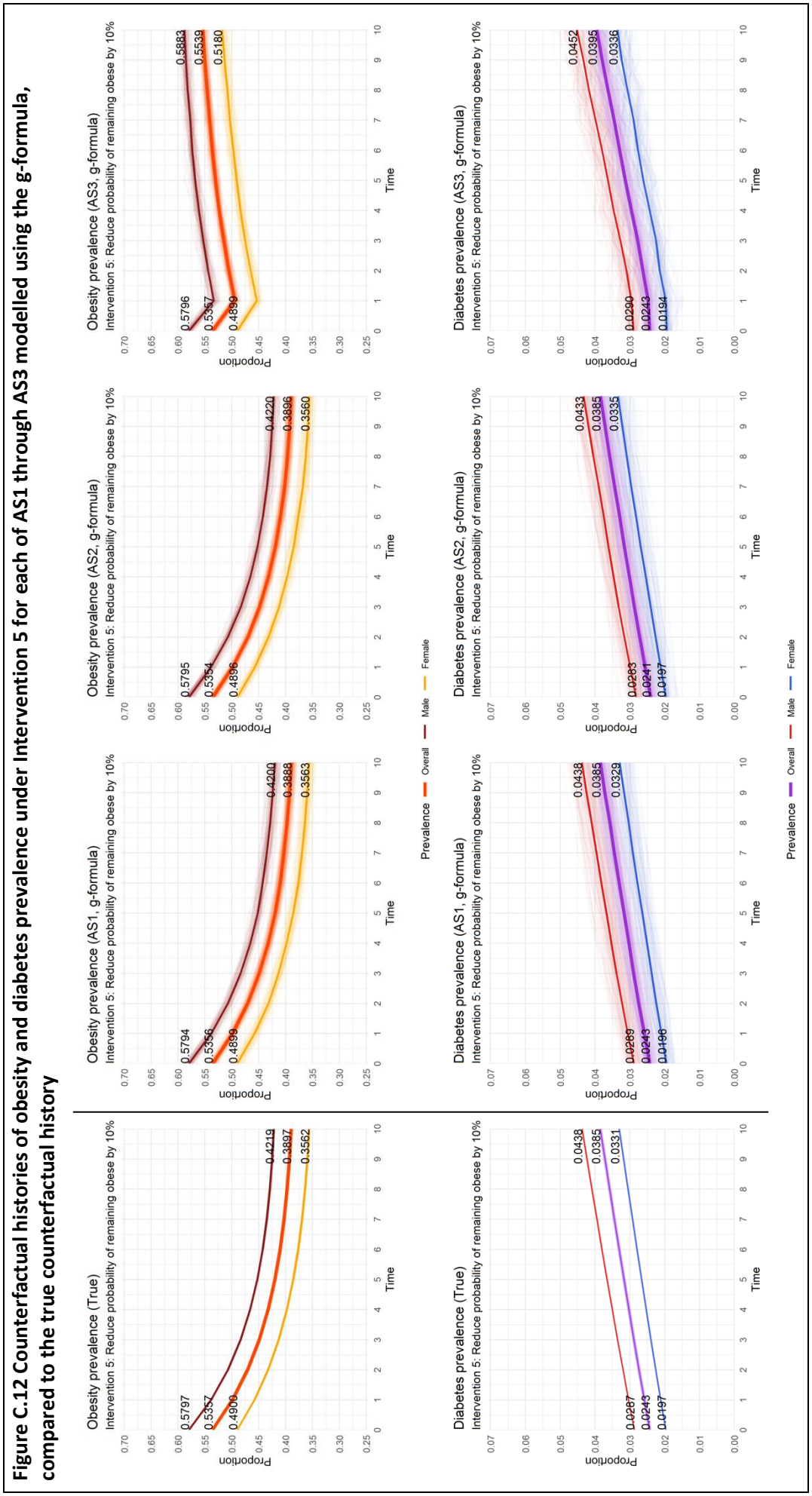
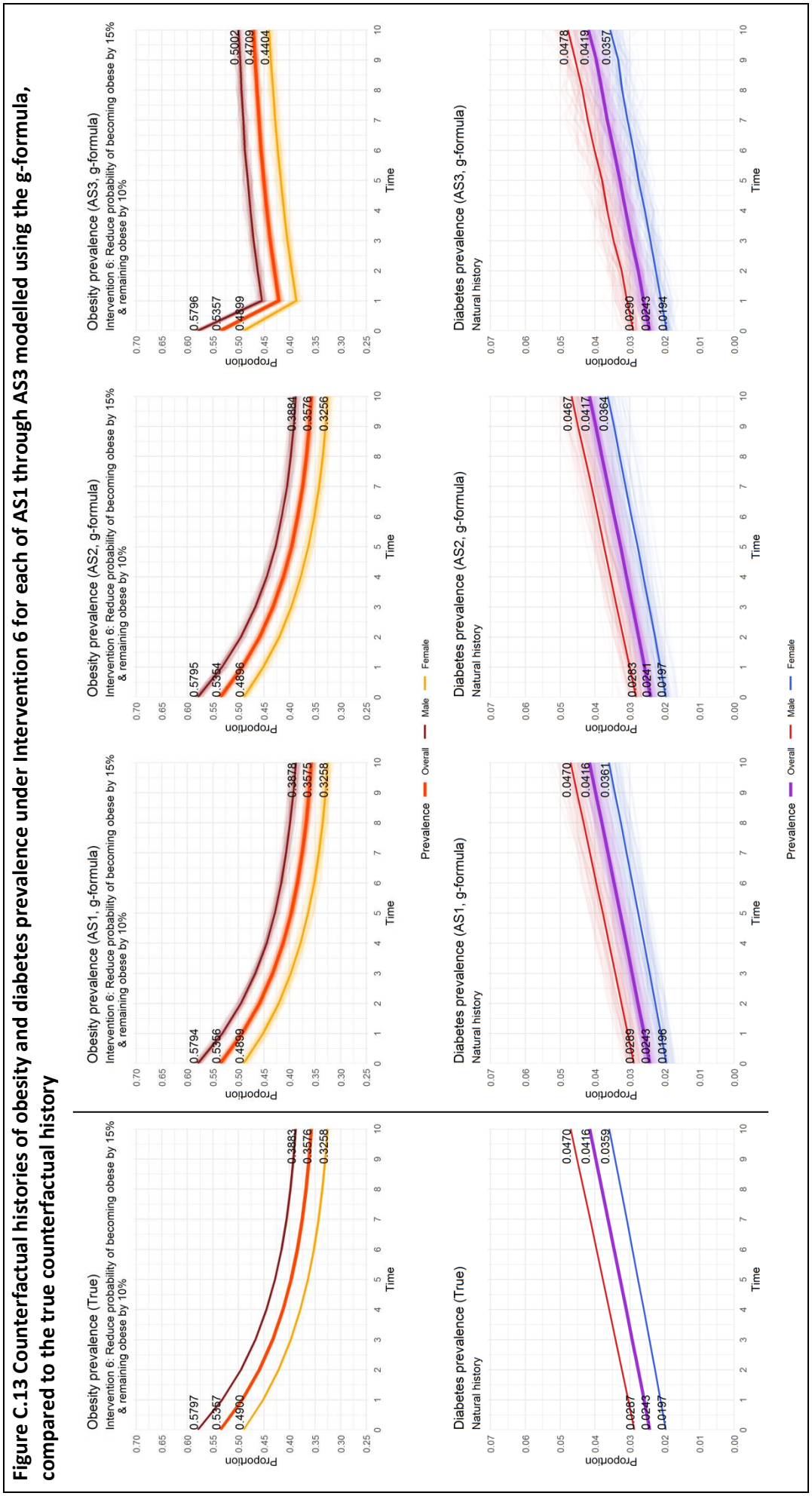


Figure C.13 Counterfactual histories of obesity and diabetes prevalence under Intervention 6 for each of AS1 through AS3 modelled using the g-formula, compared to the true counterfactual history



C.2.2.1.2 Annotated R code

```
################################################################################
## MSM: AUTOCORRELATION STRUCTURE 1 ############################################
################################################################################

# This code simulates the 'natural history' of an artificial longitudinal population
# sampled from a population of 5 million ('Population simulation - vectorised.R')
# using a TIME-based, discrete time microsimulation model

# The true autocorrelation structure of the source population is modelled

# Simulated individuals have the following 3 attributes:
# Sex (time-fixed): 0 = female, 1 = male
# Obesity (time-varying): 0 = nonobese, 1 = obese
# Diabetes (time-varying): 0 = nondiabetic, 1 = diabetic

# It then simulates the effects on diabetes prevalence at time 10
# of the following interventions:

# (1) Preventing anyone from being obese
# (2) Making everyone obese
# (3) Preventing any new obese individuals
# (4) Reducing the probability of becoming obese by 15%
# (5) Reducing the probability of remaining obese by 10%
# (6) Reducing the probability of becoming obese by 15% and remaining obese by 10%

################################################################################
## (1) SET UP ------------------------------------------------------------------

# Clear workspace
rm(list = ls())

# Load all required packages
library(readxl); library(stringr); library(Hmisc); library(plyr)
library(scales); library(ggplot2); library(gridExtra); library(HydeNet)
library(data.table)

### (a) Population/sample/simulation parameters & population data --------------

# Import individual-level population dataset
Population <- read.csv("./Population simulation - vectorised/PopData.csv",
                       header = TRUE, row.names = 1)

# Define population parameters
N.i.pop <- nrow(Population)  # number of individuals in population
N.t.pop <- 11  # number of time points (including baseline)

# Define sample parameters
N.i.sam <- 20000  # number of individuals to sample from population
N.t.sam <- 11  # number of time points (including baseline)
Time.sam <- as.vector(seq(from = 0, to = (N.t.sam - 1), by = 1),
                      mode = "integer")  # time vector

# Define simulation parameters
N.sim <- 100  # number of simulation runs (per intervention/natural history)
N.int <- 6  # number of interventions (not including natural history)

### (b) Tables to store sample simulation data --------------------------------

# Create empty matrices to store individual-level sample data
# Each row represents 1 individual (N.i.sam rows)
# Each column represents 1 time point (N.t.sam columns)
Sex.sam1 <- matrix(nrow = N.i.sam, ncol = 1,
                   dimnames = list(paste0("ind", 1:N.i.sam), "Sex"))
Obes.sam1 <- matrix(nrow = N.i.sam, ncol = N.t.sam,
                    dimnames = list(paste0("ind", 1:N.i.sam),
                                    paste0("O.t", Time.sam)))
Diab.sam1 <- matrix(nrow = N.i.sam, ncol = N.t.sam,
                    dimnames = list(paste0("ind", 1:N.i.sam),
                                    paste0("D.t", Time.sam)))

### (c) Tables to store summary data ------------------------------------------

# Create empty cross-sectional frequency table
Frequency.cs.sam1 <- data.frame(Sim = numeric(), Time = numeric(),
                                Sex = numeric(), O.t = numeric(), D.t = numeric(),
                                freq = numeric())

# Create empty tables to record obesity & diabetes prevalence from sample
# (overall and disaggregated by sex)
Obes.prev.sam1 <- data.frame(Sim = numeric(), Time = numeric(),
                             Subgroup = factor(), prev = numeric())
Diab.prev.sam1 <- data.frame(Sim = numeric(), Time = numeric(),
                             Subgroup = factor(), prev = numeric())

# Create empty tables to record cross-sectional conditional probabilities of obesity & diabetes
CProbability.Obes.cs.sam1 <- data.frame(Sim = numeric(), Time = numeric(),
                                        Sex = factor(), O.t = factor(),
                                        prob = numeric())
CProbability.Diab.cs.sam1 <- data.frame(Sim = numeric(), Time = numeric(),
                                        Sex = factor(), O.t = factor(),
                                        D.t = factor(),
                                        prob = numeric())

# Create empty tables to record cross-time conditional probabilities of obesity & diabetes
CProbability.Obes.ct.sam1 <- data.frame(Sim = numeric(), Time = numeric(),
```

```
                                      Sex = factor(), O.tminus1 = factor(),
                                      D.tminus1 = factor(), O.t = factor(),
                                      prob = numeric())
CProbability.Diab.ct.sam1 <- data.frame(Sim = numeric(), Time = numeric(),
                                      Sex = factor(), D.tminus1 = factor(),
                                      O.t = factor(), D.t = factor(),
                                      prob = numeric())

# Create empty table to record estimated parameters governing transition probabilities
TransitionParameters.sam1 <- data.frame(Sim = numeric(), Time = numeric(),
                                       Parameter = factor(), value = numeric())


# Create empty tables to store mean frequency, prevalence, conditional probability, and transition
parameter trends
## Cross-sectional frequencies
Mean.frequency.sam1 <- expand.grid(D.t = c(0, 1), O.t = c(0, 1), Sex = c(0, 1),
                                   Time = Time.sam)
Mean.frequency.sam1 <- cbind(Sim = "mean",
                             Mean.frequency.sam1[, c("Time", "Sex", "O.t", "D.t")],
                             freq = 0)  # initialise frequencies with 0
## Obesity prevalence
Mean.obes.prev.sam1 <- expand.grid(Subgroup = c("Obes.prev", "Obes.prev.f", "Obes.prev.m"),
                                   Time = Time.sam)
Mean.obes.prev.sam1 <- cbind(Sim = "mean",
                             Mean.obes.prev.sam1[, c("Time", "Subgroup")],
                             prev = 0)  # initialise prevalence with 0
## Diabetes prevalence
Mean.diab.prev.sam1 <- expand.grid(Subgroup = c("Diab.prev", "Diab.prev.f", "Diab.prev.m"),
                                   Time = Time.sam)
Mean.diab.prev.sam1 <- cbind(Sim = "mean",
                             Mean.diab.prev.sam1[, c("Time", "Subgroup")],
                             prev = 0)  # initialise prevalence with 0
## CP obesity - cross-sectional
Mean.CP.Obes.cs.sam1 <- expand.grid(O.t = c(0, 1), Sex = c(0, 1), Time = Time.sam)
Mean.CP.Obes.cs.sam1 <- cbind(Sim = "mean",
                              Mean.CP.Obes.cs.sam1[, c("Time", "Sex", "O.t")],
                              prob = 0)  # initialise probs with 0
## CP obesity - cross-time
Mean.CP.Obes.ct.sam1 <- expand.grid(O.t = c(0, 1), D.tminus1 = c(0, 1),
                                    O.tminus1 = c(0, 1), Sex = c(0, 1),
                                    Time = Time.sam[-1])
Mean.CP.Obes.ct.sam1 <- cbind(Sim = "mean",
                              Mean.CP.Obes.ct.sam1[, c("Time", "Sex", "O.tminus1", "D.tminus1",
"O.t")],
                              prob = 0)  # initialise probs with 0
## CP diabetes - cross-sectional
Mean.CP.Diab.cs.sam1 <- expand.grid(D.t = c(0, 1), O.t = c(0, 1),
                                    Sex = c(0, 1), Time = Time.sam)
Mean.CP.Diab.cs.sam1 <- cbind(Sim = "mean",
                              Mean.CP.Diab.cs.sam1[, c("Time", "Sex", "O.t", "D.t")],
                              prob = 0)  # initialise probs with 0
## CP diabetes - cross-time
Mean.CP.Diab.ct.sam1 <- expand.grid(D.t = c(0, 1), O.t = c(0, 1),
                                    D.tminus1 = c(0, 1), Sex = c(0, 1),
                                    Time = Time.sam[-1])
Mean.CP.Diab.ct.sam1 <- cbind(Sim = "mean",
                              Mean.CP.Diab.ct.sam1[, c("Time", "Sex", "D.tminus1", "O.t", "D.t")],
                              prob = 0)  # initialise probs with 0
## Transition parameters
Mean.TP.sam1 <- expand.grid(Parameter = c(paste0("a", 0:7), paste0("b", 0:7)),
                            Time = Time.sam[-1])
Mean.TP.sam1 <- cbind(Sim = "mean",
                      Mean.TP.sam1[, c("Time", "Parameter")],
                      value = 0)  # initialise values with 0

### (d) Functions ------------------------------------------------------------

#### (i) SampleV function ----------------------------------------------------

# samplev() function
# efficient implementation of the rMultinom() function of the Hmisc package
# from Krijkamp et al (2018)
samplev <- function(probs, m) {
  d <- dim(probs)  # (dimensions of probability matrix)
  n <- d[1]  # (number of rows, i.e. individuals)
  k <- d[2]  # (number of columns, i.e. states)
  lev <- dimnames(probs)[[2]]  # (names of columns, i.e. state values)
  if (!length(lev))
    lev <- 1:k
  ran <- matrix(lev[1], ncol = m, nrow = n)
  U <- t(probs)
  for(i in 2:k) {
    U[i, ] <- U[i, ] + U[i - 1, ]
  }
  if (any((U[k, ] - 1) > 1e-05))
    stop("error in multinom: probabilities do not sum to 1")

  for (j in 1:m) {
    un <- rep(runif(n), rep(k, n))
    ran[, j] <- lev[1 + colSums(un > U)]
  }
  ran
}

#### (ii) Calculate prevalence proportions -----------------------------------

# Calculate Obesity prevalence
# args: group = subgroup, freqtable = frequency table (numeric)
# returns single number (prevalence)
CalculatePrevObesityT <- function (group, freqtable) {

  if (group == "overall") {
```

```
    prevalence <- sum(subset(freqtable, Sim == s & Time == (t-1) & O.t == 1)$freq) /
      N.i.sam
    return(prevalence)

  } else if (group == "female") {

    prevalence <- sum(subset(freqtable, Sim == s & Time == (t-1) & Sex == 0 & O.t == 1)$freq) /
      sum(subset(freqtable, Sim == s & Time == (t-1) & Sex == 0)$freq)
    return(prevalence)

  } else if (group == "male") {

    prevalence <- sum(subset(freqtable, Sim == s & Time == (t-1) & Sex == 1 & O.t == 1)$freq) /
      sum(subset(freqtable, Sim == s & Time == (t-1) & Sex == 1)$freq)
    return(prevalence)

  }

}  # (close function loop)


# Calculate Diabetes prevalence
# args: group = subgroup, freqtable = frequency table (numeric)
# returns single number (prevalence)
CalculatePrevDiabetesT <- function (group, freqtable) {

  if (group == "overall") {

    prevalence <- sum(subset(freqtable, Sim == s & Time == (t-1) & D.t == 1)$freq) /
      N.i.sam
    return(prevalence)

  } else if (group == "female") {

    prevalence <- sum(subset(freqtable, Sim == s & Time == (t-1) & Sex == 0 & D.t == 1)$freq) /
      sum(subset(freqtable, Sim == s & Time == (t-1) & Sex == 0)$freq)
    return(prevalence)

  } else if (group == "male") {

    prevalence <- sum(subset(freqtable, Sim == s & Time == (t-1) & Sex == 1 & D.t == 1)$freq) /
      sum(subset(freqtable, Sim == s & Time == (t-1) & Sex == 1)$freq)
    return(prevalence)

  }

}  # (close function loop)

#### (iii) Calculate conditional probabilities ----------------------------------

# Calculate conditional probability table at time t
# args: dv = dependent variable, iv = independent variable(s), dataset = data frame (factorised)
# returns conditional probability table (cprob.t)
CalculateCPT <- function(dv, iv, dataset) {

  # Define formula for use in cpt function (from HydeNet package)
  formula <- as.formula(paste(dv, paste(iv, collapse = " + "), sep = " ~ "))

  # Create conditional probability table
  cprob.t <- cbind(Time = (t-1), am_adt(cpt(formula, data = dataset)))

  return(cprob.t)

}

# Function for converting multidimensional arrays to tables
# (from https://github.com/Rdatatable/data.table/issues/1418)
am_adt <- function(inarray) {
  if (!is.array(inarray)) stop("input must be an array")
  dims <- dim(inarray)
  if (is.null(dimnames(inarray))) {
    inarray <- provideDimnames(inarray, base = list(as.character(seq_len(max(dims)))))
  }
  FT <- if (any(class(inarray) %in% "ftable")) inarray else ftable(inarray)
  out <- data.table(as.table(ftable(FT)))
  nam <- names(out)[seq_along(dims)]
  setorderv(out[, (nam) := lapply(.SD, type.convert), .SDcols = nam], nam)[]
}



################################################################################
## (2) MSM: AUTOCORRELATION STRUCTURE 1 ----------------------------------------

# Set seed
set.seed(101)

### (a) Define conditional probabilities at time t ------------------------------

#### (i) Obesity ----------------------------------------------------------------

# Function to calculate P(Obesity = 1 | Sex, Prev obesity, Prev diabetes) at time t

CalculateProbObesityT <- function(Sex, PrevObes, PrevDiab) {

  # Inc probability (PrevObes = 0): a0 + (a2-a0)*Sex + (a1-a0)*PrevDiab + (a3-a2-a1+a0)*Sex*PrevDiab
  # Prev probability (PrevObes = 1): a4 + (a6-a4)*Sex + (a5-a4)*PrevDiab + (a7-a6-
a5+a4)*Sex*PrevDiab

  p.obes.t <- a0 + (a2-a0)*Sex + (a1-a0)*PrevDiab + (a3-a2-a1+a0)*Sex*PrevDiab +
    PrevObes*((a4-a0) + (a6-a2-(a4-a0))*Sex + (a5-a1-(a4-a0))*PrevDiab +
                (a7-a6-a5-a3+a2+a1+(a4-a0))*Sex*PrevDiab)
```

```
  return(p.obes.t)

}

# (a0,...,a7 will be estimated from sample of population for each simulation run...
# ... using EstimateTransitionProbs function)

#### (ii) Diabetes ----------------------------------------------------------

# Function to calculate P(Diabetes = 1 | Sex, Obesity, Prev diabetes) at time t

CalculateProbDiabetesT <- function(Sex, PrevDiab, Obes) {

  # Inc probability: b0 + (b2-b0)*Sex + (b1-b0)*Obes + (b3-b2-b1+b0)*Sex*Obes
  # Prev probability: b4 + (b6-b4)*Sex + (b5-b4)*Obes + (b7-b6-b5+b4)*Sex*Obes

  p.diab.t <- b0 + (b2-b0)*Sex + (b1-b0)*Obes + (b3-b2-b1+b0)*Sex*Obes +
    PrevDiab*((b4-b0) + (b6-b2-(b4-b0))*Sex + (b5-b1-(b4-b0))*Obes +
                (b7-b6-b5-b3+b2+b1+(b4-b0))*Sex*Obes)
  return(p.diab.t)

}

# (b0,...,b7 will be estimated from sample of population for each simulation run...
# ... using EstimateTransitionProbs function)

#### (iii) Calculate parameters governing transition probabilities -------------

# Function to estimate transition probabilities from a sample of individuals for time t
# (will change based on autocorrelation structure)
# args: sampledata = individual-level sample dataset (numeric)

EstimateTransitionProbs <- function(sampledata) {

  # Create dataframe for sample data (baseline, time t-1, & time t)
  vars <- c("Sex", paste0(c("O.t", "D.t"), (t-2)), paste0(c("O.t", "D.t"), (t-1)))
  sampledata.t <- data.frame(cbind(Time = (t-1), sampledata[, vars]))
  names(sampledata.t) <- c("Time", "Sex", "O.tminus1", "D.tminus1", "O.t", "D.t")
  sampledata.t[, -1] <- data.frame(apply(sampledata.t[, -1], 2, factor))

  # Calculate cross-time conditional probabilities & define transition parameters
  # (1) Obesity
  var.d <- "O.t"  # define dependent variable
  var.i <- c("Sex", "O.tminus1", "D.tminus1")  # define independent variables
  formula <- as.formula(paste(var.d, paste(var.i, collapse = " + "), sep = " ~ "))
  CP.Obes <- data.frame(am_adt(cpt(formula, data = sampledata.t)))
  CP.Obes <- rename(CP.Obes, replace = c("N" = "prob"))  # rename prob column
  CP.Obes <- subset(CP.Obes, O.t == "1")  # remove 'complement' rows
  CP.Obes <- subset(CP.Obes, select = -O.t)  # remove O.t column
  a0 <<- CP.Obes[CP.Obes[, "O.tminus1"] == 0 &
                   CP.Obes[, "Sex"] == 0 &
                   CP.Obes[, "D.tminus1"] == 0, "prob"]
  a1 <<- CP.Obes[CP.Obes[, "O.tminus1"] == 0 &
                   CP.Obes[, "Sex"] == 0 &
                   CP.Obes[, "D.tminus1"] == 1, "prob"]
  a2 <<- CP.Obes[CP.Obes[, "O.tminus1"] == 0 &
                   CP.Obes[, "Sex"] == 1 &
                   CP.Obes[, "D.tminus1"] == 0, "prob"]
  a3 <<- CP.Obes[CP.Obes[, "O.tminus1"] == 0 &
                   CP.Obes[, "Sex"] == 1 &
                   CP.Obes[, "D.tminus1"] == 1, "prob"]
  a4 <<- CP.Obes[CP.Obes[, "O.tminus1"] == 1 &
                   CP.Obes[, "Sex"] == 0 &
                   CP.Obes[, "D.tminus1"] == 0, "prob"]
  a5 <<- CP.Obes[CP.Obes[, "O.tminus1"] == 1 &
                   CP.Obes[, "Sex"] == 0 &
                   CP.Obes[, "D.tminus1"] == 1, "prob"]
  a6 <<- CP.Obes[CP.Obes[, "O.tminus1"] == 1 &
                   CP.Obes[, "Sex"] == 1 &
                   CP.Obes[, "D.tminus1"] == 0, "prob"]
  a7 <<- CP.Obes[CP.Obes[, "O.tminus1"] == 1 &
                   CP.Obes[, "Sex"] == 1 &
                   CP.Obes[, "D.tminus1"] == 1, "prob"]
  # (2) Diabetes
  var.d <- "D.t"  # define dependent variable
  var.i <- c("Sex", "D.tminus1", "O.t")  # define independent variables
  formula <- as.formula(paste(var.d, paste(var.i, collapse = " + "), sep = " ~ "))
  CP.Diab <- data.frame(am_adt(cpt(formula, data = sampledata.t)))
  CP.Diab <- rename(CP.Diab, replace = c("N" = "prob"))  # rename prob column
  CP.Diab <- subset(CP.Diab, D.t == "1")  # remove 'complement' rows
  CP.Diab <- subset(CP.Diab, select = -D.t)  # remove D.t column
  b0 <<- CP.Diab[CP.Diab[, "D.tminus1"] == 0 &
                   CP.Diab[, "Sex"] == 0 &
                   CP.Diab[, "O.t"] == 0, "prob"]
  b1 <<- CP.Diab[CP.Diab[, "D.tminus1"] == 0 &
                   CP.Diab[, "Sex"] == 0 &
                   CP.Diab[, "O.t"] == 1, "prob"]
  b2 <<- CP.Diab[CP.Diab[, "D.tminus1"] == 0 &
                   CP.Diab[, "Sex"] == 1 &
                   CP.Diab[, "O.t"] == 0, "prob"]
  b3 <<- CP.Diab[CP.Diab[, "D.tminus1"] == 0 &
                   CP.Diab[, "Sex"] == 1 &
                   CP.Diab[, "O.t"] == 1, "prob"]
  b3 <<- CP.Diab[CP.Diab[, "D.tminus1"] == 0 &
                   CP.Diab[, "Sex"] == 1 &
                   CP.Diab[, "O.t"] == 1, "prob"]
  b4 <<- CP.Diab[CP.Diab[, "D.tminus1"] == 1 &
                   CP.Diab[, "Sex"] == 0 &
                   CP.Diab[, "O.t"] == 0, "prob"]
  b5 <<- CP.Diab[CP.Diab[, "D.tminus1"] == 1 &
                   CP.Diab[, "Sex"] == 0 &
                   CP.Diab[, "O.t"] == 1, "prob"]
  b6 <<- CP.Diab[CP.Diab[, "D.tminus1"] == 1 &
```

```
                    CP.Diab[, "Sex"] == 1 &
                    CP.Diab[, "O.t"] == 0, "prob"]
  b7 <<- CP.Diab[CP.Diab[, "D.tminus1"] == 1 &
                    CP.Diab[, "Sex"] == 1 &
                    CP.Diab[, "O.t"] == 1, "prob"]

}

### (b) Simulation ------------------------------------------------------------

# Initialise obesity (a) and diabetes (b) parameters with 0
a0 <- a1 <- a2 <- a3 <- a4 <- a5 <- a6 <- a7 <- 0
b0 <- b1 <- b2 <- b3 <- b4 <- b5 <- b6 <- b7 <- 0

# Draw N.sim random numbers from N.i.sam
# These represent the random samples that will be drawn from the population
select <- matrix(nrow = N.i.sam, ncol = N.sim,
                 dimnames = list(paste0("ind", 1:N.i.sam),
                                 paste0("sample", 1:N.sim)))
for (s in 1:N.sim) {
  select[, s] <- sample(x = c(1:N.i.pop), size = N.i.sam, replace = FALSE)
}

# Record start time of simulation
v <- Sys.time()

# (1) Loop through natural history & interventions ----------------------------
# (v = 0 represents natural history, v = 1-6 represent interventions 1-6)
for (z in 0:N.int) {

  # Reset summary tables
  # (each intervention (or natural history) has a separate summary table)
  Frequency.cs.sam1 <- Frequency.cs.sam1[0, ]
  Obes.prev.sam1 <- Obes.prev.sam1[0, ]
  Diab.prev.sam1 <- Diab.prev.sam1[0, ]
  CProbability.Obes.cs.sam1 <- CProbability.Obes.cs.sam1[0, ]
  CProbability.Diab.cs.sam1 <- CProbability.Diab.cs.sam1[0, ]
  CProbability.Obes.ct.sam1 <- CProbability.Obes.ct.sam1[0, ]
  CProbability.Diab.ct.sam1 <- CProbability.Diab.ct.sam1[0, ]
  TransitionParameters.sam1 <- TransitionParameters.sam1[0, ]

  # (2) Loop through simulation runs
  for (s in 1:N.sim) {

    # Sample N.i.sam individuals from population
    # Store data in Sample dataframe (all longitudinal vars in order)
    Sample <- Population[select[, s], ]

    # Fill Sex.sim, Obes.sim, & Diab.sim matrices with baseline data
    Sex.sam1[, "Sex"] <- Sample[, "Sex"]
    Obes.sam1[, "O.t0"] <- Sample[, "O.t0"]
    Diab.sam1[, "D.t0"] <- Sample[, "D.t0"]

    # (3) Loop through time points
    for (t in 1:N.t.sam) {

      ## Record summary data at baseline
      if (t == 1) {

        # Record summary data -------------------------------------------------

        # Bind variables from time t and baseline together
        Sample.t <- data.frame(cbind(Sex.sam1[, 1], Obes.sam1[, t], Diab.sam1[, t]))
        vars.cs <- c("Sex", paste0(c("O.t", "D.t"), (t-1)))
        names(Sample.t) <- vars.cs

        # (a) Cross-sectional frequency table ---------------------------------

        freq.t <- cbind(Sim = s, Time = (t-1), count(Sample.t[, vars.cs]))
        names(freq.t) <- names(Frequency.cs.sam1)  # rename columns to match Frequency table
        Frequency.cs.sam1 <- rbind(Frequency.cs.sam1, freq.t)

        # (b) Prevalence ------------------------------------------------------

        ## Obesity
        prev.O <- cbind.data.frame(Sim = s, Time = (t-1), Subgroup = "Obes.prev",
                                   prev = CalculatePrevObesityT("overall", Frequency.cs.sam1))
        prev.O.f <- cbind.data.frame(Sim = s, Time = (t-1), Subgroup = "Obes.prev.f",
                                     prev = CalculatePrevObesityT("female", Frequency.cs.sam1))
        prev.O.m <- cbind.data.frame(Sim = s, Time = (t-1), Subgroup = "Obes.prev.m",
                                     prev = CalculatePrevObesityT("male", Frequency.cs.sam1))
        Obes.prev.sam1 <- rbind.data.frame(Obes.prev.sam1, prev.O, prev.O.f, prev.O.m)

        ## Diabetes
        prev.D <- cbind.data.frame(Sim = s, Time = (t-1), Subgroup = "Diab.prev",
                                   prev = CalculatePrevDiabetesT("overall", Frequency.cs.sam1))
        prev.D.f <- cbind.data.frame(Sim = s, Time = (t-1), Subgroup = "Diab.prev.f",
                                     prev = CalculatePrevDiabetesT("female", Frequency.cs.sam1))
        prev.D.m <- cbind.data.frame(Sim = s, Time = (t-1), Subgroup = "Diab.prev.m",
                                     prev = CalculatePrevDiabetesT("male", Frequency.cs.sam1))
        Diab.prev.sam1 <- rbind.data.frame(Diab.prev.sam1, prev.D, prev.D.f, prev.D.m)

        # (c) Conditional probabilities ---------------------------------------

        # Convert variables in Sample.t dataset to factors
        # (requred for calculating conditonal probabilties)
        Sample.t <- data.frame(lapply(Sample.t, factor, levels = c("0", "1")))

        ## (i) Cross-sectional -----------------------------------------------

        ## Obesity
        var.d <- paste0("O.t", (t-1))  # (define dependent variable)
        var.i <- "Sex"  # (define independent variable)
```

```
        cprob.t <- cbind(Sim = s, CalculateCPT(dv = var.d, iv = var.i, dataset = Sample.t))
        names(cprob.t) <- names(CProbability.Obes.cs.sam1)
        CProbability.Obes.cs.sam1 <- rbind.data.frame(CProbability.Obes.cs.sam1, cprob.t)

        ## Diabetes
        var.d <- paste0("D.t", (t-1))
        var.i <- c("Sex", paste0("O.t", (t-1)))
        cprob.t <- cbind(Sim = s, CalculateCPT(dv = var.d, iv = var.i, dataset = Sample.t))
        names(cprob.t) <- names(CProbability.Diab.cs.sam1)
        CProbability.Diab.cs.sam1 <- rbind(CProbability.Diab.cs.sam1, cprob.t)


      } else {  ## Estimate transition probs, update time-varying characteristics, & record summary
data at time t

        # Estimate transition probabilities for time t -------------------------

        # Estimate transition probabilities (for natural history) using Sample dataframe
        EstimateTransitionProbs(sampledata = Sample)

        # Define transition probabilities for obesity under Interventions 1-6
        if (z != 0) {
          if (z == 1) {  # under Intervention 1, no individuals may be obese
            a0 <- a1 <- a2 <- a3 <- a4 <- a5 <- a6 <- a7 <- 0
          } else if (z == 2) {  # under Intervention 2, all individuals are obese
            a0 <- a1 <- a2 <- a3 <- a4 <- a5 <- a6 <- a7 <- 1
          } else if (z == 3) {  # under Intervention 3, no individuals may become obese
            a0 <- a1 <- a2 <- a3 <- 0
          } else if (z == 4) {  # under Intervention 4, reduce incident prob of obesity by 15%
            a0 <- 0.85*a0; a1 <- 0.85*a1; a2 <- 0.85*a2; a3 <- 0.85*a3
          } else if (z == 5) {  # under Intervention 5, reduce prevalent prob of obesity by 10%
            a4 <- 0.9*a4; a5 <- 0.9*a5; a6 <- 0.9*a6; a7 <- 0.9*a7
          } else if (z == 6) {  # under Intervention 6, reduce incident prob of obesity by 15% &
prevalent prob of obesity by 10%
            a0 <- 0.85*a0; a1 <- 0.85*a1; a2 <- 0.85*a2; a3 <- 0.85*a3
            a4 <- 0.9*a4; a5 <- 0.9*a5; a6 <- 0.9*a6; a7 <- 0.9*a7
          }
        }

        # Record parameters governing transition probabilities
        par.t <- cbind.data.frame(Sim = s, Time = (t-1),
                        Parameter = c(paste0("a", 0:7), paste0("b", 0:7)),
                        value = c(a0, a1, a2, a3, a4, a5, a6, a7, b0, b1, b2, b3, b4, b5, b6, b7))
        TransitionParameters.sam1 <- rbind.data.frame(TransitionParameters.sam1, par.t)

        # Update time-varying characteristics ----------------------------------

        # (a) Obesity ----------------------------------------------------------
        p.obes.t <- cbind(1 - CalculateProbObesityT(Sex = Sex.sam1[, 1],
                                                    PrevObes = Obes.sam1[, (t-1)],
                                                    PrevDiab = Diab.sam1[, (t-1)]),
                          CalculateProbObesityT(Sex = Sex.sam1[, 1],
                                                PrevObes = Obes.sam1[, (t-1)],
                                                PrevDiab = Diab.sam1[, (t-1)]))
        Obes.sam1[, t] <- samplev(probs = p.obes.t, m = 1)
        Obes.sam1[, t] <- Obes.sam1[, t] - 1  # (factor levels should be 0 and 1)

        # (b) Diabetes ---------------------------------------------------------
        p.diab.t <- cbind(1 - CalculateProbDiabetesT(Sex = Sex.sam1[, 1],
                                                     PrevDiab = Diab.sam1[, (t-1)],
                                                     Obes = Obes.sam1[, t]),
                          CalculateProbDiabetesT(Sex = Sex.sam1[, 1],
                                                 PrevDiab = Diab.sam1[, (t-1)],
                                                 Obes = Obes.sam1[, t]))
        Diab.sam1[, t] <- samplev(probs = p.diab.t, m = 1)
        Diab.sam1[, t] <- Diab.sam1[, t] - 1  # (factor levels should be 0 and 1)

        # Record summary data --------------------------------------------------

        # Bind variables from time t, time t-1, and baseline together
        Sample.t <- data.frame(cbind(Sex.sam1[, 1], Obes.sam1[, (t-1)], Diab.sam1[, (t-1)]),
                               Obes.sam1[, t], Diab.sam1[, t])
        vars.cs <- c("Sex", paste0(c("O.t", "D.t"), (t-1)))
        vars.ct <- c("Sex", paste0(c("O.t", "D.t"), (t-2)), paste0(c("O.t", "D.t"), (t-1)))
        names(Sample.t) <- vars.ct

        # (a) Cross-sectional frequency table ----------------------------------
        freq.t <- cbind(Sim = s, Time = (t-1), count(Sample.t[, vars.cs]))
        names(freq.t) <- names(Frequency.cs.sam1)
        Frequency.cs.sam1 <- rbind(Frequency.cs.sam1, freq.t)

        # (b) Prevalence -------------------------------------------------------

        ## Obesity
        prev.O <- cbind.data.frame(Sim = s, Time = (t-1), Subgroup = "Obes.prev",
                                   prev = CalculatePrevObesityT("overall", Frequency.cs.sam1))
        prev.O.f <- cbind.data.frame(Sim = s, Time = (t-1), Subgroup = "Obes.prev.f",
                                     prev = CalculatePrevObesityT("female", Frequency.cs.sam1))
        prev.O.m <- cbind.data.frame(Sim = s, Time = (t-1), Subgroup = "Obes.prev.m",
                                     prev = CalculatePrevObesityT("male", Frequency.cs.sam1))
        Obes.prev.sam1 <- rbind.data.frame(Obes.prev.sam1, prev.O, prev.O.f, prev.O.m)

        ## Diabetes
        prev.D <- cbind.data.frame(Sim = s, Time = (t-1), Subgroup = "Diab.prev",
                                   prev = CalculatePrevDiabetesT("overall", Frequency.cs.sam1))
        prev.D.f <- cbind.data.frame(Sim = s, Time = (t-1), Subgroup = "Diab.prev.f",
                                     prev = CalculatePrevDiabetesT("female", Frequency.cs.sam1))
        prev.D.m <- cbind.data.frame(Sim = s, Time = (t-1), Subgroup = "Diab.prev.m",
                                     prev = CalculatePrevDiabetesT("male", Frequency.cs.sam1))
        Diab.prev.sam1 <- rbind.data.frame(Diab.prev.sam1, prev.D, prev.D.f, prev.D.m)

        # (c) Conditional probabilities ----------------------------------------

```

```
        # Convert variables in Sample.t dataset to factors
        # (requred for calculating conditonal probabilties)
        Sample.t <- data.frame(lapply(Sample.t, factor, levels = c("0", "1")))

        ## (i) Cross-sectional ----------------------------------------------------

        ## Obesity
        var.d <- paste0("O.t", (t-1))  # (define dependent variable)
        var.i <- "Sex"  # (define independent variable)
        cprob.t <- cbind(Sim = s, CalculateCPT(dv = var.d, iv = var.i, dataset = Sample.t))
        names(cprob.t) <- names(CProbability.Obes.cs.sam1)
        CProbability.Obes.cs.sam1 <- rbind.data.frame(CProbability.Obes.cs.sam1, cprob.t)

        ## Diabetes
        var.d <- paste0("D.t", (t-1))
        var.i <- c("Sex", paste0("O.t", (t-1)))
        cprob.t <- cbind(Sim = s, CalculateCPT(dv = var.d, iv = var.i, dataset = Sample.t))
        names(cprob.t) <- names(CProbability.Diab.cs.sam1)
        CProbability.Diab.cs.sam1 <- rbind(CProbability.Diab.cs.sam1, cprob.t)

        ## (ii) Cross-time --------------------------------------------------------

        ## Obesity
        var.d <- paste0("O.t", (t-1))
        var.i <- c("Sex", paste0(c("O.t", "D.t"), (t-2)))
        cprob.t <- cbind(Sim = s, CalculateCPT(dv = var.d, iv = var.i, dataset = Sample.t))
        names(cprob.t) <- names(CProbability.Obes.ct.sam1)
        CProbability.Obes.ct.sam1 <- rbind.data.frame(CProbability.Obes.ct.sam1, cprob.t)

        ## Diabetes
        var.d <- paste0("D.t", (t-1))
        var.i <- c("Sex", paste0("D.t", (t-2)), paste0("O.t", (t-1)))
        cprob.t <- cbind(Sim = s, CalculateCPT(dv = var.d, iv = var.i, dataset = Sample.t))
        names(cprob.t) <- names(CProbability.Diab.ct.sam1)
        CProbability.Diab.ct.sam1 <- rbind.data.frame(CProbability.Diab.ct.sam1, cprob.t)

      }

      # Display progress of simulation
      cat('\r', paste(round((t / N.t.sam * 100), 0),
                      "% done of simulation", s, "of", N.sim, "of Intervention", z, "of", N.int, "
", sep = " "))

    }  # (close time loop - 3)

  }  # (close simulation run loop - 2)

  # Calculate mean trends for Intervention z -------------------------------------

  # (a) Cross-sectional frequencies ----------------------------------------------

  # Use expand.grid function to create full frequency table
  # (Frequency table doesn't show combinations with empty cells)
  f <- expand.grid(D.t = c(0, 1), O.t = c(0, 1), Sex = c(0, 1), Time = Time.sam,
                   Sim = seq(from = 1, to = N.sim, by = 1))
  f <- cbind(f[, c("Sim", "Time", "Sex", "O.t", "D.t")], freq = 0)  # initialise frequencies with 0
  # Fill f with data from (incomplete) Frequency table
  for (i in 1:nrow(Frequency.cs.sam1)) {

    sim <- Frequency.cs.sam1[i, "Sim"]
    time <- Frequency.cs.sam1[i, "Time"]
    sex <- Frequency.cs.sam1[i, "Sex"]
    o.t <- Frequency.cs.sam1[i, "O.t"]
    d.t <- Frequency.cs.sam1[i, "D.t"]
    freq <- Frequency.cs.sam1[i, "freq"]

    f[f[, "Sim"] == sim &
        f[, "Time"] == time &
        f[, "Sex"] == sex &
        f[, "O.t"] == o.t &
        f[, "D.t"] == d.t,
      "freq"] <- freq

  }

  # Overwrite incomplete Frequency table
  Frequency.cs.sam1 <- f; rm(f)

  # Calculate mean frequency at each time
  for (i in 1:nrow(Mean.frequency.sam1)) {

    time <- Mean.frequency.sam1[i, "Time"]
    sex <- Mean.frequency.sam1[i, "Sex"]
    o.t <- Mean.frequency.sam1[i, "O.t"]
    d.t <- Mean.frequency.sam1[i, "D.t"]

    avg <- mean(subset(Frequency.cs.sam1, Time == time &
                         Sex == sex &
                         O.t == o.t &
                         D.t == d.t)$freq)

    Mean.frequency.sam1[Mean.frequency.sam1[, "Time"] == time &
                          Mean.frequency.sam1[, "Sex"] == sex &
                          Mean.frequency.sam1[, "O.t"] == o.t &
                          Mean.frequency.sam1[, "D.t"] == d.t,
                        "freq"] <- avg

  }

  # (b) Prevalence ---------------------------------------------------------------

  # Calculate mean obesity prevalence at each time
  for (i in 1:nrow(Mean.obes.prev.sam1)) {
```

```
  time <- Mean.obes.prev.sam1[i, "Time"]
  sub <- Mean.obes.prev.sam1[i, "Subgroup"]

  avg <- mean(subset(Obes.prev.sam1, Time == time & Subgroup == sub)$prev)

  Mean.obes.prev.sam1[Mean.obes.prev.sam1[, "Time"] == time &
                        Mean.obes.prev.sam1[, "Subgroup"] == sub,
                      "prev"] <- avg

}

# Calculate mean diabetes prevalence at each time
for (i in 1:nrow(Mean.diab.prev.sam1)) {

  time <- Mean.diab.prev.sam1[i, "Time"]
  sub <- Mean.diab.prev.sam1[i, "Subgroup"]

  avg <- mean(subset(Diab.prev.sam1, Time == time & Subgroup == sub)$prev)

  Mean.diab.prev.sam1[Mean.diab.prev.sam1[, "Time"] == time &
                        Mean.diab.prev.sam1[, "Subgroup"] == sub,
                      "prev"] <- avg

}

# (c) Conditional probabilities ------------------------------------------------

## (i) Cross-sectional ---------------------------------------------------------

# Calculate mean CP of obesity at each time point
for (i in 1:nrow(Mean.CP.Obes.cs.sam1)) {

  time <- Mean.CP.Obes.cs.sam1[i, "Time"]
  sex <- Mean.CP.Obes.cs.sam1[i, "Sex"]
  o.t <- Mean.CP.Obes.cs.sam1[i, "O.t"]

  avg <- mean(subset(CProbability.Obes.cs.sam1, Time == time &
                       Sex == sex &
                       O.t == o.t)$prob)

  Mean.CP.Obes.cs.sam1[Mean.CP.Obes.cs.sam1[, "Time"] == time &
                         Mean.CP.Obes.cs.sam1[, "Sex"] == sex &
                         Mean.CP.Obes.cs.sam1[, "O.t"] == o.t,
                       "prob"] <- avg

}

# Calculate mean CP of diabetes at each time point
for (i in 1:nrow(Mean.CP.Diab.cs.sam1)) {

  time <- Mean.CP.Diab.cs.sam1[i, "Time"]
  sex <- Mean.CP.Diab.cs.sam1[i, "Sex"]
  o.t <- Mean.CP.Diab.cs.sam1[i, "O.t"]
  d.t <- Mean.CP.Diab.cs.sam1[i, "D.t"]

  avg <- mean(subset(CProbability.Diab.cs.sam1, Time == time &
                       Sex == sex &
                       O.t == o.t &
                       D.t == d.t)$prob)

  Mean.CP.Diab.cs.sam1[Mean.CP.Diab.cs.sam1[, "Time"] == time &
                         Mean.CP.Diab.cs.sam1[, "Sex"] == sex &
                         Mean.CP.Diab.cs.sam1[, "O.t"] == o.t &
                         Mean.CP.Diab.cs.sam1[, "D.t"]== d.t,
                       "prob"] <- avg

}

## (ii) Cross-time -------------------------------------------------------------

# Calculate mean CP of obesity at each time point
for (i in 1:nrow(Mean.CP.Obes.ct.sam1)) {

  time <- Mean.CP.Obes.ct.sam1[i, "Time"]
  sex <- Mean.CP.Obes.ct.sam1[i, "Sex"]
  o.tminus1 <- Mean.CP.Obes.ct.sam1[i, "O.tminus1"]
  d.tminus1 <- Mean.CP.Obes.ct.sam1[i, "D.tminus1"]
  o.t <- Mean.CP.Obes.ct.sam1[i, "O.t"]

  avg <- mean(subset(CProbability.Obes.ct.sam1, Time == time &
                       Sex == sex &
                       O.tminus1 == o.tminus1 &
                       D.tminus1 == d.tminus1 &
                       O.t == o.t)$prob)

  Mean.CP.Obes.ct.sam1[Mean.CP.Obes.ct.sam1[, "Time"] == time &
                         Mean.CP.Obes.ct.sam1[, "Sex"] == sex &
                         Mean.CP.Obes.ct.sam1[, "O.tminus1"] == o.tminus1 &
                         Mean.CP.Obes.ct.sam1[, "D.tminus1"] == d.tminus1 &
                         Mean.CP.Obes.ct.sam1[, "O.t"] == o.t,
                       "prob"] <- avg

}

# Calculate mean CP of diabetes at each time point
for (i in 1:nrow(Mean.CP.Diab.ct.sam1)) {

  time <- Mean.CP.Diab.ct.sam1[i, "Time"]
  sex <- Mean.CP.Diab.ct.sam1[i, "Sex"]
  d.tminus1 <- Mean.CP.Diab.ct.sam1[i, "D.tminus1"]
  o.t <- Mean.CP.Diab.ct.sam1[i, "O.t"]
  d.t <- Mean.CP.Diab.ct.sam1[i, "D.t"]
```

```
817
818       avg <- mean(subset(CProbability.Diab.ct.sam1, Time == time &
819                            Sex == sex &
820                            D.tminus1 == d.tminus1 &
821                            O.t == o.t &
822                            D.t == d.t)$prob)
823
824       Mean.CP.Diab.ct.sam1[Mean.CP.Diab.ct.sam1[, "Time"] == time &
825                              Mean.CP.Diab.ct.sam1[, "Sex"] == sex &
826                              Mean.CP.Diab.ct.sam1[, "D.tminus1"] == d.tminus1 &
827                              Mean.CP.Diab.ct.sam1[, "O.t"] == o.t &
828                              Mean.CP.Diab.ct.sam1[, "D.t"] == d.t,
829                            "prob"] <- avg
830
831     }
832
833     # (d) Transition parameters ------------------------------------------------
834
835     # Calculate mean TP at each time point
836     for (i in 1:nrow(Mean.TP.sam1)) {
837
838       time <- Mean.TP.sam1[i, "Time"]
839       par <- Mean.TP.sam1[i, "Parameter"]
840
841       avg <- mean(subset(TransitionParameters.sam1, Time == time & Parameter == par)$value)
842
843       Mean.TP.sam1[Mean.TP.sam1[, "Time"] == time &
844                      Mean.TP.sam1[, "Parameter"] == par,
845                    "value"] <- avg
846
847     }
848
849     # Save summary tables for Intervention z -----------------------------------
850
851     # Define file location
852     if (z == 0) {  # (natural history)
853       path <- paste0("./Microsimulation models/Time t transition probs/MSM 1/Natural history/")
854     } else {  # (intervention z)
855       path <- paste0("./Microsimulation models/Time t transition probs/MSM 1/Intervention ", z, "/")
856     }
857
858     # Define file names
859     if (z == 0) {  # (natural history)
860       file.freq <- paste0("Sam1Freq.csv")  # frequency
861       file.prev.o <- paste0("Sam1ObesPrev.csv")  # obesity prevalence
862       file.prev.d <- paste0("Sam1DiabPrev.csv")  # diabetes prevalence
863       file.cpcs.o <- paste0("Sam1ObesCPcs.csv")  # CP obesity (cross-sectional)
864       file.cpct.o <- paste0("Sam1ObesCPct.csv")  # CP obesity (cross-time)
865       file.cpcs.d <- paste0("Sam1DiabCPcs.csv")  # CP diabetes (cross-sectional)
866       file.cpct.d <- paste0("Sam1DiabCPct.csv")  # CP diabetes (cross-time)
867       file.tp <- paste0("Sam1TP.csv")  # transition parameters
868       file.m.freq <- paste0("Sam1FreqMean.csv")  # mean frequency
869       file.m.prev.o <- paste0("Sam1ObesPrevMean.csv")  # mean obesity prevalence
870       file.m.prev.d <- paste0("Sam1DiabPrevMean.csv")  # mean diabetes prevalence
871       file.m.cpcs.o <- paste0("Sam1ObesCPcsMean.csv")  # mean CP obesity (cross-sectional)
872       file.m.cpct.o <- paste0("Sam1ObesCPctMean.csv")  # mean CP obesity (cross-time)
873       file.m.cpcs.d <- paste0("Sam1DiabCPcsMean.csv")  # mean CP diabetes (cross-sectional)
874       file.m.cpct.d <- paste0("Sam1DiabCPctMean.csv")  # mean CP diabetes (cross-time)
875       file.m.tp <- paste0("Sam1TPMean.csv")  # mean transition parameters
876     } else {  # (intervention z)
877       file.freq <- paste0("Sam1FreqInt", z, ".csv")  # frequency
878       file.prev.o <- paste0("Sam1ObesPrevInt", z, ".csv")  # obesity prevalence
879       file.prev.d <- paste0("Sam1DiabPrevInt", z, ".csv")  # diabetes prevalence
880       file.cpcs.o <- paste0("Sam1ObesCPcsInt", z, ".csv")  # CP obesity (cross-sectional)
881       file.cpct.o <- paste0("Sam1ObesCPctInt", z, ".csv")  # CP obesity (cross-time)
882       file.cpcs.d <- paste0("Sam1DiabCPcsInt", z, ".csv")  # CP diabetes (cross-sectional)
883       file.cpct.d <- paste0("Sam1DiabCPctInt", z, ".csv")  # CP diabetes (cross-time)
884       file.tp <- paste0("Sam1TPInt", z, ".csv")  # transition parameters
885       file.m.freq <- paste0("Sam1FreqMeanInt", z, ".csv")  # mean frequency
886       file.m.prev.o <- paste0("Sam1ObesPrevMeanInt", z, ".csv")  # mean obesity prevalence
887       file.m.prev.d <- paste0("Sam1DiabPrevMeanInt", z, ".csv")  # mean diabetes prevalence
888       file.m.cpcs.o <- paste0("Sam1ObesCPcsMeanInt", z, ".csv")  # mean CP obesity (cross-sectional)
889       file.m.cpct.o <- paste0("Sam1ObesCPctMeanInt", z, ".csv")  # mean CP obesity (cross-time)
890       file.m.cpcs.d <- paste0("Sam1DiabCPcsMeanInt", z, ".csv")  # mean CP diabetes (cross-sectional)
891       file.m.cpct.d <- paste0("Sam1DiabCPctMeanInt", z, ".csv")  # mean CP diabetes (cross-time)
892       file.m.tp <- paste0("Sam1TPMeanInt", z, ".csv")  # mean transition parameters
893     }
894
895     # Export frequency table
896     write.csv(Frequency.cs.sam1, file = paste0(path, file.freq), row.names = FALSE)
897
898     # Export prevalence tables
899     write.csv(Obes.prev.sam1, file = paste0(path, file.prev.o), row.names = FALSE)
900     write.csv(Diab.prev.sam1, file = paste0(path, file.prev.d), row.names = FALSE)
901
902     # Export conditional probability tables
903     write.csv(CProbability.Obes.cs.sam1, file = paste0(path, file.cpcs.o), row.names = FALSE)
904     write.csv(CProbability.Obes.ct.sam1, file = paste0(path, file.cpct.o), row.names = FALSE)
905     write.csv(CProbability.Diab.cs.sam1, file = paste0(path, file.cpcs.d), row.names = FALSE)
906     write.csv(CProbability.Diab.ct.sam1, file = paste0(path, file.cpct.d), row.names = FALSE)
907
908     # Export transition parameter tables
909     write.csv(TransitionParameters.sam1, file = paste0(path, file.tp), row.names = FALSE)
910
911     # Export mean trend tables
912     write.csv(Mean.frequency.sam1, file = paste0(path, file.m.freq), row.names = FALSE)
913     write.csv(Mean.obes.prev.sam1, file = paste0(path, file.m.prev.o), row.names = FALSE)
914     write.csv(Mean.diab.prev.sam1, file = paste0(path, file.m.prev.d), row.names = FALSE)
915     write.csv(Mean.CP.Obes.cs.sam1, file = paste0(path, file.m.cpcs.o), row.names = FALSE)
916     write.csv(Mean.CP.Obes.ct.sam1, file = paste0(path, file.m.cpct.o), row.names = FALSE)
917     write.csv(Mean.CP.Diab.cs.sam1, file = paste0(path, file.m.cpcs.d), row.names = FALSE)
918     write.csv(Mean.CP.Diab.ct.sam1, file = paste0(path, file.m.cpct.d), row.names = FALSE)
919     write.csv(Mean.TP.sam1, file = paste0(path, file.m.tp), row.names = FALSE)
```

```
}  # (close intervention loop - 1)

comp.time <- Sys.time() - v; comp.time  # print total simulation time
# (~39 seconds per simulation run of 20000 individuals)
# (~37 mins per 100 simulation runs of 20000 individuals)
```

Note that the above code relates to AS1 (i.e. the true data-generating process of the population); for all other autocorrelation structures, the function which estimates the transition probabilities at time *t* from a sample of individuals from the population (lines 334 – 411) changes based on the autocorrelation structure that is modelled. For AS2, this function is:

```
EstimateTransitionProbs <- function(sampledata) {

  # Create dataframe for sample data (baseline, time t-1, & time t)
  vars <- c("Sex", paste0(c("O.t", "D.t"), (t-2)), paste0(c("O.t", "D.t"), (t-1)))  # define
variables
  sampledata.t <- data.frame(cbind(Time = (t-1), sampledata[, vars]))
  names(sampledata.t) <- c("Time", "Sex", "O.tminus1", "D.tminus1", "O.t", "D.t")  # rename
variables
  sampledata.t[, -1] <- data.frame(apply(sampledata.t[, -1], 2, factor))  # convert vars to factors

  # Calculate cross-time conditional probabilities & define transition parameters
  # (1) Obesity
  var.d <- "O.t"  # define dependent variable
  var.i <- c("Sex", "O.tminus1")  # define independent variables
  formula <- as.formula(paste(var.d, paste(var.i, collapse = " + "), sep = " ~ "))
  CP.Obes <- data.frame(am_adt(cpt(formula, data = sampledata.t)))
  CP.Obes <- rename(CP.Obes, replace = c("N" = "prob"))  # rename prob column
  CP.Obes <- subset(CP.Obes, O.t == "1")  # remove 'complement' rows
  CP.Obes <- subset(CP.Obes, select = -O.t)  # remove O.t column
  a0 <<- CP.Obes[CP.Obes[, "O.tminus1"] == 0 &
                           CP.Obes[, "Sex"] == 0, "prob"]
  a2 <<- CP.Obes[CP.Obes[, "O.tminus1"] == 0 &
                           CP.Obes[, "Sex"] == 1, "prob"]
  a4 <<- CP.Obes[CP.Obes[, "O.tminus1"] == 1 &
                           CP.Obes[, "Sex"] == 0, "prob"]
  a6 <<- CP.Obes[CP.Obes[, "O.tminus1"] == 1 &
                           CP.Obes[, "Sex"] == 1, "prob"]
  # (2) Diabetes
  var.d <- "D.t"  # define dependent variable
  var.i <- c("Sex", "D.tminus1", "O.t")  # define independent variables
  formula <- as.formula(paste(var.d, paste(var.i, collapse = " + "), sep = " ~ "))
  CP.Diab <- data.frame(am_adt(cpt(formula, data = sampledata.t)))
  CP.Diab <- rename(CP.Diab, replace = c("N" = "prob"))  # rename prob column
  CP.Diab <- subset(CP.Diab, D.t == "1")  # remove 'complement' rows
  CP.Diab <- subset(CP.Diab, select = -D.t)  # remove D.t column
  b0 <<- CP.Diab[CP.Diab[, "D.tminus1"] == 0 &
                   CP.Diab[, "Sex"] == 0 &
                   CP.Diab[, "O.t"] == 0, "prob"]
  b1 <<- CP.Diab[CP.Diab[, "D.tminus1"] == 0 &
                   CP.Diab[, "Sex"] == 0 &
                   CP.Diab[, "O.t"] == 1, "prob"]
  b2 <<- CP.Diab[CP.Diab[, "D.tminus1"] == 0 &
                   CP.Diab[, "Sex"] == 1 &
                   CP.Diab[, "O.t"] == 0, "prob"]
  b3 <<- CP.Diab[CP.Diab[, "D.tminus1"] == 0 &
                   CP.Diab[, "Sex"] == 1 &
                   CP.Diab[, "O.t"] == 1, "prob"]
  b4 <<- CP.Diab[CP.Diab[, "D.tminus1"] == 1 &
                   CP.Diab[, "Sex"] == 0 &
                   CP.Diab[, "O.t"] == 0, "prob"]
  b5 <<- CP.Diab[CP.Diab[, "D.tminus1"] == 1 &
                   CP.Diab[, "Sex"] == 0 &
                   CP.Diab[, "O.t"] == 1, "prob"]
  b6 <<- CP.Diab[CP.Diab[, "D.tminus1"] == 1 &
                   CP.Diab[, "Sex"] == 1 &
                   CP.Diab[, "O.t"] == 0, "prob"]
  b7 <<- CP.Diab[CP.Diab[, "D.tminus1"] == 1 &
                   CP.Diab[, "Sex"] == 1 &
                   CP.Diab[, "O.t"] == 1, "prob"]

}
```

For AS3, this function is:

```
EstimateTransitionProbs <- function(sampledata) {

  # Create dataframe for sample data (baseline & time t)
  vars <- c("Sex", paste0(c("O.t", "D.t"), (t-1)))  # define variables
  sampledata.t <- data.frame(cbind(Time = (t-1), sampledata[, vars]))
  names(sampledata.t) <- c("Time", "Sex", "O.t", "D.t")  # rename variables
  sampledata.t[, -1] <- data.frame(apply(sampledata.t[, -1], 2, factor))  # convert vars to factors

  # Calculate cross-sectional conditional probabilities & define transition parameters
  # (1) Obesity
  var.d <- "O.t"  # define dependent variable
  var.i <- "Sex"  # define independent variable
  formula <- as.formula(paste(var.d, paste(var.i, collapse = " + "), sep = " ~ "))
  CP.Obes <- data.frame(am_adt(cpt(formula, data = sampledata.t)))
  CP.Obes <- rename(CP.Obes, replace = c("N" = "prob"))  # rename prob column
  CP.Obes <- subset(CP.Obes, O.t == "1")  # remove 'complement' rows
  CP.Obes <- subset(CP.Obes, select = -O.t)  # remove O.t column
  a0 <<- CP.Obes[CP.Obes[, "Sex"] == 0, "prob"]
  a2 <<- CP.Obes[CP.Obes[, "Sex"] == 1, "prob"]
  # (2) Diabetes
```

```
  var.d <- "D.t"  # define dependent variable
  var.i <- c("Sex", "O.t")  # define independent variables
  formula <- as.formula(paste(var.d, paste(var.i, collapse = " + "), sep = " ~ "))
  CP.Diab <- data.frame(am_adt(cpt(formula, data = sampledata.t)))
  CP.Diab <- rename(CP.Diab, replace = c("N" = "prob"))  # rename prob column
  CP.Diab <- subset(CP.Diab, D.t == "1")  # remove 'complement' rows
  CP.Diab <- subset(CP.Diab, select = -D.t)  # remove D.t column
  b0 <<- CP.Diab[CP.Diab[, "Sex"] == 0 &
                 CP.Diab[, "O.t"] == 0, "prob"]
  b1 <<- CP.Diab[CP.Diab[, "Sex"] == 0 &
                 CP.Diab[, "O.t"] == 1, "prob"]
  b2 <<- CP.Diab[CP.Diab[, "Sex"] == 1 &
                 CP.Diab[, "O.t"] == 0, "prob"]
  b3 <<- CP.Diab[CP.Diab[, "Sex"] == 1 &
                 CP.Diab[, "O.t"] == 1, "prob"]

}
```

The output from each simulation is then saved to its subfolder ('MSM 2' and 'MSM 3', respectively).

C.2.2.2 Microsimulation

Since AS1 represents the true autocorrelation structure of the population, it is expected that using microsimulation to model AS1 will replicate the true natural history of the population and produce unbiased estimates of obesity and diabetes prevalence in the population at all time points. AS1 is also expected to replicate the true counterfactual histories under Interventions 1 through 6, thereby producing unbiased estimates of all intervention effects.

However, no other autocorrelation structures (i.e. neither AS2 nor AS3) are expected to faithfully replicate the natural history of the population because the MSM does not have access to the entire joint distribution of all variables (i.e. all variables across all time periods. Modelling AS2 and AS3 is therefore expected to produce biased estimates of all intervention effects.

C.2.2.2.1 Counterfactual histories under hypothetical intervention

Here, we present the results of using microsimulation to model the counterfactual histories for Interventions 2 through 6 (the results of Intervention 1 are presented in Chapter 6, Section 6.4.2.3.3), according to each of the three autocorrelation structures (AS1 through AS3).

Figures which compare the true effect of each intervention on obesity and diabetes prevalence in the population with those modelled by the g-formula are shown in Figure C.14 (Intervention 2), Figure C.15 (Intervention 3), Figure C.16 (Intervention 4), Figure C.17 (Intervention 5), and Figure C.18 (Intervention 6).

Figure C.14 Counterfactual histories of obesity and diabetes prevalence under Intervention 2 for each of AS1 through AS3 modelled using microsimulation, compared to the true counterfactual history

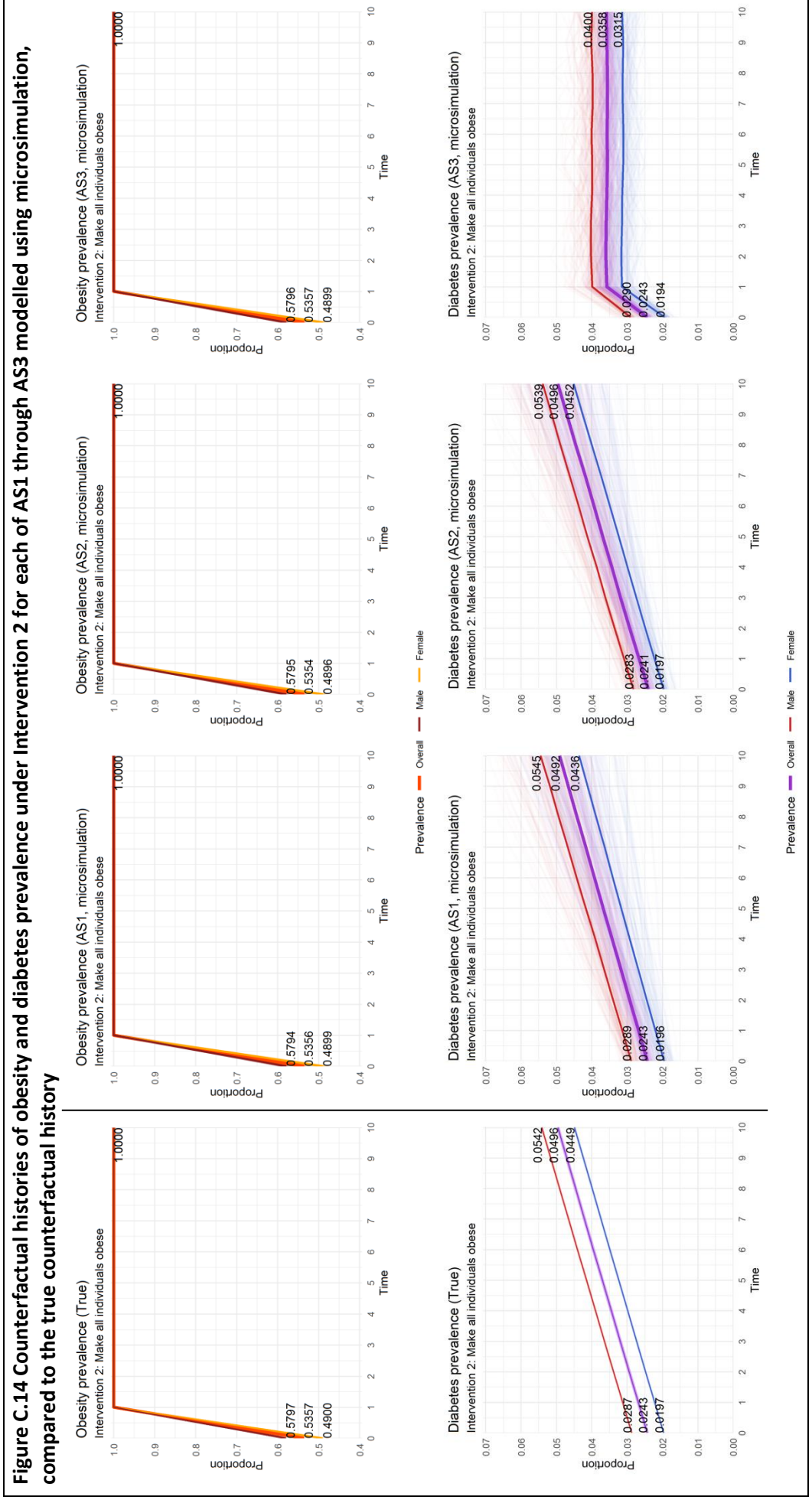


Figure C.15 Counterfactual histories of obesity and diabetes prevalence under Intervention 3 for each of AS1 through AS3 modelled using microsimulation, compared to the true counterfactual history

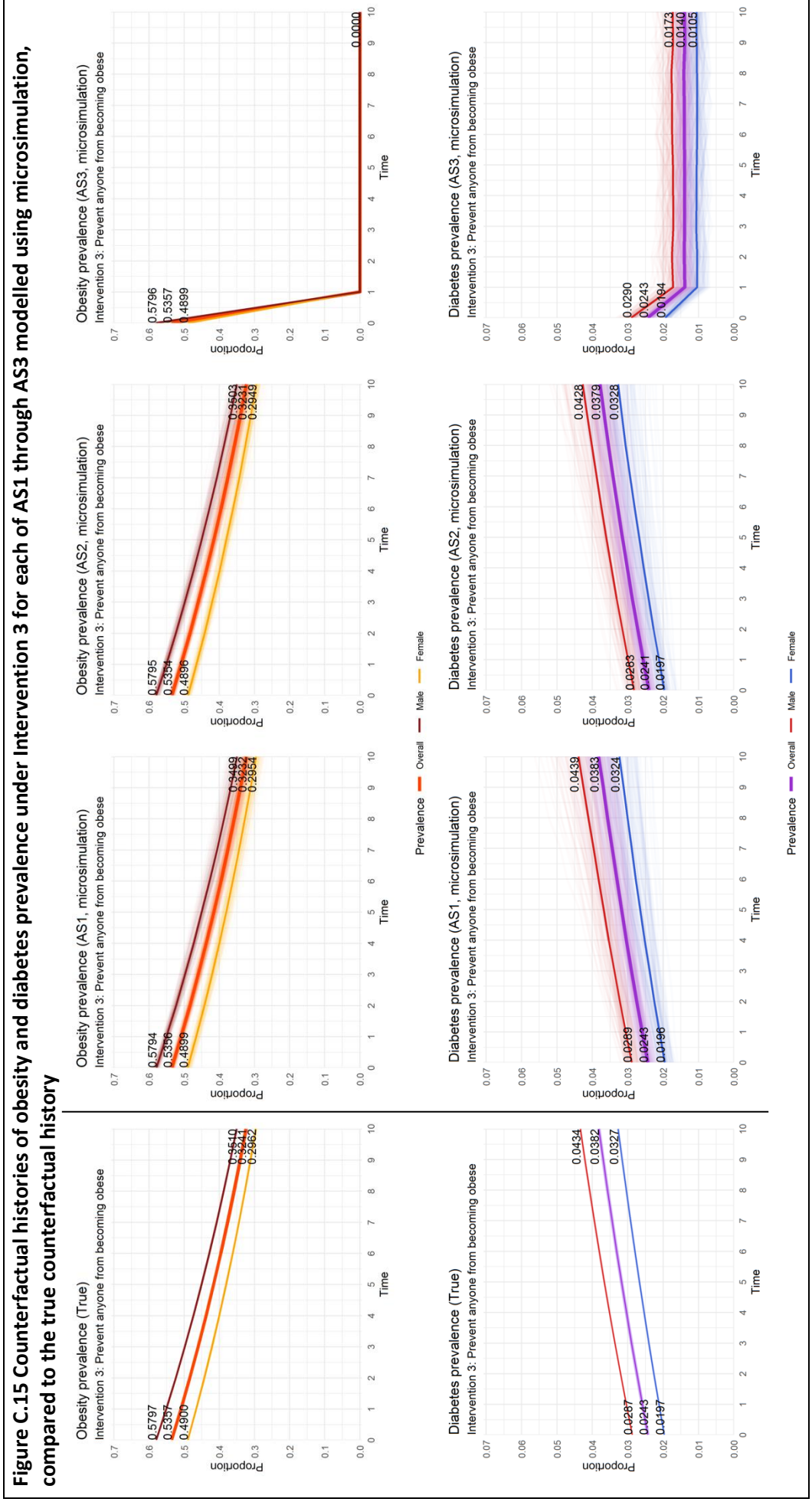


Figure C.16 Counterfactual histories of obesity and diabetes prevalence under Intervention 4 for each of AS1 through AS3 modelled using microsimulation, compared to the true counterfactual history

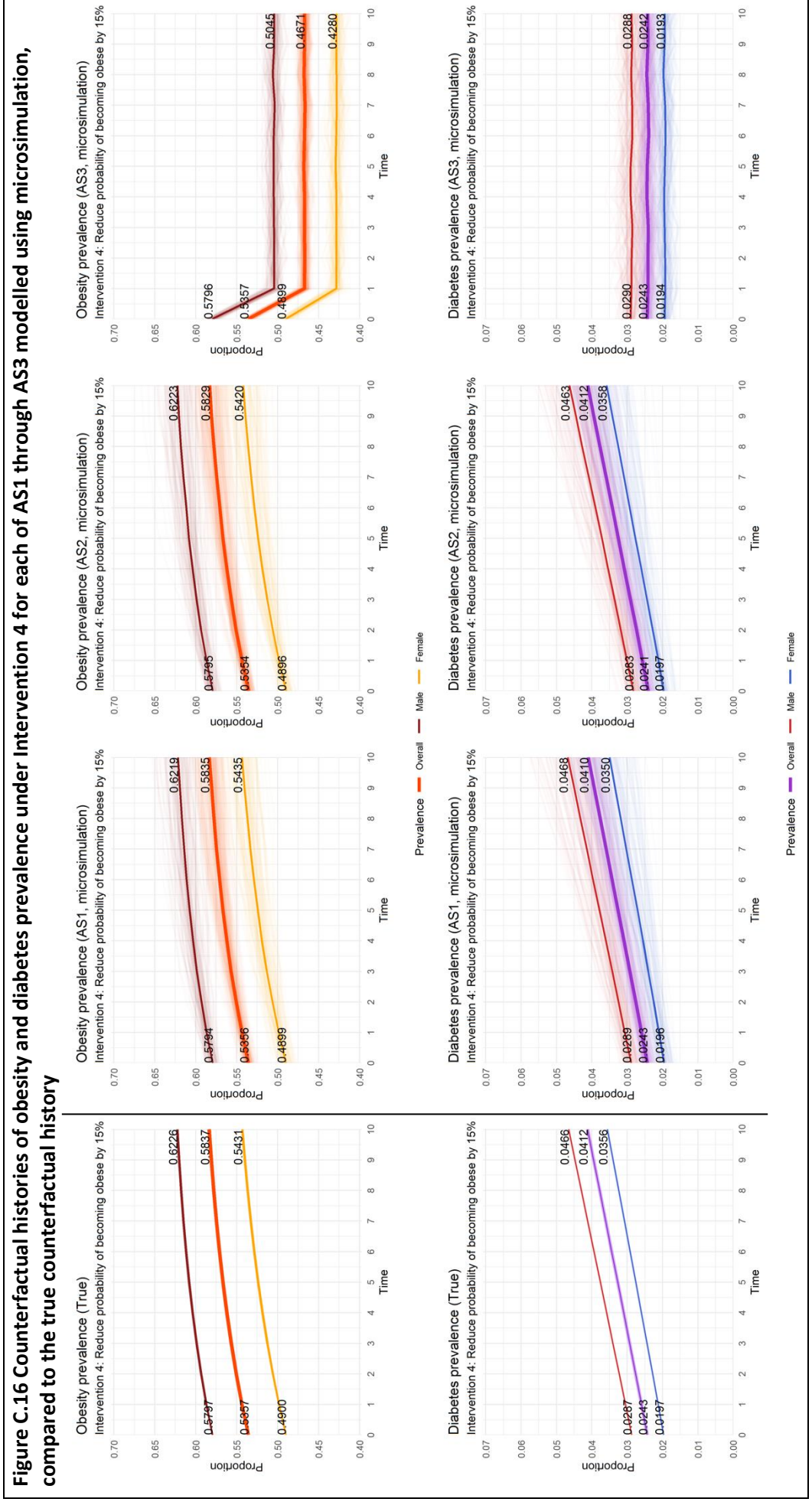


Figure C.17 Counterfactual histories of obesity and diabetes prevalence under Intervention 5 for each of AS1 through AS3 modelled using microsimulation, compared to the true counterfactual history

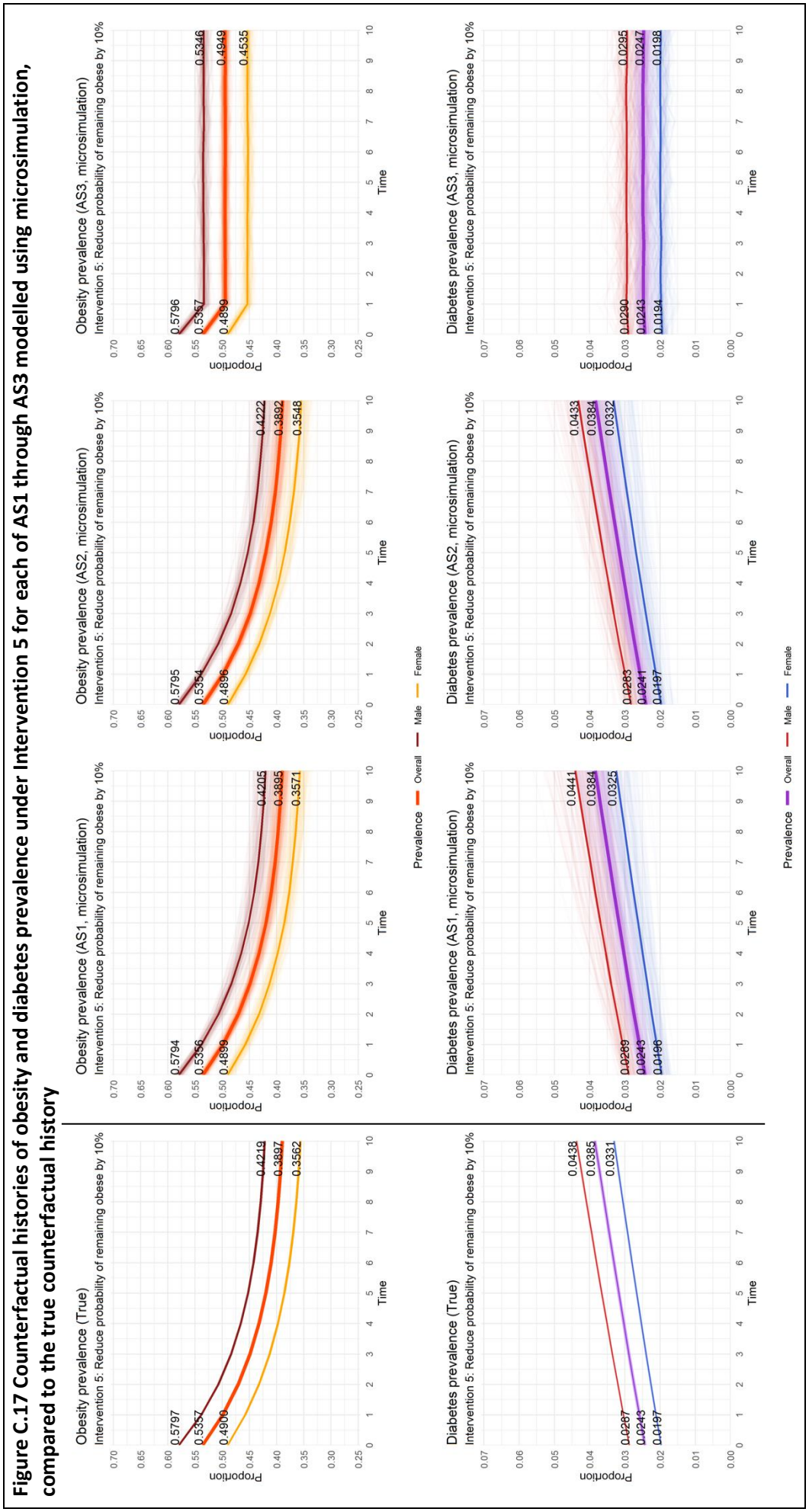
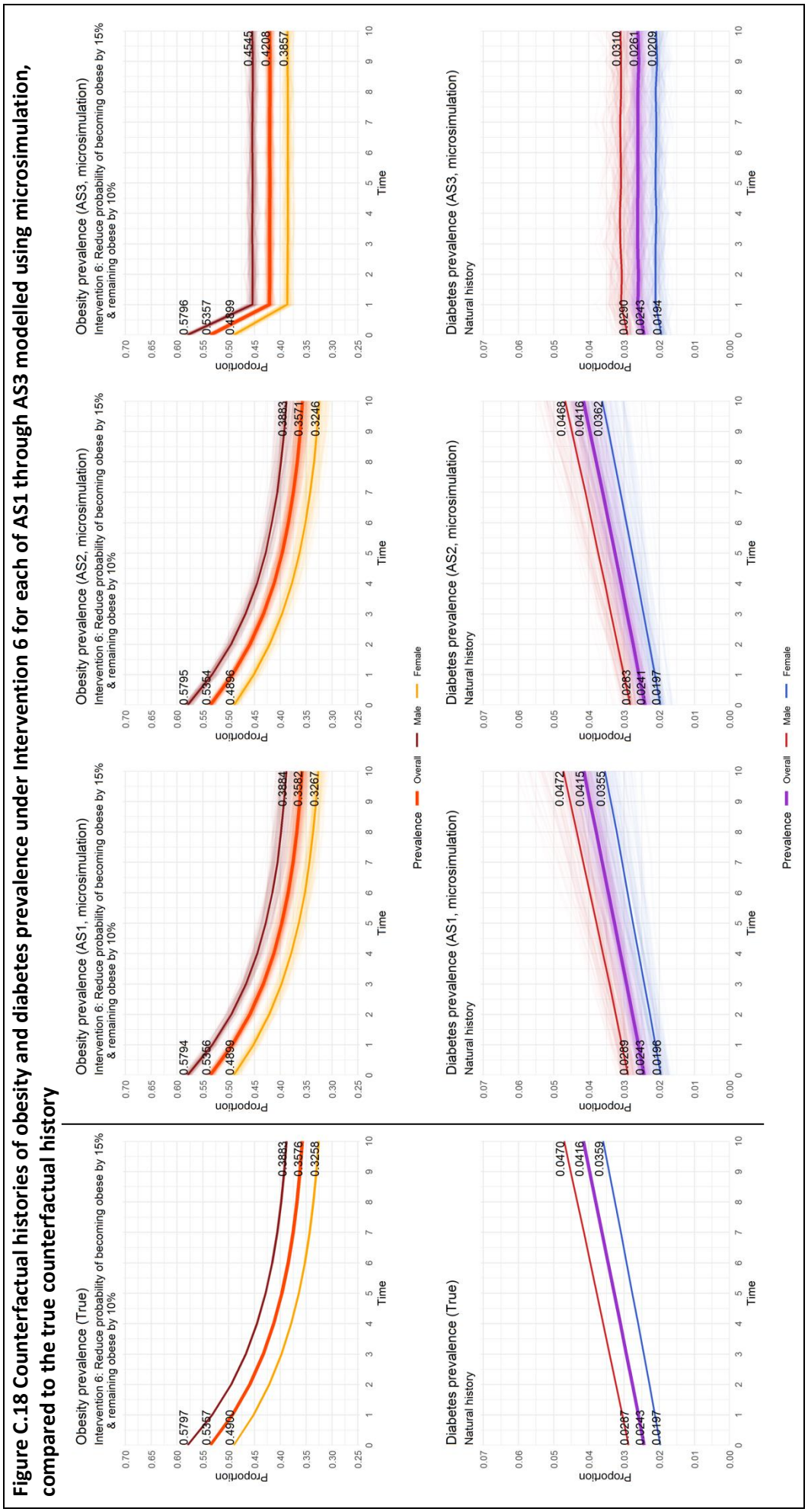


Figure C.18 Counterfactual histories of obesity and diabetes prevalence under Intervention 6 for each of AS1 through AS3 modelled using microsimulation, compared to the true counterfactual history



C.2.2.3 Annotated R code

```
################################################################################
## MSM: AUTOCORRELATION STRUCTURE 1 ############################################
################################################################################

# This code simulates the 'natural history' of an artificial longitudinal population
# sampled from a population of 5 million ('Population simulation - vectorised.R')
# using a TIME-based, discrete time microsimulation model

# The true autocorrelation structure of the source population is modelled

# Simulated individuals have the following 3 attributes:
# Sex (time-fixed): 0 = female, 1 = male
# Obesity (time-varying): 0 = nonobese, 1 = obese
# Diabetes (time-varying): 0 = nondiabetic, 1 = diabetic

# It then simulates the effects on diabetes prevalence at time 10
# of the following interventions:

# (1) Preventing anyone from being obese
# (2) Making everyone obese
# (3) Preventing any new obese individuals
# (4) Reducing the probability of becoming obese by 15%
# (5) Reducing the probability of remaining obese by 10%
# (6) Reducing the probability of becoming obese by 15% and remaining obese by 10%

################################################################################
## (1) SET UP ----------------------------------------------------------------

# Clear workspace
rm(list = ls())

# Load all required packages
library(readxl); library(stringr); library(Hmisc); library(plyr)
library(scales); library(ggplot2); library(gridExtra); library(HydeNet)
library(data.table)

### (a) Population/sample/simulation parameters & population data --------------

# Import individual-level population dataset
Population <- read.csv("./Population simulation - vectorised/PopData.csv",
                       header = TRUE, row.names = 1)

# Define population parameters
N.i.pop <- nrow(Population)  # number of individuals in population
N.t.pop <- 11  # number of time points (including baseline)

# Define sample parameters
N.i.sam <- 20000  # number of individuals to sample from population
N.t.sam <- 11  # number of time points (including baseline)
Time.sam <- as.vector(seq(from = 0, to = (N.t.sam - 1), by = 1),
                      mode = "integer")  # time vector

# Define simulation parameters
N.sim <- 100  # number of simulation runs (per intervention/natural history)
N.int <- 6  # number of interventions (not including natural history)

### (b) Tables to store sample simulation data ---------------------------------

# Create empty matrices to store individual-level sample data
# Each row represents 1 individual (N.i.sam rows)
# Each column represents 1 time point (N.t.sam columns)
Sex.sam1 <- matrix(nrow = N.i.sam, ncol = 1,
                   dimnames = list(paste0("ind", 1:N.i.sam), "Sex"))
Obes.sam1 <- matrix(nrow = N.i.sam, ncol = N.t.sam,
                    dimnames = list(paste0("ind", 1:N.i.sam),
                                    paste0("O.t", Time.sam)))
Diab.sam1 <- matrix(nrow = N.i.sam, ncol = N.t.sam,
                    dimnames = list(paste0("ind", 1:N.i.sam),
                                    paste0("D.t", Time.sam)))

### (c) Tables to store summary data -------------------------------------------

# Create empty cross-sectional frequency table
Frequency.cs.sam1 <- data.frame(Sim = numeric(), Time = numeric(),
                                Sex = numeric(), O.t = numeric(), D.t = numeric(),
                                freq = numeric())

# Create empty tables to record obesity & diabetes prevalence from sample
# (overall and disaggregated by sex)
Obes.prev.sam1 <- data.frame(Sim = numeric(), Time = numeric(),
                             Subgroup = factor(), prev = numeric())
Diab.prev.sam1 <- data.frame(Sim = numeric(), Time = numeric(),
                             Subgroup = factor(), prev = numeric())

# Create empty tables to record cross-sectional conditional probabilities of obesity & diabetes
CProbability.Obes.cs.sam1 <- data.frame(Sim = numeric(), Time = numeric(),
                                        Sex = factor(), O.t = factor(),
                                        prob = numeric())
CProbability.Diab.cs.sam1 <- data.frame(Sim = numeric(), Time = numeric(),
                                        Sex = factor(), O.t = factor(),
                                        D.t = factor(),
                                        prob = numeric())

# Create empty tables to record cross-time conditional probabilities of obesity & diabetes
CProbability.Obes.ct.sam1 <- data.frame(Sim = numeric(), Time = numeric(),
                                       Sex = factor(), O.tminus1 = factor(),
```

```
                                      D.tminus1 = factor(), O.t = factor(),
                                      prob = numeric())
CProbability.Diab.ct.sam1 <- data.frame(Sim = numeric(), Time = numeric(),
                                      Sex = factor(), D.tminus1 = factor(),
                                      O.t = factor(), D.t = factor(),
                                      prob = numeric())

# Create empty table to record estimated parameters governing transition probabilities
TransitionParameters.sam1 <- data.frame(Sim = numeric(), Parameter = factor(),
                                       value = numeric())

# Create empty tables to store mean frequency, prevalence, & conditional probability trends
## Cross-sectional frequencies
Mean.frequency.sam1 <- expand.grid(D.t = c(0, 1), O.t = c(0, 1), Sex = c(0, 1),
                                   Time = Time.sam)
Mean.frequency.sam1 <- cbind(Sim = "mean",
                             Mean.frequency.sam1[, c("Time", "Sex", "O.t", "D.t")],
                             freq = 0)  # initialise frequencies with 0
## Obesity prevalence
Mean.obes.prev.sam1 <- expand.grid(Subgroup = c("Obes.prev", "Obes.prev.f", "Obes.prev.m"),
                                   Time = Time.sam)
Mean.obes.prev.sam1 <- cbind(Sim = "mean",
                             Mean.obes.prev.sam1[, c("Time", "Subgroup")],
                             prev = 0)  # initialise prevalence with 0
## Diabetes prevalence
Mean.diab.prev.sam1 <- expand.grid(Subgroup = c("Diab.prev", "Diab.prev.f", "Diab.prev.m"),
                                   Time = Time.sam)
Mean.diab.prev.sam1 <- cbind(Sim = "mean",
                             Mean.diab.prev.sam1[, c("Time", "Subgroup")],
                             prev = 0)  # initialise prevalence with 0
## CP obesity - cross-sectional
Mean.CP.Obes.cs.sam1 <- expand.grid(O.t = c(0, 1), Sex = c(0, 1), Time = Time.sam)
Mean.CP.Obes.cs.sam1 <- cbind(Sim = "mean",
                              Mean.CP.Obes.cs.sam1[, c("Time", "Sex", "O.t")],
                              prob = 0)  # initialise probs with 0
## CP obesity - cross-time
Mean.CP.Obes.ct.sam1 <- expand.grid(O.t = c(0, 1), D.tminus1 = c(0, 1),
                                    O.tminus1 = c(0, 1), Sex = c(0, 1),
                                    Time = Time.sam[-1])
Mean.CP.Obes.ct.sam1 <- cbind(Sim = "mean",
                              Mean.CP.Obes.ct.sam1[, c("Time", "Sex", "O.tminus1", "D.tminus1",
"O.t")],
                              prob = 0)  # initialise probs with 0
## CP diabetes - cross-sectional
Mean.CP.Diab.cs.sam1 <- expand.grid(D.t = c(0, 1), O.t = c(0, 1),
                                    Sex = c(0, 1), Time = Time.sam)
Mean.CP.Diab.cs.sam1 <- cbind(Sim = "mean",
                              Mean.CP.Diab.cs.sam1[, c("Time", "Sex", "O.t", "D.t")],
                              prob = 0)  # initialise probs with 0
## CP diabetes - cross-time
Mean.CP.Diab.ct.sam1 <- expand.grid(D.t = c(0, 1), O.t = c(0, 1),
                                    D.tminus1 = c(0, 1), Sex = c(0, 1),
                                    Time = Time.sam[-1])
Mean.CP.Diab.ct.sam1 <- cbind(Sim = "mean",
                              Mean.CP.Diab.ct.sam1[, c("Time", "Sex", "D.tminus1", "O.t", "D.t")],
                              prob = 0)  # initialise probs with 0
## Transition parameters
Mean.TP.sam1 <- cbind.data.frame(Sim = "mean",
                                 Parameter = c(paste0("a", 0:7), paste0("b", 0:7)),
                                 value = 0)  # initialise values with 0

### (d) Functions ------------------------------------------------------------

#### (i) SampleV function ----------------------------------------------------

# samplev() function
# efficient implementation of the rMultinom() function of the Hmisc package
# from Krijkamp et al (2018)
samplev <- function(probs, m) {
  d <- dim(probs)  # (dimensions of probability matrix)
  n <- d[1]  # (number of rows, i.e. individuals)
  k <- d[2]  # (number of columns, i.e. states)
  lev <- dimnames(probs)[[2]]  # (names of columns, i.e. state values)
  if (!length(lev))
    lev <- 1:k
  ran <- matrix(lev[1], ncol = m, nrow = n)
  U <- t(probs)
  for(i in 2:k) {
    U[i, ] <- U[i, ] + U[i - 1, ]
  }
  if (any((U[k, ] - 1) > 1e-05))
    stop("error in multinom: probabilities do not sum to 1")

  for (j in 1:m) {
    un <- rep(runif(n), rep(k, n))
    ran[, j] <- lev[1 + colSums(un > U)]
  }
  ran
}

#### (ii) Calculate prevalence proportions -----------------------------------

# Calculate Obesity prevalence
# args: group = subgroup, freqtable = frequency table (numeric)
# returns single number (prevalence)
CalculatePrevObesityT <- function (group, freqtable) {

  if (group == "overall") {

    prevalence <- sum(subset(freqtable, Sim == s & Time == (t-1) & O.t == 1)$freq) /
      N.i.sam
    return(prevalence)

```

```
  } else if (group == "female") {

    prevalence <- sum(subset(freqtable, Sim == s & Time == (t-1) & Sex == 0 & O.t == 1)$freq) /
      sum(subset(freqtable, Sim == s & Time == (t-1) & Sex == 0)$freq)
    return(prevalence)

  } else if (group == "male") {

    prevalence <- sum(subset(freqtable, Sim == s & Time == (t-1) & Sex == 1 & O.t == 1)$freq) /
      sum(subset(freqtable, Sim == s & Time == (t-1) & Sex == 1)$freq)
    return(prevalence)

  }

}  # (close function loop)


# Calculate Diabetes prevalence
# args: group = subgroup, freqtable = frequency table (numeric)
# returns single number (prevalence)
CalculatePrevDiabetesT <- function (group, freqtable) {

  if (group == "overall") {

    prevalence <- sum(subset(freqtable, Sim == s & Time == (t-1) & D.t == 1)$freq) /
      N.i.sam
    return(prevalence)

  } else if (group == "female") {

    prevalence <- sum(subset(freqtable, Sim == s & Time == (t-1) & Sex == 0 & D.t == 1)$freq) /
      sum(subset(freqtable, Sim == s & Time == (t-1) & Sex == 0)$freq)
    return(prevalence)

  } else if (group == "male") {

    prevalence <- sum(subset(freqtable, Sim == s & Time == (t-1) & Sex == 1 & D.t == 1)$freq) /
      sum(subset(freqtable, Sim == s & Time == (t-1) & Sex == 1)$freq)
    return(prevalence)

  }

}  # (close function loop)

#### (iii) Calculate conditional probabilities ---------------------------------

# Calculate conditional probability table at time t
# args: dv = dependent variable, iv = independent variable(s), dataset = data frame (factorised)
# returns conditional probability table (cprob.t)
CalculateCPT <- function(dv, iv, dataset) {

  # Define formula for use in cpt function (from HydeNet package)
  formula <- as.formula(paste(dv, paste(iv, collapse = " + "), sep = " ~ "))

  # Create conditional probability table
  cprob.t <- cbind(Time = (t-1), am_adt(cpt(formula, data = dataset)))

  return(cprob.t)

}

# Function for converting multidimensional arrays to tables
# (from https://github.com/Rdatatable/data.table/issues/1418)
am_adt <- function(inarray) {
  if (!is.array(inarray)) stop("input must be an array")
  dims <- dim(inarray)
  if (is.null(dimnames(inarray))) {
    inarray <- provideDimnames(inarray, base = list(as.character(seq_len(max(dims)))))
  }
  FT <- if (any(class(inarray) %in% "ftable")) inarray else ftable(inarray)
  out <- data.table(as.table(ftable(FT)))
  nam <- names(out)[seq_along(dims)]
  setorderv(out[, (nam) := lapply(.SD, type.convert), .SDcols = nam], nam)[]
}



################################################################################
## (2) MSM: AUTOCORRELATION STRUCTURE 1 ----------------------------------------

# Set seed
set.seed(101)

### (a) Define conditional probabilities at time t -----------------------------

#### (i) Obesity ---------------------------------------------------------------

# Function to calculate P(Obesity = 1 | Sex, Prev obesity, Prev diabetes) at time t

CalculateProbObesityT <- function(Sex, PrevObes, PrevDiab) {

  # Incident probability (PrevObes = 0): a0 + (a2-a0)*Sex + (a1-a0)*PrevDiab + (a3-a2-
a1+a0)*Sex*PrevDiab
  # Prevalent probability (PrevObes = 1): a4 + (a6-a4)*Sex + (a5-a4)*PrevDiab + (a7-a6-
a5+a4)*Sex*PrevDiab

  p.obes.t <- a0 + (a2-a0)*Sex + (a1-a0)*PrevDiab + (a3-a2-a1+a0)*Sex*PrevDiab +
    PrevObes*((a4-a0) + (a6-a2-(a4-a0))*Sex + (a5-a1-(a4-a0))*PrevDiab + (a7-a6-a5-a3+a2+a1+(a4-
a0))*Sex*PrevDiab)
  return(p.obes.t)

}
```

```
# (a0,...,a7 will be estimated from sample of population for each simulation run...
# ... using EstimateTransitionProbs function)

#### (ii) Diabetes ----------------------------------------------------------

# Function to calculate P(Diabetes = 1 | Sex, Obesity, Prev diabetes) at time t

CalculateProbDiabetesT <- function(Sex, PrevDiab, Obes) {

  # Incident probability: b0 + (b2-b0)*Sex + (b1-b0)*Obes + (b3-b2-b1+b0)*Sex*Obes
  # Prevalent probability: b4 + (b6-b4)*Sex + (b5-b4)*Obes + (b7-b6-b5+b4)*Sex*Obes

  p.diab.t <- b0 + (b2-b0)*Sex + (b1-b0)*Obes + (b3-b2-b1+b0)*Sex*Obes +
    PrevDiab*((b4-b0) + (b6-b2-(b4-b0))*Sex + (b5-b1-(b4-b0))*Obes + (b7-b6-b5-b3+b2+b1+(b4-
b0))*Sex*Obes)
  return(p.diab.t)

}

# (b0,...,b7 will be estimated from sample of population for each simulation run...
# ... using EstimateTransitionProbs function)

#### (iii) Calculate parameters governing transition probabilities -------------

# Function to estimate transition probabilities from a sample of individuals for time 1 - will
change based on autocorrelation structure
# args: sampledata = individual-level sample dataset (numeric)

EstimateTransitionProbs <- function(sampledata) {

  # Create dataframe for sample data (baseline & time 1)
  vars <- c("Sex", "O.t0", "D.t0", "O.t1", "D.t1")  # define variables
  sampledata.t1 <- data.frame(sampledata[, vars])
  names(sampledata.t1) <- c("Sex", "O.tminus1", "D.tminus1", "O.t", "D.t")  # rename variables
  sampledata.t1 <- data.frame(apply(sampledata.t1, 2, factor))  # convert vars to factors

  # Calculate cross-time conditional probabilities & define transition parameters
  # (1) Obesity
  var.d <- "O.t"  # define dependent variable
  var.i <- c("Sex", "O.tminus1", "D.tminus1")  # define independent variables
  formula <- as.formula(paste(var.d, paste(var.i, collapse = " + "), sep = " ~ "))
  CP.Obes <- data.frame(am_adt(cpt(formula, data = sampledata.t1)))
  CP.Obes <- rename(CP.Obes, replace = c("N" = "prob"))  # rename prob column
  CP.Obes <- subset(CP.Obes, O.t == "1")  # remove 'complement' rows
  CP.Obes <- subset(CP.Obes, select = -O.t)  # remove O.t column
  a0 <<- CP.Obes[CP.Obes[, "O.tminus1"] == 0 &
                   CP.Obes[, "Sex"] == 0 &
                   CP.Obes[, "D.tminus1"] == 0, "prob"]
  a1 <<- CP.Obes[CP.Obes[, "O.tminus1"] == 0 &
                   CP.Obes[, "Sex"] == 0 &
                   CP.Obes[, "D.tminus1"] == 1, "prob"]
  a2 <<- CP.Obes[CP.Obes[, "O.tminus1"] == 0 &
                   CP.Obes[, "Sex"] == 1 &
                   CP.Obes[, "D.tminus1"] == 0, "prob"]
  a3 <<- CP.Obes[CP.Obes[, "O.tminus1"] == 0 &
                   CP.Obes[, "Sex"] == 1 &
                   CP.Obes[, "D.tminus1"] == 1, "prob"]
  a4 <<- CP.Obes[CP.Obes[, "O.tminus1"] == 1 &
                   CP.Obes[, "Sex"] == 0 &
                   CP.Obes[, "D.tminus1"] == 0, "prob"]
  a5 <<- CP.Obes[CP.Obes[, "O.tminus1"] == 1 &
                   CP.Obes[, "Sex"] == 0 &
                   CP.Obes[, "D.tminus1"] == 1, "prob"]
  a6 <<- CP.Obes[CP.Obes[, "O.tminus1"] == 1 &
                   CP.Obes[, "Sex"] == 1 &
                   CP.Obes[, "D.tminus1"] == 0, "prob"]
  a7 <<- CP.Obes[CP.Obes[, "O.tminus1"] == 1 &
                   CP.Obes[, "Sex"] == 1 &
                   CP.Obes[, "D.tminus1"] == 1, "prob"]
  # (2) Diabetes
  var.d <- "D.t"  # define dependent variable
  var.i <- c("Sex", "D.tminus1", "O.t")  # define independent variables
  formula <- as.formula(paste(var.d, paste(var.i, collapse = " + "), sep = " ~ "))
  CP.Diab <- data.frame(am_adt(cpt(formula, data = sampledata.t1)))
  CP.Diab <- rename(CP.Diab, replace = c("N" = "prob"))  # rename prob column
  CP.Diab <- subset(CP.Diab, D.t == "1")  # remove 'complement' rows
  CP.Diab <- subset(CP.Diab, select = -D.t)  # remove D.t column
  b0 <<- CP.Diab[CP.Diab[, "D.tminus1"] == 0 &
                   CP.Diab[, "Sex"] == 0 &
                   CP.Diab[, "O.t"] == 0, "prob"]
  b1 <<- CP.Diab[CP.Diab[, "D.tminus1"] == 0 &
                   CP.Diab[, "Sex"] == 0 &
                   CP.Diab[, "O.t"] == 1, "prob"]
  b2 <<- CP.Diab[CP.Diab[, "D.tminus1"] == 0 &
                   CP.Diab[, "Sex"] == 1 &
                   CP.Diab[, "O.t"] == 0, "prob"]
  b3 <<- CP.Diab[CP.Diab[, "D.tminus1"] == 0 &
                   CP.Diab[, "Sex"] == 1 &
                   CP.Diab[, "O.t"] == 1, "prob"]
  b4 <<- CP.Diab[CP.Diab[, "D.tminus1"] == 1 &
                   CP.Diab[, "Sex"] == 0 &
                   CP.Diab[, "O.t"] == 0, "prob"]
  b5 <<- CP.Diab[CP.Diab[, "D.tminus1"] == 1 &
                   CP.Diab[, "Sex"] == 0 &
                   CP.Diab[, "O.t"] == 1, "prob"]
  b6 <<- CP.Diab[CP.Diab[, "D.tminus1"] == 1 &
                   CP.Diab[, "Sex"] == 1 &
                   CP.Diab[, "O.t"] == 0, "prob"]
  b7 <<- CP.Diab[CP.Diab[, "D.tminus1"] == 1 &
                   CP.Diab[, "Sex"] == 1 &
                   CP.Diab[, "O.t"] == 1, "prob"]

}
```

```
### (b) Simulation -----------------------------------------------------------

# Initialise obesity (a) and diabetes (b) parameters with 0
a0 <- a1 <- a2 <- a3 <- a4 <- a5 <- a6 <- a7 <- 0
b0 <- b1 <- b2 <- b3 <- b4 <- b5 <- b6 <- b7 <- 0

# Draw N.sim random numbers from N.i.sam
# These represent the random samples that will be drawn from the population
select <- matrix(nrow = N.i.sam, ncol = N.sim,
                 dimnames = list(paste0("ind", 1:N.i.sam),
                                 paste0("sample", 1:N.sim)))
for (s in 1:N.sim) {
  select[, s] <- sample(x = c(1:N.i.pop), size = N.i.sam, replace = FALSE)
}

# Record start time of simulation
v <- Sys.time()

# (1) Loop through natural history & interventions ---------------------------
# (v = 0 represents natural history, v = 1-6 represent interventions 1-6)
for (z in 0:N.int) {

  # Reset summary tables
  # (each intervention (or natural history) has a separate summary table)
  Frequency.cs.sam1 <- Frequency.cs.sam1[0, ]
  Obes.prev.sam1 <- Obes.prev.sam1[0, ]
  Diab.prev.sam1 <- Diab.prev.sam1[0, ]
  CProbability.Obes.cs.sam1 <- CProbability.Obes.cs.sam1[0, ]
  CProbability.Diab.cs.sam1 <- CProbability.Diab.cs.sam1[0, ]
  CProbability.Obes.ct.sam1 <- CProbability.Obes.ct.sam1[0, ]
  CProbability.Diab.ct.sam1 <- CProbability.Diab.ct.sam1[0, ]
  TransitionParameters.sam1 <- TransitionParameters.sam1[0, ]

  # (2) Loop through simulation runs
  for (s in 1:N.sim) {

    # Sample N.i.sam individuals from population
    # Store data in Sample dataframe (all longitudinal vars in order)
    Sample <- Population[select[, s], ]

    # Estimate transition probabilities (for natural history) using Sample dataframe
    EstimateTransitionProbs(sampledata = Sample)

    # Define transition probabilities for obesity under Interventions 1-6
    if (z != 0) {
      if (z == 1) {  # under Intervention 1, no individuals may be obese
        a0 <- a1 <- a2 <- a3 <- a4 <- a5 <- a6 <- a7 <- 0
      } else if (z == 2) {  # under Intervention 2, all individuals are obese
        a0 <- a1 <- a2 <- a3 <- a4 <- a5 <- a6 <- a7 <- 1
      } else if (z == 3) {  # under Intervention 3, no individuals may become obese
        a0 <- a1 <- a2 <- a3 <- 0
      } else if (z == 4) {  # under Intervention 4, reduce incident prob of obesity by 15%
        a0 <- 0.85*a0; a1 <- 0.85*a1; a2 <- 0.85*a2; a3 <- 0.85*a3
      } else if (z == 5) {  # under Intervention 5, reduce prevalent prob of obesity by 10%
        a4 <- 0.9*a4; a5 <- 0.9*a5; a6 <- 0.9*a6; a7 <- 0.9*a7
      } else if (z == 6) {  # under Intervention 6, reduce incident prob of obesity by 15% &
prevalent prob of obesity by 10%
        a0 <- 0.85*a0; a1 <- 0.85*a1; a2 <- 0.85*a2; a3 <- 0.85*a3
        a4 <- 0.9*a4; a5 <- 0.9*a5; a6 <- 0.9*a6; a7 <- 0.9*a7
      }
    }

    # Record parameters governing transition probabilities
    par.t <- cbind.data.frame(Sim = s,
                              Parameter = c(paste0("a", 0:7), paste0("b", 0:7)),
                              value = c(a0, a1, a2, a3, a4, a5, a6, a7, b0, b1, b2, b3, b4, b5, b6,
b7)) # create parameter table for time t
    TransitionParameters.sam1 <- rbind.data.frame(TransitionParameters.sam1, par.t)

    # Fill Sex.sim, Obes.sim, & Diab.sim matrices with baseline data
    Sex.sam1[, "Sex"] <- Sample[, "Sex"]
    Obes.sam1[, "O.t0"] <- Sample[, "O.t0"]
    Diab.sam1[, "D.t0"] <- Sample[, "D.t0"]

    # (3) Loop through time points
    for (t in 1:N.t.sam) {

      ## Record summary data at baseline
      if (t == 1) {

        # Record summary data ------------------------------------------------

        # Bind variables from time t and baseline together
        Sample.t <- data.frame(cbind(Sex.sam1[, 1], Obes.sam1[, t], Diab.sam1[, t]))
        vars.cs <- c("Sex", paste0(c("O.t", "D.t"), (t-1)))  # define variables
        names(Sample.t) <- vars.cs

        # (a) Cross-sectional frequency table --------------------------------

        freq.t <- cbind(Sim = s, Time = (t-1), count(Sample.t[, vars.cs]))
        names(freq.t) <- names(Frequency.cs.sam1)
        Frequency.cs.sam1 <- rbind(Frequency.cs.sam1, freq.t)

        # (b) Prevalence -----------------------------------------------------

        ## Obesity
        prev.O <- cbind.data.frame(Sim = s, Time = (t-1), Subgroup = "Obes.prev",
                                   prev = CalculatePrevObesityT("overall", Frequency.cs.sam1))
        prev.O.f <- cbind.data.frame(Sim = s, Time = (t-1), Subgroup = "Obes.prev.f",
                                     prev = CalculatePrevObesityT("female", Frequency.cs.sam1))
        prev.O.m <- cbind.data.frame(Sim = s, Time = (t-1), Subgroup = "Obes.prev.m",
                                     prev = CalculatePrevObesityT("male", Frequency.cs.sam1))
```

```
        Obes.prev.sam1 <- rbind.data.frame(Obes.prev.sam1, prev.O, prev.O.f, prev.O.m)

        ## Diabetes
        prev.D <- cbind.data.frame(Sim = s, Time = (t-1), Subgroup = "Diab.prev",
                                   prev = CalculatePrevDiabetesT("overall", Frequency.cs.sam1))
        prev.D.f <- cbind.data.frame(Sim = s, Time = (t-1), Subgroup = "Diab.prev.f",
                                     prev = CalculatePrevDiabetesT("female", Frequency.cs.sam1))
        prev.D.m <- cbind.data.frame(Sim = s, Time = (t-1), Subgroup = "Diab.prev.m",
                                     prev = CalculatePrevDiabetesT("male", Frequency.cs.sam1))
        Diab.prev.sam1 <- rbind.data.frame(Diab.prev.sam1, prev.D, prev.D.f, prev.D.m)

        # (c) Conditional probabilities ----------------------------------------

        # Convert variables in Sample.t dataset to factors
        # (requred for calculating conditonal probabilties)
        Sample.t <- data.frame(lapply(Sample.t, factor, levels = c("0", "1")))

        ## (i) Cross-sectional ------------------------------------------------

        ## Obesity
        var.d <- paste0("O.t", (t-1))  # (define dependent variable)
        var.i <- "Sex"  # (define independent variable)
        cprob.t <- cbind(Sim = s, CalculateCPT(dv = var.d, iv = var.i, dataset = Sample.t))
        names(cprob.t) <- names(CProbability.Obes.cs.sam1)  # rename columns to match CP table
        CProbability.Obes.cs.sam1 <- rbind.data.frame(CProbability.Obes.cs.sam1, cprob.t)

        ## Diabetes
        var.d <- paste0("D.t", (t-1))
        var.i <- c("Sex", paste0("O.t", (t-1)))
        cprob.t <- cbind(Sim = s, CalculateCPT(dv = var.d, iv = var.i, dataset = Sample.t))
        names(cprob.t) <- names(CProbability.Diab.cs.sam1)
        CProbability.Diab.cs.sam1 <- rbind(CProbability.Diab.cs.sam1, cprob.t)


      } else {  ## Update time-varying characteristics & record summary data at time t

        # Update time-varying characteristics ----------------------------------

        # (a) Obesity ----------------------------------------------------------
        p.obes.t <- cbind(1 - CalculateProbObesityT(Sex = Sex.sam1[, 1],
                                                    PrevObes = Obes.sam1[, (t-1)],
                                                    PrevDiab = Diab.sam1[, (t-1)]),
                          CalculateProbObesityT(Sex = Sex.sam1[, 1],
                                                PrevObes = Obes.sam1[, (t-1)],
                                                PrevDiab = Diab.sam1[, (t-1)]))
        Obes.sam1[, t] <- samplev(probs = p.obes.t, m = 1)
        Obes.sam1[, t] <- Obes.sam1[, t] - 1  # (factor levels should be 0 and 1)

        # (b) Diabetes ---------------------------------------------------------
        p.diab.t <- cbind(1 - CalculateProbDiabetesT(Sex = Sex.sam1[, 1],
                                                     PrevDiab = Diab.sam1[, (t-1)],
                                                     Obes = Obes.sam1[, t]),
                          CalculateProbDiabetesT(Sex = Sex.sam1[, 1],
                                                 PrevDiab = Diab.sam1[, (t-1)],
                                                 Obes = Obes.sam1[, t]))
        Diab.sam1[, t] <- samplev(probs = p.diab.t, m = 1)
        Diab.sam1[, t] <- Diab.sam1[, t] - 1  # (factor levels should be 0 and 1)

        # Record summary data --------------------------------------------------

        # Bind variables from time t, time t-1, and baseline together
        Sample.t <- data.frame(cbind(Sex.sam1[, 1], Obes.sam1[, (t-1)], Diab.sam1[, (t-1)]),
                               Obes.sam1[, t], Diab.sam1[, t])
        vars.cs <- c("Sex", paste0(c("O.t", "D.t"), (t-1)))
        vars.ct <- c("Sex", paste0(c("O.t", "D.t"), (t-2)), paste0(c("O.t", "D.t"), (t-1)))
        names(Sample.t) <- vars.ct

        # (a) Cross-sectional frequency table ----------------------------------
        freq.t <- cbind(Sim = s, Time = (t-1), count(Sample.t[, vars.cs]))
        names(freq.t) <- names(Frequency.cs.sam1)  # rename columns to match Frequency table
        Frequency.cs.sam1 <- rbind(Frequency.cs.sam1, freq.t)

        # (b) Prevalence -------------------------------------------------------

        ## Obesity
        prev.O <- cbind.data.frame(Sim = s, Time = (t-1), Subgroup = "Obes.prev",
                                   prev = CalculatePrevObesityT("overall", Frequency.cs.sam1))
        prev.O.f <- cbind.data.frame(Sim = s, Time = (t-1), Subgroup = "Obes.prev.f",
                                     prev = CalculatePrevObesityT("female", Frequency.cs.sam1))
        prev.O.m <- cbind.data.frame(Sim = s, Time = (t-1), Subgroup = "Obes.prev.m",
                                     prev = CalculatePrevObesityT("male", Frequency.cs.sam1))
        Obes.prev.sam1 <- rbind.data.frame(Obes.prev.sam1, prev.O, prev.O.f, prev.O.m)

        ## Diabetes
        prev.D <- cbind.data.frame(Sim = s, Time = (t-1), Subgroup = "Diab.prev",
                                   prev = CalculatePrevDiabetesT("overall", Frequency.cs.sam1))
        prev.D.f <- cbind.data.frame(Sim = s, Time = (t-1), Subgroup = "Diab.prev.f",
                                     prev = CalculatePrevDiabetesT("female", Frequency.cs.sam1))
        prev.D.m <- cbind.data.frame(Sim = s, Time = (t-1), Subgroup = "Diab.prev.m",
                                     prev = CalculatePrevDiabetesT("male", Frequency.cs.sam1))
        Diab.prev.sam1 <- rbind.data.frame(Diab.prev.sam1, prev.D, prev.D.f, prev.D.m)

        # (c) Conditional probabilities ----------------------------------------

        # Convert variables in Sample.t dataset to factors
        # (requred for calculating conditonal probabilties)
        Sample.t <- data.frame(lapply(Sample.t, factor, levels = c("0", "1")))

        ## (i) Cross-sectional ------------------------------------------------

        ## Obesity
        var.d <- paste0("O.t", (t-1))  # (define dependent variable)
        var.i <- "Sex"  # (define independent variable)
```

```
        cprob.t <- cbind(Sim = s, CalculateCPT(dv = var.d, iv = var.i, dataset = Sample.t))
        names(cprob.t) <- names(CProbability.Obes.cs.sam1)  # rename columns to match CP table
        CProbability.Obes.cs.sam1 <- rbind.data.frame(CProbability.Obes.cs.sam1, cprob.t)

        ## Diabetes
        var.d <- paste0("D.t", (t-1))
        var.i <- c("Sex", paste0("O.t", (t-1)))
        cprob.t <- cbind(Sim = s, CalculateCPT(dv = var.d, iv = var.i, dataset = Sample.t))
        names(cprob.t) <- names(CProbability.Diab.cs.sam1)
        CProbability.Diab.cs.sam1 <- rbind(CProbability.Diab.cs.sam1, cprob.t)

        ## (ii) Cross-time ------------------------------------------------------

        ## Obesity
        var.d <- paste0("O.t", (t-1))
        var.i <- c("Sex", paste0(c("O.t", "D.t"), (t-2)))
        cprob.t <- cbind(Sim = s, CalculateCPT(dv = var.d, iv = var.i, dataset = Sample.t))
        names(cprob.t) <- names(CProbability.Obes.ct.sam1)
        CProbability.Obes.ct.sam1 <- rbind.data.frame(CProbability.Obes.ct.sam1, cprob.t)

        ## Diabetes
        var.d <- paste0("D.t", (t-1))
        var.i <- c("Sex", paste0("D.t", (t-2)), paste0("O.t", (t-1)))
        cprob.t <- cbind(Sim = s, CalculateCPT(dv = var.d, iv = var.i, dataset = Sample.t))
        names(cprob.t) <- names(CProbability.Diab.ct.sam1)
        CProbability.Diab.ct.sam1 <- rbind.data.frame(CProbability.Diab.ct.sam1, cprob.t)

      }

      # Display progress of simulation
      cat('\r', paste(round((t / N.t.sam * 100), 0),
                      "% done of simulation", s, "of", N.sim, "of Intervention", z, "of", N.int, "
", sep = " "))

    }  # (close time loop - 3)

  }  # (close simulation run loop - 2)

  # Calculate mean trends for Intervention z -----------------------------------

  # (a) Cross-sectional frequencies --------------------------------------------

  # Use expand.grid function to create full frequency table
  # (Frequency table doesn't show combinations with empty cells)
  f <- expand.grid(D.t = c(0, 1), O.t = c(0, 1), Sex = c(0, 1), Time = Time.sam,
                   Sim = seq(from = 1, to = N.sim, by = 1))
  f <- cbind(f[, c("Sim", "Time", "Sex", "O.t", "D.t")], freq = 0)  # initialise frequencies with 0
  # Fill f with data from (incomplete) Frequency table
  for (i in 1:nrow(Frequency.cs.sam1)) {

    sim <- Frequency.cs.sam1[i, "Sim"]
    time <- Frequency.cs.sam1[i, "Time"]
    sex <- Frequency.cs.sam1[i, "Sex"]
    o.t <- Frequency.cs.sam1[i, "O.t"]
    d.t <- Frequency.cs.sam1[i, "D.t"]
    freq <- Frequency.cs.sam1[i, "freq"]

    f[f[, "Sim"] == sim &
        f[, "Time"] == time &
        f[, "Sex"] == sex &
        f[, "O.t"] == o.t &
        f[, "D.t"] == d.t,
      "freq"] <- freq

  }

  # Overwrite incomplete Frequency table
  Frequency.cs.sam1 <- f; rm(f)

  # Calculate mean frequency at each time
  for (i in 1:nrow(Mean.frequency.sam1)) {

    time <- Mean.frequency.sam1[i, "Time"]
    sex <- Mean.frequency.sam1[i, "Sex"]
    o.t <- Mean.frequency.sam1[i, "O.t"]
    d.t <- Mean.frequency.sam1[i, "D.t"]

    avg <- mean(subset(Frequency.cs.sam1, Time == time &
                         Sex == sex &
                         O.t == o.t &
                         D.t == d.t)$freq)

    Mean.frequency.sam1[Mean.frequency.sam1[, "Time"] == time &
                          Mean.frequency.sam1[, "Sex"] == sex &
                          Mean.frequency.sam1[, "O.t"] == o.t &
                          Mean.frequency.sam1[, "D.t"] == d.t,
                        "freq"] <- avg

  }

  # (b) Prevalence -------------------------------------------------------------

  # Calculate mean obesity prevalence at each time
  for (i in 1:nrow(Mean.obes.prev.sam1)) {

    time <- Mean.obes.prev.sam1[i, "Time"]
    sub <- Mean.obes.prev.sam1[i, "Subgroup"]

    avg <- mean(subset(Obes.prev.sam1, Time == time & Subgroup == sub)$prev)

    Mean.obes.prev.sam1[Mean.obes.prev.sam1[, "Time"] == time &
                          Mean.obes.prev.sam1[, "Subgroup"] == sub,
                        "prev"] <- avg
```

```
}

# Calculate mean diabetes prevalence at each time
for (i in 1:nrow(Mean.diab.prev.sam1)) {

  time <- Mean.diab.prev.sam1[i, "Time"]
  sub <- Mean.diab.prev.sam1[i, "Subgroup"]

  avg <- mean(subset(Diab.prev.sam1, Time == time & Subgroup == sub)$prev)

  Mean.diab.prev.sam1[Mean.diab.prev.sam1[, "Time"] == time &
                        Mean.diab.prev.sam1[, "Subgroup"] == sub,
                      "prev"] <- avg

}

# (c) Conditional probabilities ------------------------------------------------

## (i) Cross-sectional ---------------------------------------------------------

# Calculate mean CP of obesity at each time point
for (i in 1:nrow(Mean.CP.Obes.cs.sam1)) {

  time <- Mean.CP.Obes.cs.sam1[i, "Time"]
  sex <- Mean.CP.Obes.cs.sam1[i, "Sex"]
  o.t <- Mean.CP.Obes.cs.sam1[i, "O.t"]

  avg <- mean(subset(CProbability.Obes.cs.sam1, Time == time &
                       Sex == sex &
                       O.t == o.t)$prob)

  Mean.CP.Obes.cs.sam1[Mean.CP.Obes.cs.sam1[, "Time"] == time &
                         Mean.CP.Obes.cs.sam1[, "Sex"] == sex &
                         Mean.CP.Obes.cs.sam1[, "O.t"] == o.t,
                       "prob"] <- avg

}

# Calculate mean CP of diabetes at each time point
for (i in 1:nrow(Mean.CP.Diab.cs.sam1)) {

  time <- Mean.CP.Diab.cs.sam1[i, "Time"]
  sex <- Mean.CP.Diab.cs.sam1[i, "Sex"]
  o.t <- Mean.CP.Diab.cs.sam1[i, "O.t"]
  d.t <- Mean.CP.Diab.cs.sam1[i, "D.t"]

  avg <- mean(subset(CProbability.Diab.cs.sam1, Time == time &
                       Sex == sex &
                       O.t == o.t &
                       D.t == d.t)$prob)

  Mean.CP.Diab.cs.sam1[Mean.CP.Diab.cs.sam1[, "Time"] == time &
                         Mean.CP.Diab.cs.sam1[, "Sex"] == sex &
                         Mean.CP.Diab.cs.sam1[, "O.t"] == o.t &
                         Mean.CP.Diab.cs.sam1[, "D.t"]== d.t,
                       "prob"] <- avg

}

## (ii) Cross-time -------------------------------------------------------------

# Calculate mean CP of obesity at each time point
for (i in 1:nrow(Mean.CP.Obes.ct.sam1)) {

  time <- Mean.CP.Obes.ct.sam1[i, "Time"]
  sex <- Mean.CP.Obes.ct.sam1[i, "Sex"]
  o.tminus1 <- Mean.CP.Obes.ct.sam1[i, "O.tminus1"]
  d.tminus1 <- Mean.CP.Obes.ct.sam1[i, "D.tminus1"]
  o.t <- Mean.CP.Obes.ct.sam1[i, "O.t"]

  avg <- mean(subset(CProbability.Obes.ct.sam1, Time == time &
                       Sex == sex &
                       O.tminus1 == o.tminus1 &
                       D.tminus1 == d.tminus1 &
                       O.t == o.t)$prob)

  Mean.CP.Obes.ct.sam1[Mean.CP.Obes.ct.sam1[, "Time"] == time &
                         Mean.CP.Obes.ct.sam1[, "Sex"] == sex &
                         Mean.CP.Obes.ct.sam1[, "O.tminus1"] == o.tminus1 &
                         Mean.CP.Obes.ct.sam1[, "D.tminus1"] == d.tminus1 &
                         Mean.CP.Obes.ct.sam1[, "O.t"] == o.t,
                       "prob"] <- avg

}

# Calculate mean CP of diabetes at each time point
for (i in 1:nrow(Mean.CP.Diab.ct.sam1)) {

  time <- Mean.CP.Diab.ct.sam1[i, "Time"]
  sex <- Mean.CP.Diab.ct.sam1[i, "Sex"]
  d.tminus1 <- Mean.CP.Diab.ct.sam1[i, "D.tminus1"]
  o.t <- Mean.CP.Diab.ct.sam1[i, "O.t"]
  d.t <- Mean.CP.Diab.ct.sam1[i, "D.t"]

  avg <- mean(subset(CProbability.Diab.ct.sam1, Time == time &
                       Sex == sex &
                       D.tminus1 == d.tminus1 &
                       O.t == o.t &
                       D.t == d.t)$prob)

  Mean.CP.Diab.ct.sam1[Mean.CP.Diab.ct.sam1[, "Time"] == time &
                         Mean.CP.Diab.ct.sam1[, "Sex"] == sex &
```

```
                            Mean.CP.Diab.ct.sam1[, "D.tminus1"] == d.tminus1 &
                            Mean.CP.Diab.ct.sam1[, "O.t"] == o.t &
                            Mean.CP.Diab.ct.sam1[, "D.t"] == d.t,
                          "prob"] <- avg

  }

  # (d) Transition parameters ------------------------------------------------

  # Calculate mean TP at each time point
  for (i in 1:nrow(Mean.TP.sam1)) {

    par <- Mean.TP.sam1[i, "Parameter"]

    avg <- mean(subset(TransitionParameters.sam1, Parameter == par)$value)

    Mean.TP.sam1[Mean.TP.sam1[, "Parameter"] == par,
                 "value"] <- avg

  }

  # Save summary tables for Intervention z -----------------------------------

  # Define file location
  if (z == 0) {  # (natural history)
    path <- paste0("./Microsimulation models/Time 1 transition probs/MSM 1/Natural history/")
  } else {  # (intervention z)
    path <- paste0("./Microsimulation models/Time 1 transition probs/MSM 1/Intervention ", z, "/")
  }

  # Define file names
  if (z == 0) {  # (natural history)
    file.freq <- paste0("Sam1Freq.csv")  # frequency
    file.prev.o <- paste0("Sam1ObesPrev.csv")  # obesity prevalence
    file.prev.d <- paste0("Sam1DiabPrev.csv")  # diabetes prevalence
    file.cpcs.o <- paste0("Sam1ObesCPcs.csv")  # CP obesity (cross-sectional)
    file.cpct.o <- paste0("Sam1ObesCPct.csv")  # CP obesity (cross-time)
    file.cpcs.d <- paste0("Sam1DiabCPcs.csv")  # CP diabetes (cross-sectional)
    file.cpct.d <- paste0("Sam1DiabCPct.csv")  # CP diabetes (cross-time)
    file.tp <- paste0("Sam1TP.csv")  # transition parameters
    file.m.freq <- paste0("Sam1FreqMean.csv")  # mean frequency
    file.m.prev.o <- paste0("Sam1ObesPrevMean.csv")  # mean obesity prevalence
    file.m.prev.d <- paste0("Sam1DiabPrevMean.csv")  # mean diabetes prevalence
    file.m.cpcs.o <- paste0("Sam1ObesCPcsMean.csv")  # mean CP obesity (cross-sectional)
    file.m.cpct.o <- paste0("Sam1ObesCPctMean.csv")  # mean CP obesity (cross-time)
    file.m.cpcs.d <- paste0("Sam1DiabCPcsMean.csv")  # mean CP diabetes (cross-sectional)
    file.m.cpct.d <- paste0("Sam1DiabCPctMean.csv")  # mean CP diabetes (cross-time)
    file.m.tp <- paste0("Sam1TPMean.csv")  # mean transition parameters
  } else {  # (intervention z)
    file.freq <- paste0("Sam1FreqInt", z, ".csv")  # frequency
    file.prev.o <- paste0("Sam1ObesPrevInt", z, ".csv")  # obesity prevalence
    file.prev.d <- paste0("Sam1DiabPrevInt", z, ".csv")  # diabetes prevalence
    file.cpcs.o <- paste0("Sam1ObesCPcsInt", z, ".csv")  # CP obesity (cross-sectional)
    file.cpct.o <- paste0("Sam1ObesCPctInt", z, ".csv")  # CP obesity (cross-time)
    file.cpcs.d <- paste0("Sam1DiabCPcsInt", z, ".csv")  # CP diabetes (cross-sectional)
    file.cpct.d <- paste0("Sam1DiabCPctInt", z, ".csv")  # CP diabetes (cross-time)
    file.tp <- paste0("Sam1TPInt", z, ".csv")  # transition parameters
    file.m.freq <- paste0("Sam1FreqMeanInt", z, ".csv")  # mean frequency
    file.m.prev.o <- paste0("Sam1ObesPrevMeanInt", z, ".csv")  # mean obesity prevalence
    file.m.prev.d <- paste0("Sam1DiabPrevMeanInt", z, ".csv")  # mean diabetes prevalence
    file.m.cpcs.o <- paste0("Sam1ObesCPcsMeanInt", z, ".csv")  # mean CP obesity (cross-sectional)
    file.m.cpct.o <- paste0("Sam1ObesCPctMeanInt", z, ".csv")  # mean CP obesity (cross-time)
    file.m.cpcs.d <- paste0("Sam1DiabCPcsMeanInt", z, ".csv")  # mean CP diabetes (cross-sectional)
    file.m.cpct.d <- paste0("Sam1DiabCPctMeanInt", z, ".csv")  # mean CP diabetes (cross-time)
    file.m.tp <- paste0("Sam1TPMeanInt", z, ".csv")  # mean transition parameters
  }

  # Export frequency table
  write.csv(Frequency.cs.sam1, file = paste0(path, file.freq), row.names = FALSE)

  # Export prevalence tables
  write.csv(Obes.prev.sam1, file = paste0(path, file.prev.o), row.names = FALSE)
  write.csv(Diab.prev.sam1, file = paste0(path, file.prev.d), row.names = FALSE)

  # Export conditional probability tables
  write.csv(CProbability.Obes.cs.sam1, file = paste0(path, file.cpcs.o), row.names = FALSE)
  write.csv(CProbability.Obes.ct.sam1, file = paste0(path, file.cpct.o), row.names = FALSE)
  write.csv(CProbability.Diab.cs.sam1, file = paste0(path, file.cpcs.d), row.names = FALSE)
  write.csv(CProbability.Diab.ct.sam1, file = paste0(path, file.cpct.d), row.names = FALSE)

  # Export transition parameter tables
  write.csv(TransitionParameters.sam1, file = paste0(path, file.tp), row.names = FALSE)

  # Export mean trend tables
  write.csv(Mean.frequency.sam1, file = paste0(path, file.m.freq), row.names = FALSE)
  write.csv(Mean.obes.prev.sam1, file = paste0(path, file.m.prev.o), row.names = FALSE)
  write.csv(Mean.diab.prev.sam1, file = paste0(path, file.m.prev.d), row.names = FALSE)
  write.csv(Mean.CP.Obes.cs.sam1, file = paste0(path, file.m.cpcs.o), row.names = FALSE)
  write.csv(Mean.CP.Obes.ct.sam1, file = paste0(path, file.m.cpct.o), row.names = FALSE)
  write.csv(Mean.CP.Diab.cs.sam1, file = paste0(path, file.m.cpcs.d), row.names = FALSE)
  write.csv(Mean.CP.Diab.ct.sam1, file = paste0(path, file.m.cpct.d), row.names = FALSE)
  write.csv(Mean.TP.sam1, file = paste0(path, file.m.tp), row.names = FALSE)

}  # (close intervention loop - 1)

comp.time <- Sys.time() - v; comp.time  # print total simulation time
# (~39 seconds per simulation run of 20000 individuals)
# (~30 mins per 100 simulation runs of 20000 individuals)
```

Note that the above code relates to AS1 (i.e. the true data-generating process of the population); for all other autocorrelation structures, the function which estimates the transition probabilities at time *t* from a sample of individuals from the population (lines 331 – 405) changes based on the autocorrelation structure that is modelled. For AS2, this function is:

```
331 EstimateTransitionProbs <- function(sampledata) {
332
333   # Create dataframe for sample data (baseline & time 1)
334   vars <- c("Sex", "O.t0", "D.t0", "O.t1", "D.t1")  # define variables
335   sampledata.t1 <- data.frame(sampledata[, vars])
336   names(sampledata.t1) <- c("Sex", "O.tminus1", "D.tminus1", "O.t", "D.t")  # rename variables
337   sampledata.t1 <- data.frame(apply(sampledata.t1, 2, factor))  # convert vars to factors
338
339   # Calculate cross-time conditional probabilities & define transition parameters
340   # (1) Obesity
341   var.d <- "O.t"  # define dependent variable
342   var.i <- c("Sex", "O.tminus1")  # define independent variables
343   formula <- as.formula(paste(var.d, paste(var.i, collapse = " + "), sep = " ~ "))
344   CP.Obes <- data.frame(am_adt(cpt(formula, data = sampledata.t1)))
345   CP.Obes <- rename(CP.Obes, replace = c("N" = "prob"))  # rename prob column
346   CP.Obes <- subset(CP.Obes, O.t == "1")  # remove 'complement' rows
347   CP.Obes <- subset(CP.Obes, select = -O.t)  # remove O.t column
348   a0 <<- CP.Obes[CP.Obes[, "O.tminus1"] == 0 &
349                  CP.Obes[, "Sex"] == 0, "prob"]
350   a2 <<- CP.Obes[CP.Obes[, "O.tminus1"] == 0 &
351                  CP.Obes[, "Sex"] == 1, "prob"]
352   a4 <<- CP.Obes[CP.Obes[, "O.tminus1"] == 1 &
353                  CP.Obes[, "Sex"] == 0, "prob"]
354   a6 <<- CP.Obes[CP.Obes[, "O.tminus1"] == 1 &
355                  CP.Obes[, "Sex"] == 1, "prob"]
356   # (2) Diabetes
357   var.d <- "D.t"  # define dependent variable
358   var.i <- c("Sex", "D.tminus1", "O.t")  # define independent variables
359   formula <- as.formula(paste(var.d, paste(var.i, collapse = " + "), sep = " ~ "))
360   CP.Diab <- data.frame(am_adt(cpt(formula, data = sampledata.t1)))
361   CP.Diab <- rename(CP.Diab, replace = c("N" = "prob"))  # rename prob column
362   CP.Diab <- subset(CP.Diab, D.t == "1")  # remove 'complement' rows
363   CP.Diab <- subset(CP.Diab, select = -D.t)  # remove D.t column
364   b0 <<- CP.Diab[CP.Diab[, "D.tminus1"] == 0 &
365                  CP.Diab[, "Sex"] == 0 &
366                  CP.Diab[, "O.t"] == 0, "prob"]
367   b1 <<- CP.Diab[CP.Diab[, "D.tminus1"] == 0 &
368                  CP.Diab[, "Sex"] == 0 &
369                  CP.Diab[, "O.t"] == 1, "prob"]
370   b2 <<- CP.Diab[CP.Diab[, "D.tminus1"] == 0 &
371                  CP.Diab[, "Sex"] == 1 &
372                  CP.Diab[, "O.t"] == 0, "prob"]
373   b3 <<- CP.Diab[CP.Diab[, "D.tminus1"] == 0 &
374                  CP.Diab[, "Sex"] == 1 &
375                  CP.Diab[, "O.t"] == 1, "prob"]
376   b4 <<- CP.Diab[CP.Diab[, "D.tminus1"] == 1 &
377                  CP.Diab[, "Sex"] == 0 &
378                  CP.Diab[, "O.t"] == 0, "prob"]
379   b5 <<- CP.Diab[CP.Diab[, "D.tminus1"] == 1 &
380                  CP.Diab[, "Sex"] == 0 &
381                  CP.Diab[, "O.t"] == 1, "prob"]
382   b6 <<- CP.Diab[CP.Diab[, "D.tminus1"] == 1 &
383                  CP.Diab[, "Sex"] == 1 &
384                  CP.Diab[, "O.t"] == 0, "prob"]
385   b7 <<- CP.Diab[CP.Diab[, "D.tminus1"] == 1 &
386                  CP.Diab[, "Sex"] == 1 &
387                  CP.Diab[, "O.t"] == 1, "prob"]
388
389 }
```

For AS3, this function is:

```
331 EstimateTransitionProbs <- function(sampledata) {
332
333   # Create dataframe for sample data (time 1)
334   vars <- c("Sex", "O.t1", "D.t1")  # define variables
335   sampledata.t1 <- data.frame(sampledata[, vars])
336   names(sampledata.t1) <- c("Sex", "O.t", "D.t")  # rename variables
337   sampledata.t1 <- data.frame(apply(sampledata.t1, 2, factor))  # convert vars to factors
338
339   # Calculate cross-time conditional probabilities & define transition parameters
340   # (1) Obesity
341   var.d <- "O.t"  # define dependent variable
342   var.i <- "Sex"  # define independent variable
343   formula <- as.formula(paste(var.d, paste(var.i, collapse = " + "), sep = " ~ "))
344   CP.Obes <- data.frame(am_adt(cpt(formula, data = sampledata.t1)))
345   CP.Obes <- rename(CP.Obes, replace = c("N" = "prob"))  # rename prob column
346   CP.Obes <- subset(CP.Obes, O.t == "1")  # remove 'complement' rows
347   CP.Obes <- subset(CP.Obes, select = -O.t)  # remove O.t column
348   a0 <<- CP.Obes[CP.Obes[, "Sex"] == 0, "prob"]
349   a2 <<- CP.Obes[CP.Obes[, "Sex"] == 1, "prob"]
350   # (2) Diabetes
351   var.d <- "D.t"  # define dependent variable
352   var.i <- c("Sex", "O.t")  # define independent variables
353   formula <- as.formula(paste(var.d, paste(var.i, collapse = " + "), sep = " ~ "))
354   CP.Diab <- data.frame(am_adt(cpt(formula, data = sampledata.t1)))
355   CP.Diab <- rename(CP.Diab, replace = c("N" = "prob"))  # rename prob column
356   CP.Diab <- subset(CP.Diab, D.t == "1")  # remove 'complement' rows
```

```
  CP.Diab <- subset(CP.Diab, select = -D.t)  # remove D.t column
  b0 <<- CP.Diab[CP.Diab[, "Sex"] == 0 &
                           CP.Diab[, "O.t"] == 0, "prob"]
  b1 <<- CP.Diab[CP.Diab[, "Sex"] == 0 &
                           CP.Diab[, "O.t"] == 1, "prob"]
  b2 <<- CP.Diab[CP.Diab[, "Sex"] == 1 &
                           CP.Diab[, "O.t"] == 0, "prob"]
  b3 <<- CP.Diab[CP.Diab[, "Sex"] == 1 &
                           CP.Diab[, "O.t"] == 1, "prob"]

}
```

The output from each simulation is then saved to its subfolder ('MSM 2' and 'MSM 3', respectively).

C.2.3 Sensitivity analyses

In this subsection, we provide details relating to the sensitivity analyses performed, which are described in Section 6.4.4. We describe the simulation parameters for all sensitivity analyses (§C.2.3.1), and we present the results of all sensitivity analyses (§C.2.3.2).

C.2.3.1 Simulation parameters

Five sensitivity analyses were performed, in which simulation parameters governing the natural history of the population were altered.

In Sensitivity analysis 1, baseline diabetes prevalence was increased by three times across all subgroups. In Sensitivity analysis 2, diabetes incidence was increased by four times across all subgroups. In sensitivity analysis 3, the effect of previous diabetes on obesity was increased fifteen times for non-obese individuals and two times for obese individuals. Sensitivity analysis 4 combined the changes of Sensitivity analyses 1 and 2, whereas Sensitivity analysis 5 combined the changes of Sensitivity analyses 2 and 3.

Parameters describing the distribution of sex, obesity, and diabetes at baseline (i.e. $t = 0$) for all sensitivity analyses are given in Table C.4. Parameters describing the evolution of the baseline population (i.e. the transition parameters) for all subsequent time points (i.e. time t, for $1 \leq t \leq 10$) for all sensitivity analyses are given in Table C.5.

Table C.4 Parameters describing the joint distribution of sex, obesity, and diabetes in the baseline population (i.e. time $t = 0$) for each sensitivity analysis, compared to the original simulation

		Probability					
Status	**Covariate(s)**	**Original simulation**	**Sensitivity analysis 1**	**Sensitivity analysis 2**	**Sensitivity analysis 3**	**Sensitivity analysis 4**	**Sensitivity analysis 5**
Male	n/a	0.521	0.521	0.521	0.521	0.521	0.521
Obese	Female	0.490	0.490	0.490	0.490	0.490	0.490
	Male	0.580	0.580	0.580	0.580	0.580	0.580
Diabetic	Female, non-obese	0.010	0.030	0.010	0.010	0.030	0.010
	Female, obese	0.030	0.090	0.030	0.030	0.090	0.030
	Male, non-obese	0.017	0.051	0.017	0.017	0.051	0.017
	Male, obese	0.037	0.111	0.037	0.037	0.111	0.037

Changes to the original simulation parameters (from Table C.1) are highlighted in red for easy identification.

Table C.5 Transition parameters describing the evolution of the baseline population (i.e. time t, for $1 \leq t \leq 10$) for each sensitivity analysis, compared to the original simulation

			Probability					
Updated status	**Current status**	**Current covariates**	**Original simulation**	**Sensitivity analysis 1**	**Sensitivity analysis 2**	**Sensitivity analysis 3**	**Sensitivity analysis 4**	**Sensitivity analysis 5**
Obese	Non-obese	Female, non-diabetic	0.07500	0.07500	0.07500	0.07500	0.07500	0.07500
		Female, diabetic	0.10500	0.10500	0.10500	0.52500	0.52500	0.52500
		Male, non-diabetic	0.10000	0.10000	0.10000	0.10000	0.10000	0.10000
		Male, diabetic	0.13000	0.13000	0.13000	0.55000	0.55000	0.55000
	Obese	Female, non-diabetic	0.95000	0.95000	0.95000	0.95000	0.95000	0.95000
		Female, diabetic	0.97000	0.97000	0.97000	0.99000	0.99000	0.99000
		Male, non-diabetic	0.95000	0.95000	0.95000	0.95000	0.95000	0.95000
		Male, diabetic	0.97000	0.97000	0.97000	0.99000	0.99000	0.99000
Diabetic	Non-diabetic	Female, non-obese	0.00060	0.00060	0.0024	0.00060	0.00060	0.0024
		Female, obese	0.00260	0.00260	0.0104	0.00260	0.00260	0.0104
		Male, non-obese	0.00065	0.00065	0.0026	0.00065	0.00065	0.0026
		Male, obese	0.00265	0.00265	0.0106	0.00265	0.00265	0.0106
	Diabetic	Female, non-obese	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
		Female, obese	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
		Male, non-obese	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
		Male, obese	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000

Changes to the original simulation parameters (from Table C.2) are highlighted in red for easy identification.

C.2.3.2 Results

Here, we present the results obtained by using the g-formula and microsimulation to estimate the causal effects of each intervention for Sensitivity analyses 1 through 4 (the results of Sensitivity analysis 5 are presented in Chapter 6, Section 6.4.4.2).

Tables which compare the true causal effect of each intervention to those estimated by the g-formula and microsimulation, for each of AS1 through AS3), are shown in Table C.6 (Sensitivity analysis 1), Table C.7 (Sensitivity analysis 2), Table C.8 (Sensitivity analysis 3), and Table C.9 (Sensitivity analysis 4).

Table C.6 Table describing the estimated causal effect of each intervention on diabetes prevalence for each autocorrelation structure modelled using the g-formula and microsimulation, compared to the true effect in the population (Sensitivity analysis 1)

		AS1		AS2		AS3	
Effect	**True**	**g-formula**	**MSM**	**g-formula**	**MSM**	**g-formula**	**MSM**
Effect of Intervention 1 *(prevent individuals from being obese)*	**-0.0105** ***(-11.8%)***	-0.0107 *(-12.0%)*	-0.0103 *(-11.6%)*	-0.0110 *(-12.3%)*	-0.0112 *(-12.6%)*	-0.0398 *(-44.7%)*	-0.0327 *(-44.1%)*
Effect of Intervention 2 *(make all individuals obese)*	**0.0077** ***(8.7%)***	0.0077 *(8.6%)*	0.0075 *(8.4%)*	0.0075 *(8.4%)*	0.0076 *(8.5%)*	0.0248 *(27.9%)*	0.0275 *(37.0%)*
Effect of Intervention 3 *(prevent anyone from becoming obese)*	**-0.0033** ***(-3.7%)***	-0.0032 *(-3.6%)*	-0.0030 *(-3.4%)*	-0.0036 *(-4.0%)*	-0.0037 *(-4.2%)*	-0.0402 *(-45.2%)*	-0.0330 *(-44.4%)*
Effect of Intervention 4 *(reduce probability of becoming obese by 15%)*	**-0.0004** ***(-0.4%)***	-0.0005 *(-0.5%)*	-0.0004 *(-0.4%)*	-0.0004 *(-0.4%)*	-0.0005 *(-0.5%)*	-0.0055 *(-6.2%)*	-0.0045 *(-6.1%)*
Effect of Intervention 5 *(reduce probability of remaining obese by 10%)*	**-0.0029** ***(-3.2%)***	-0.0029 *(-3.3%)*	-0.0028 *(-3.2%)*	-0.0030 *(-3.4%)*	-0.0031 *(-3.5%)*	-0.0036 *(-4.1%)*	-0.0032 *(-4.3%)*
Effect of Intervention 6 *(reduce probability of becoming obese by 15% and remaining obese by 10%)*	**-0.0033** ***(-3.7%)***	-0.0034 *(-3.8%)*	-0.0033 *(-3.7%)*	-0.0035 *(-3.9%)*	-0.0037 *(-4.1%)*	-0.0091 *(-10.2%)*	-0.0074 *(-10.0%)*
Total causal effect (TCE)	**0.0183** ***(23.3%)***	0.0183 *(23.4%)*	0.0178 *(22.7%)*	0.0184 *(23.5%)*	0.0188 *(24.2%)*	0.0646 *(131.5%)*	0.0602 *(145.0%)*

The estimated causal effect of each intervention (1 through 6) on diabetes prevalence was calculated by subtracting the average observed diabetes prevalence at time 10 under the natural history from the average diabetes prevalence at time 10 that was observed when the given intervention was applied to a random sample of 20,000 individuals. The TCE was calculated by subtracting the average diabetes prevalence at time 10 that was observed when Intervention 1 was applied from that which was observed when Intervention 2 was applied. All effects are additionally expressed as percentage changes.

Table C.7 Table describing the estimated causal effect of each intervention on diabetes prevalence for each autocorrelation structure modelled using the g-formula and microsimulation, compared to the true effect in the population (Sensitivity analysis 2)

		AS1		AS2		AS3	
Effect	**True**	**g-formula**	**MSM**	**g-formula**	**MSM**	**g-formula**	**MSM**
Effect of Intervention 1 *(prevent individuals from being obese)*	**-0.0432** ***(-47.2%)***	-0.0434 *(-47.3%)*	-0.0434 *(-47.3%)*	-0.0437 *(-47.5%)*	-0.0443 *(-48.4%)*	-0.0470 *(-51.5%)*	-0.0153 *(-49.2%)*
Effect of Intervention 2 *(make all individuals obese)*	**0.0304** ***(33.1%)***	0.0309 *(33.7%)*	0.0308 *(33.6%)*	0.0301 *(32.8%)*	0.0310 *(33.9%)*	0.0296 *(32.4%)*	0.0123 *(39.3%)*
Effect of Intervention 3 *(prevent anyone from becoming obese)*	**-0.0134** ***(-14.6%)***	-0.0129 *(-14.1%)*	-0.013 *(-14.2%)*	-0.0134 *(-14.6%)*	-0.0135 *(-14.8%)*	-0.0474 *(-51.9%)*	-0.0153 *(-49.2%)*
Effect of Intervention 4 *(reduce probability of becoming obese by 15%)*	**-0.0017** ***(-1.8%)***	-0.0015 *(-1.7%)*	-0.0016 *(-1.8%)*	-0.0013 *(-1.4%)*	-0.0013 *(-1.4%)*	-0.0066 *(-7.3%)*	-0.0023 *(-7.5%)*
Effect of Intervention 5 *(reduce probability of remaining obese by 10%)*	**-0.0118** ***(-12.9%)***	-0.0116 *(-12.6%)*	-0.0116 *(-12.6%)*	-0.0118 *(-12.8%)*	-0.0118 *(-12.9%)*	-0.0043 *(-4.7%)*	-0.0018 *(-5.6%)*
Effect of Intervention 6 *(reduce probability of becoming obese by 15% and remaining obese by 10%)*	**-0.0134** ***(-14.6%)***	-0.0133 *(-14.5%)*	-0.0133 *(-14.5%)*	-0.0136 *(-14.8%)*	-0.0136 *(-14.9%)*	-0.0108 *(-11.9%)*	-0.0038 *(-12.2%)*
Total causal effect (TCE)	**0.0736** ***(151.9%)***	0.0743 *(153.8%)*	0.0743 *(153.5%)*	0.0738 *(153.1%)*	0.0752 *(159.6%)*	0.0766 *(172.9%)*	0.0276 *(174.4%)*

The estimated causal effect of each intervention (1 through 6) on diabetes prevalence was calculated by subtracting the average observed diabetes prevalence at time 10 under the natural history from the average diabetes prevalence at time 10 that was observed when the given intervention was applied to a random sample of 20,000 individuals. The TCE was calculated by subtracting the average diabetes prevalence at time 10 that was observed when Intervention 1 was applied from that which was observed when Intervention 2 was applied. All effects are additionally expressed as percentage changes.

Table C.8 Table describing the estimated causal effect of each intervention on diabetes prevalence for each autocorrelation structure modelled using the g-formula and microsimulation, compared to the true effect in the population (Sensitivity analysis 3)

		AS1		AS2		AS3	
Effect	**True**	**g-formula**	**MSM**	**g-formula**	**MSM**	**g-formula**	**MSM**
Effect of Intervention 1 *(prevent individuals from being obese)*	**-0.0112** ***(26.9%)***	-0.0116 *(-27.8%)*	-0.0110 *(-26.6%)*	-0.0116 *(-27.8%)*	-0.0120 *(-28.7%)*	-0.0387 *(-92.8%)*	-0.0185 *(-71.0%)*
Effect of Intervention 2 *(make all individuals obese)*	**0.0080** ***(19.3%)***	0.0080 *(19.3%)*	0.0077 *(18.5%)*	0.0078 *(18.7%)*	0.0078 *(18.7%)*	0.0232 *(55.7%)*	0.0147 *(56.6%)*
Effect of Intervention 3 *(prevent anyone from becoming obese)*	**-0.0034** ***(-8.2%)***	-0.0034 *(-8.1%)*	-0.0032 *(-7.7%)*	-0.0036 *(-8.7%)*	-0.0039 *(-9.4%)*	-0.0388 *(-93.0%)*	-0.0186 *(-71.3%)*
Effect of Intervention 4 *(reduce probability of becoming obese by 15%)*	**-0.0004** ***(-0.9%)***	-0.0005 *(-1.2%)*	-0.0005 *(-1.1%)*	-0.0003 *(-0.8%)*	-0.0005 *(-1.1%)*	-0.0057 *(-13.8%)*	-0.0029 *(-11.0%)*
Effect of Intervention 5 *(reduce probability of remaining obese by 10%)*	**-0.0031** ***(-7.4%)***	-0.0032 *(-7.6%)*	-0.0031 *(-7.4%)*	-0.0032 *(-7.6%)*	-0.0033 *(-7.8%)*	-0.0038 *(-9.1%)*	-0.0019 *(-7.3%)*
Effect of Intervention 6 *(reduce probability of becoming obese by 15% and remaining obese by 10%)*	**-0.0035** ***(-8.3%)***	-0.0037 *(-8.8%)*	-0.0035 *(-8.5%)*	-0.0037 *(-8.8%)*	-0.0039 *(-9.2%)*	-0.0089 *(-21.4%)*	-0.0043 *(-16.7%)*
Total causal effect (TCE)	**0.0192** ***(63.2%)***	0.0196 *(65.2%)*	0.0187 *(61.5%)*	0.0194 *(65.4%)*	0.0198 *(66.5%)*	0.0619 *(2055.4%)*	0.0332 *(440.5%)*

The estimated causal effect of each intervention (1 through 6) on diabetes prevalence was calculated by subtracting the average observed diabetes prevalence at time 10 under the natural history from the average diabetes prevalence at time 10 that was observed when the given intervention was applied to a random sample of 20,000 individuals. The TCE was calculated by subtracting the average diabetes prevalence at time 10 that was observed when Intervention 1 was applied from that which was observed when Intervention 2 was applied. All effects are additionally expressed as percentage changes.

Table C.9 Table describing the estimated causal effect of each intervention on diabetes prevalence for each autocorrelation structure modelled using the g-formula and microsimulation, compared to the true effect in the population (Sensitivity analysis 4)

		AS1		AS2		AS3	
Effect	**True**	**g-formula**	**MSM**	**g-formula**	**MSM**	**g-formula**	**MSM**
Effect of Intervention 1 *(prevent individuals from being obese)*	**-0.0105** ***(-11.8%)***	-0.0107 *(-12.0%)*	-0.0103 *(-11.6%)*	-0.0112 *(-12.6%)*	-0.0118 *(-13.2%)*	-0.0836 *(-93.8%)*	-0.0525 *(70.6%)*
Effect of Intervention 2 *(make all individuals obese)*	**0.0077** ***(8.7%)***	0.0077 *(8.6%)*	0.0075 *(8.4%)*	0.0072 *(8.0%)*	0.0070 *(7.8%)*	0.0469 *(52.6%)*	0.0412 *(55.4%)*
Effect of Intervention 3 *(prevent anyone from becoming obese)*	**-0.0033** ***(-3.7%)***	-0.0032 *(-3.6%)*	-0.0030 *(-3.4%)*	-0.0037 *(-4.2%)*	-0.0042 *(-4.7%)*	-0.0837 *(-93.8%)*	-0.0526 *(-70.8%)*
Effect of Intervention 4 *(reduce probability of becoming obese by 15%)*	**-0.0004** ***(-0.4%)***	-0.0005 *(-0.5%)*	-0.0004 *(-0.4%)*	-0.0004 *(-0.5%)*	-0.0005 *(-0.6%)*	-0.0125 *(14.0%)*	-0.0077 *(-10.4%)*
Effect of Intervention 5 *(reduce probability of remaining obese by 10%)*	**-0.0029** ***(-3.2%)***	-0.0029 *(-3.3%)*	-0.0028 *(-3.2%)*	-0.0031 *(-3.5%)*	-0.0032 *(-3.5%)*	-0.0082 *(-9.2%)*	-0.0052 *(-7.0%)*
Effect of Intervention 6 *(reduce probability of becoming obese by 15% and remaining obese by 10%)*	**-0.0033** ***(-3.7%)***	-0.0034 *(-3.8%)*	-0.0033 *(-3.7%)*	-0.0036 *(-4.0%)*	-0.0037 *(-4.2%)*	-0.0194 *(-21.7%)*	-0.0123 *(-16.5%)*
Total causal effect (TCE)	**0.0183** ***(23.3%)***	0.0183 *(23.4%)*	0.0178 *(22.7%)*	0.0184 *(23.6%)*	0.0188 *(24.2%)*	0.1306 *(2343.1%)*	0.0937 *(429.0%)*

The estimated causal effect of each intervention (1 through 6) on diabetes prevalence was calculated by subtracting the average observed diabetes prevalence at time 10 under the natural history from the average diabetes prevalence at time 10 that was observed when the given intervention was applied to a random sample of 20,000 individuals. The TCE was calculated by subtracting the average diabetes prevalence at time 10 that was observed when Intervention 1 was applied from that which was observed when Intervention 2 was applied. All effects are additionally expressed as percentage changes.

References

1. Arnold, K.F., Berrie, L., Tennant, P.W.G. and Gilthorpe, M.S. A causal inference perspective on the analysis of compositional data. *International Journal of Epidemiology.* 2020, **0**(0), pp.1-7.
2. Arnold, K.F., Davies, V., de Kamps, M., Tennant, P.W.G., Mbotwa, J. and Gilthorpe, M.S. Reflections on modern methods: Generalised linear models for prognosis and intervention – theory, practice, and implications for machine learning. *International Journal of Epidemiology.* 2020, **0**(0).
3. Arnold, K.F., Harrison, W.J., Heppenstall, A.J. and Gilthorpe, M.S. DAG-informed regression modelling, agent-based modelling and microsimulation modelling: a critical comparison of methods for causal inference. *International Journal of Epidemiology.* 2019, **48**(1), pp.243-253.
4. Tennant, P.W.G., Arnold, K.F., Ellison, G.T.H. and Gilthorpe, M.S. Analyses of 'change scores' do not estimate causal effects in observational data. *ArXiv e-prints.* [Online]. 2019. [Accessed July 01, 2019]. Available from: https://ui.adsabs.harvard.edu/abs/2019arXiv190702764T
5. Arnold, K.F., Ellison, G.T.H., Gadd, S.C., Textor, J., Tennant, P.W.G., Heppenstall, A. and Gilthorpe, M.S. Adjustment for time-invariant and time-varying confounders in 'unexplained residuals' models for longitudinal data within a causal framework and associated challenges. *Statistical Methods in Medical Research.* 2019, **28**(5), pp.1347-1364.
6. Hernán, M.A. and Robins, J.M. *Causal Inference: What If.* Boca Raton: Chapman & Hall/CRC, 2020.
7. Hill, A.B. The Environment and Disease: Association or Causation? *Proceedings of the Royal Society of Medicine.* 1965, **58**(5), pp.295-300.
8. Krieger, N. and Davey Smith, G. The tale wagged by the DAG: broadening the scope of causal inference and explanation for epidemiology. *International Journal of Epidemiology.* 2016, **45**(6), pp.1787-1808.
9. Vandenbroucke, J.P., Broadbent, A. and Pearce, N. Causality and causal inference in epidemiology: the need for a pluralistic approach. *International Journal of Epidemiology.* 2016, **22**, p.22.
10. Daniel, R.M., De Stavola, B.L. and Vansteelandt, S. The formal approach to quantitative causal inference in epidemiology: misguided or misrepresented? *International Journal of Epidemiology.* 2017, **45**(6), pp.1817-1829.
11. Weed, D.L. Causal inference in epidemiology: potential outcomes, pluralism and peer review. *International Journal of Epidemiology.* 2017, **27**, p.27.
12. VanderWeele, T.J. On Causes, Causal Inference, and Potential Outcomes. *International Journal of Epidemiology.* 2017, **27**, p.27.
13. Robins, J.M. and Weissman, M.B. Counterfactual causation and streetlamps: what is to be done? *International Journal of Epidemiology.* 2017, **27**, p.27.
14. Krieger, N. and Davey Smith, G. FACEing reality: productive tensions between our epidemiological questions, methods and mission. *International Journal of Epidemiology.* 2017, **45**(6), pp.1852–1865.
15. Broadbent, A., Vandenbroucke, J.P. and Pearce, N. Formalism or pluralism? A reply to commentaries on 'Causality and causal inference in epidemiology'. *International Journal of Epidemiology.* 2017, **27**, p.27.
16. *The Oxford Handbook of Causation.* 1 ed. New York: Oxford University Press, 2009.
17. Reiss, J. Causation in the Social Sciences: Evidence, Inference, and Purpose. *Philosophy of the Social Sciences.* 2009, **39**(1), pp.20-40.

18. Neyman, J. Sur les applications de la th´eorie des probabilit´es aux experiences agricoles: Essai des principes. In Polish, English translation by D. M. Dabrowska and T. P. Speed. . *Statistical Science.* 1923, **5**(4), pp.465-472.
19. Rubin, D.B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology.* 1974, **66**(5), pp.688-701.
20. Rubin, D.B. Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. *Journal of the American Statistical Association.* 2005, **100**(469), pp.322-331.
21. Galea, S., Riddle, M. and Kaplan, G.A. Causal thinking and complex system approaches in epidemiology. *International Journal of Epidemiology.* 2009, **39**(1), pp.97-106.
22. Mooney, S.J., Westreich, D.J. and El-Sayed, A.M. Epidemiology in the era of big data. *Epidemiology.* 2015, **26**(3), pp.390-394.
23. *A life course approach to chronic disease epidemiology.* 2 ed. Oxford: Oxford University Press, 2004.
24. Tu, Y.K., Tilling, K., Sterne, J.A.C. and Gilthorpe, M.S. A critical evaluation of statistical approaches to examining the role of growth trajectories in the developmental origins of health and disease. *International Journal of Epidemiology.* 2013, **42**(5), pp.1327-1339.
25. Kuchenbaecker, K.B., Hopper, J.L., Barnes, D.R., Phillips, K.-A., Mooij, T.M., Roos-Blom, M.-J., Jervis, S., van Leeuwen, F.E., Milne, R.L., Andrieu, N., Goldgar, D.E., Terry, M.B., Rookus, M.A., Easton, D.F., Antoniou, A.C., BRCA1, a.t. and Consortium, B.C. Risks of Breast, Ovarian, and Contralateral Breast Cancer for BRCA1 and BRCA2 Mutation Carriers. *JAMA.* 2017, **317**(23), pp.2402-2416.
26. Robins, J.M. and Hernán, M.A. Estimation of the causal effects of time-varying exposures. In: Fitzmaurice, G. et al. eds. *Longitudinal Data Analysis*. Boca Raton: Chapman & Hall/CRC, 2009, pp.553-599.
27. Pearl, J., Glymour, M. and Jewell, N.P. *Causal Inference in Statistics: A Primer.* 1 ed. Chichester: John Wiley & Sons Ltd, 2016.
28. Rubin, D.B. Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies. *Statistical Science.* 1990, **5**(4), pp.472-480.
29. Ding, P. *Exploring the Role of Randomization in Causal Inference*. PhD thesis, Harvard University, 2015.
30. Greenland, S. Randomization, Statistics, and Causal Inference. *Epidemiology.* 1990, **1**(6), pp.421-429.
31. Cartwright, N. What are randomised controlled trials good for? *Philosophical Studies.* 2009, **147**(1), p.59.
32. Senn, S. Seven myths of randomisation in clinical trials. *Statistics in Medicine.* 2013, **32**(9), pp.1439-1450.
33. Murray, E.J. and Hernán, M.A. Adherence adjustment in the Coronary Drug Project: A call for better per-protocol effect estimates in randomized trials. *Clinical Trials.* 2016, **13**(4), pp.372-378.
34. Howe, C.J., Cole, S.R., Lau, B., Napravnik, S. and Eron, J.J.J. Selection Bias Due to Loss to Follow Up in Cohort Studies. *Epidemiology.* 2016, **27**(1), pp.91-97.
35. Wright, S. On the nature of size factors. *Genetics.* 1918, **3**(1), pp.367-374.
36. Wright, S. The Method of Path Coefficients. *The Annals of Mathematical Statistics.* 1934, **5**(3), pp.161-215.
37. Tu, Y.K. Directed acyclic graphs and structural equation modelling. In: Tu, Y.K. and Greenwood, D.C. eds. *Modern Methods for Epidemiology*. Springer, 2012, pp.191-203.
38. Pearl, J. *Causality: Models, Reasoning, and Inference.* New York: Cambridge University Press, 2000.
39. Elwert, F. Graphical Causal Models. In: Morgan, S.L. ed. *Handbook of Causal Analysis for Social Research*. Dordrecht: Springer Science + Business, 2013, pp.245-273.
40. VanderWeele, T.J. *Explanation in Causal Inference: Methods for Mediation and Interaction.* 1 ed. New York: Oxford University Press, 2015.

41. Tennant, P.W.G., Arnold, K.F., Berrie, L., Ellison, G.T.H. and Gilthorpe, M.S. *Advanced Modelling Strategies: Challenges and pitfalls in robust causal inference with observational data.* [Online]. Leeds Institute for Data Analytics (LIDA), 2017.
42. Textor, J. and Gilthorpe, M.S. *Covariate Roles in DAGs.* [Online]. 2011. [Accessed 12 January 2020]. Available from: http://www.dagitty.net/learn/graphs/roles.html
43. van Leeuwen, R.W.F., Swart, E.L., Boven, E., Boom, F.A., Schuitenmaker, M.G. and Hugtenburg, J.G. Potential drug interactions in cancer therapy: a prevalence study using an advanced screening method. *Annals of Oncology.* 2011, **22**(10), pp.2334-2341.
44. *Who Gets Chemotherapy?* [Online]. 2020. [Accessed 30 January 2020].
45. Textor, J., Hardt, J. and Knüppel, S. DAGitty: A graphical tool for analyzing causal diagrams. *Epidemiology.* 2011, **22**(5).
46. Textor, J. and van der Zander, B. *dagitty: Graphical Analysis of Structural Causal Models*. R. 2016. Available from: https://CRAN.R-project.org/package=dagitty
47. Textor, J., van der Zander, B., Gilthorpe, M.S., Liskiewicz, M. and Ellison, G.T.H. Robust causal inference using directed acyclic graphs: the R package 'dagitty'. *International Journal of Epidemiology.* 2017, **15**, p.15.
48. Pearl, J. The algorithmization of counterfactuals. *Annals of Mathematics and Artificial Intelligence.* 2011, **61**(1), pp.29-39.
49. Rehkopf, D.H., Glymour, M.M. and Osypuk, T.L. The Consistency Assumption for Causal Inference in Social Epidemiology: When a Rose is Not a Rose. *Current Epidemiology Reports.* 2016, **3**(1), pp.63-71.
50. Cole, S.R. and Frangakis, C.E. Commentary: The Consistency Statement in Causal Inference: A Definition or an Assumption? *Epidemiology.* 2009, **20**(1), pp.3-5.
51. Pearl, J. On the Consistency Rule in Causal Inference: Axiom, Definition, Assumption, or Theorem? *Epidemiology.* 2010, **21**(6), pp.872-875.
52. Schwartz, S., Gatto, N.M. and Campbell, U.B. Causal identification: a charge of epidemiology in danger of marginalization. *Annals of Epidemiology.* 2016, **26**(10), pp.669-673.
53. Hernán, M.A. and Taubman, S.L. Does obesity shorten life? The importance of well-defined interventions to answer causal questions. *International Journal of Obesity.* 2008, **32**, pp.S8-S14.
54. Tchetgen, E.J.T. and VanderWeele, T.J. On causal inference in the presence of interference. *Statistical Methods in Medical Research.* 2010, **21**(1), pp.55-75.
55. Halloran, M.E. and Struchiner, C.J. Causal Inference in Infectious Diseases. *Epidemiology.* 1995, **6**(2), pp.142-151.
56. Naimi, A.I., Cole, S.R. and Kennedy, E.H. An introduction to g methods. *International Journal of Epidemiology.* 2017, **46**(2), pp.756-762.
57. Tennant, P.W., Harrison, W.J., Murray, E.J., Arnold, K.F., Berrie, L., Fox, M.P., Gadd, S.C., Keeble, C., Ranker, L.R., Textor, J., Tomova, G.D., Gilthorpe, M.S. and Ellison, G.T. Use of directed acyclic graphs (DAGs) in applied health research: review and recommendations [v1]. *medRxiv.* [Online]. 2019. [Accessed January 20, 2020]. Available from: https://www.medrxiv.org/content/medrxiv/early/2019/12/27/2019.12.20.19015511.full.pdf
58. Daniel, R.M., Cousens, S.N., De Stavola, B.L., Kenward, M.G. and Sterne, J.A.C. Methods for dealing with time-dependent confounding. *Statistics in Medicine.* 2013, **32**, pp.1584-1618.
59. Arnold, K.F. and Gilthorpe, M.S. Introduction to g-methods. In: Tennant, P.W.G. and Gilthorpe, M.S. eds. *Causal inference with observational data: The challenges and pitfalls [unpublished lecture notes]*. Leeds: Leeds Institute for Data Analytics (LIDA), 2019, pp.94-110.

60. Taubman, S.L., Robins, J.M., Mittleman, M.A. and Hernán, M.A. Intervening on risk factors for coronary heart disease: an application of the parametric g-formula. *International Journal of Epidemiology.* 2009, **38**(6), pp.1599-1611.
61. Robins, J.M., Hernán, M.Á. and Brumback, B. Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology.* 2000, **11**(5), pp.550-560.
62. Vansteelandt, S. and Joffe, M. Structural Nested Models and G-estimation: The Partially Realized Promise. *Statistical Science.* 2014, **29**(4), pp.707-731.
63. Picciotto, S. and Neophytou, A.M. G-Estimation of Structural Nested Models: Recent Applications in Two Subfields of Epidemiology. *Current Epidemiology Reports.* 2016, **3**, pp.242-251.
64. Westreich, D., Cole, S.R., Young, J.G., Palella, F., Tien, P.C., Kingsley, L., Gange, S.J. and Hernán, M.A. The parametric g-formula to estimate the effect of highly active antiretroviral therapy on incident AIDS or death. *Statistics in Medicine.* 2012, **31**(18), pp.2000-2009.
65. VanderWeele, T.J. On Causes, Causal Inference, and Potential Outcomes. *International Journal of Epidemiology.* 2017, **45**(6), pp.1809-1816.
66. Pearl, J. Understanding Simpson's paradox. *The American Statistician.* 2013, **88**, pp.8-13.
67. Pearl, J. *Lord's Paradox Revisited - (Oh Lord! Kumbaya!)*. Unpublished, 2016.
68. Arah, O.A. The role of causal reasoning in understanding Simpson's paradox, Lord's paradox, and the suppression effect: Covariate selection in the analysis of observational studies. *Emerging Themes in Epidemiology.* 2008, **5**(5), pp.1-5.
69. Tu, Y.K., Gunnell, D. and Gilthorpe, M.S. Simpson's Paradox, Lord's Paradox, and Suppression Effects are the same phenomenon - The reversal paradox. *Emerging Themes in Epidemiology.* 2008, **5**(2).
70. Arnold, K.F., Davies, V., de Kamps, M., Tennant, P.W.G., Mbotwa, J. and Gilthorpe, M.S. Generalised linear models for prognosis and intervention: Theory, practice, and implications for machine learning [v2]. *ArXiv e-prints.* [Online]. 2020. [Accessed January 20, 2020]. Available from: https://ui.adsabs.harvard.edu/abs/2019arXiv190601461A
71. Wilcox, A.J. On the importance—and the unimportance— of birthweight. *International Journal of Epidemiology.* 2001, **30**(6), pp.1233-1241.
72. Simpson, W.J. A preliminary report on cigarette smoking and the incidence of prematurity. *American Journal of Obstetrics & Gynecology.* 1957, **73**(4), pp.808-815.
73. Yerushalmy, J. The relationship of parents' cigarette smoking to outcome of pregnancy—implications as to the problem of inferring causation from observed associations1. *International Journal of Epidemiology.* 2014, **43**(5), pp.1355-1366.
74. Hernandez-Diaz, S., Schisterman, E.F. and Hernan, M.A. The Birth Weight "Paradox" Uncovered? *American Journal of Epidemiology.* 2006, **164**(11), pp.1115-1120.
75. Macmahon, B., Alpert, M. and Salber, E.J. Infant weight and parental smoking habits. *American Journal of Epidemiology.* 1965, **82**(3), pp.247-261.
76. Whitcomb, B.W., Schisterman, E.F., Perkins, N.J. and Platt, R.W. Quantification of collider-stratification bias and the birthweight paradox. *Paediatric and Perinatal Epidemiology.* 2009, **23**(5), pp.394-402.
77. VanderWeele, T.J., Mumford, S.L. and Schisterman, E.F. Conditioning on Intermediates in Perinatal Epidemiology. *Epidemiology.* 2012, **23**(1), pp.1-9.
78. VanderWeele, T.J. Commentary: Resolutions of the birthweight paradox: competing explanations and analytical insights. *International Journal of Epidemiology.* 2014, **43**(5), pp.1368-1373.
79. Kramer, M.S., Zhang, X. and Platt, R.W. Commentary: Yerushalmy, maternal cigarette smoking and the perinatal mortality crossover paradox. *International Journal of Epidemiology.* 2014, **43**(5), pp.1378-1381.

80. Aitchison, J. The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society: Series B.* 1982, **44**(2), pp.139-177.
81. Arnold, K.F., Berrie, L., Tennant, P.W.G. and Gilthorpe, M.S. A causal inference perspective on the analysis of compositional data. *International Journal of Epidemiology.* 2020, **0**(0), pp.1-11.
82. Shachter, R.D. Probabilistic Inference and Influence Diagrams. *Operations Research.* 1988, **36**(4), pp.589-604.
83. Hernán, M.A., Hsu, J. and Healy, B. A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks. *CHANCE.* 2019, **32**(1), pp.42-49.
84. Shmueli, G. To Explain or Predict? *Statistical Science.* 2010, **25**(3), pp.289-310.
85. Celis-Morales, C.A., Lyall, D.M., Welsh, P., Anderson, J., Steell, L., Guo, Y., Maldonado, R., Mackay, D.F., Pell, J.P., Sattar, N. and Gill, J.M.R. Association between active commuting and incident cardiovascular disease, cancer, and mortality: prospective cohort study. *BMJ.* 2017, **357**, p.j1456.
86. Westreich, D. and Greenland, S. The Table 2 Fallacy: Presenting and Interpreting Confounder and Modifier Coefficients. *American Journal of Epidemiology.* 2013, **177**(4), pp.292-298.
87. Breiman, L., Friedman, J., Stone, C.J. and Olshen, R.A. *Classification and Regression Trees.* 1 ed. Boca Raton: Chapman & Hall/CRC, 1984.
88. Breiman, L. Random Forests. *Machine Learning.* 2001, **45**(1), pp.5-32.
89. White, H. *Artificial Neural Networks: Approximation and Learning Theory.* Oxford: Blackwell Publishers, Inc., 1992.
90. Krieger, N. and Davey Smith, G. Reply to Pearl: Algorithm of the truth vs real-world science (letter). *International Journal of Epidemiology.* 2018, **47**(3), pp.1004-1006.
91. Vandenbroucke, J.P., Broadbent, A. and Pearce, N. Causality and causal inference in epidemiology: the need for a pluralistic approach. *International Journal of Epidemiology.* 2016, **45**(6), pp.1776–1786.
92. Van Breukelen, G.J.P. ANCOVA versus change from baseline had more power in randomized studies and more bias in nonrandomized studies. *Journal of Clinical Epidemiology.* 2006, **59**(9), pp.920-925.
93. Keijzer-Veen, M.G., Euser, A.M., Van Montfoort, N., Dekker, F.W., Vandenbroucke, J.P. and Van Houwelingen, H.C. A regression model with unexplained residuals was preferred in the analysis of the fetal origins of adult diseases hypothesis. *Journal of Clinical Epidemiology.* 2005, **58**(12), pp.1320-1324.
94. Tu, Y.K. and Gilthorpe, M.S. Unexplained residuals models are not solutions to statistical modeling of the fetal origins hypothesis. *Journal of Clinical Epidemiology.* 2007, **60**(3), pp.318-319.
95. Lovelace, R. and Dumont, M. *Spatial Microsimulation with R.* Boca Raton: Taylor & Francis Group, LLC, 2016.
96. Auchincloss, A.H. and Diez Roux, A.V. A new tool for epidemiology: the usefulness of dynamic-agent models in understanding place effects on health. *American Journal of Epidemiology.* 2008, **168**(1), pp.1-8.
97. Birkin, M. and Wu, B. A Review of Microsimulation and Hybrid Agent-based Approaches. In: Heppenstall, A. et al. eds. *Agent-Based Models of Geographical Systems*. Springer, 2012, pp.51-68.
98. Marshall, B.D. and Galea, S. Formalizing the role of agent-based modeling in causal inference and epidemiology. *American Journal of Epidemiology.* 2014, **181**(2), pp.1-8.
99. Murray, E.J., Robins, J.M., Seage III, G.R., Freedberg, K.A. and Hernán, M.A. A comparison of agent-based models and the parametric g-formula for causal inference. *American Journal of Epidemiology.* 2017, **186**(2), pp. 131-142.
100. Hammond, R.A. Complex systems modeling for obesity research. *Preventing Chronic Disease.* 2009, **6**(3), p.A97.

101. Green, L.W. Public health asks of systems science: to advance our evidence-based practice, can you help us get more practice-based evidence? *American Journal of Public Health.* 2006, **96**(3), pp.406-409.
102. Ness, R.B., Koopman, J.S. and Roberts, M.S. Causal system modeling in chronic disease epidemiology: a proposal. *Annals of Epidemiology.* 2007, **17**(7), pp.564-568.
103. Luke, D.A. and Stamatakis, K.A. Systems science methods in public health: dynamics, networks, and agents. *Annual Review of Public Health.* 2012, **33**, pp.357-376.
104. Fink, D.S., Keyes, K.M. and Cerdá, M. Social Determinants of Population Health: A Systems Sciences Approach. *Current Epidemiology Reports.* 2016, **3**, pp.98-105.
105. von Neumann, J. The general and locial theory of automata. In: Jeffress, L.A. ed. *Cerebral Mechanisms in Behavior: The Hixon Symposiom*. Oxford: Wiley, 1951, pp.1-41.
106. Orcutt, G.H. A new type of socio-economic system. *The Review of Economics and Statistics.* 1957, **39**(2), pp.116-123.
107. Schelling, T.C. Dynamic models of segregation. *Journal of Mathematical Sociology.* 1971, **1**, pp.143-186.
108. Butland, B., Jebb, S., Kopelman, P., McPherson, K., Thomas, S., Mardell, J. and Parry, V. *Foresight: Tackling Obesities: Future Choices - Project Report.* London: Government Office for Science, 2007.
109. Heppenstall, A.J., Evans, A.J. and Birkin, M.H. Genetic algorithm optimisation of an agent-based model for simulating a retail market. *Environment and Planning B: Urban Analytics and City Science.* 2007, **34**(6), pp.1051-1070.
110. Manley, E., Cheng, T., Penn, A. and Emmonds, A. A framework for simulating large-scale complex urban traffic dynamics through hybrid agent-based modelling. *Computers, Environment and Urban Systems.* 2014, **44**, pp.27-36.
111. Crooks, A., Croitoru, A., Lu, X., Wise, S., Irvine, J.M. and Stefanidis, A. Walk This Way: Improving Pedestrian Agent-Based Models through Scene Activity Analysis. *International Journal of Geo-Information.* 2015, **4**(3), pp.1627-1656.
112. Hernán, M.A. Invited Commentary: Agent-Based Models for Causal Inference - Reweighting Data and Theory in Epidemiology. *American Journal of Epidemiology.* 2014, **181**(2), pp.103-105.
113. Angrist, J.D. and Krueger, A.B. Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments. *Journal of Economic Perspectives.* 2001, **15**(4), pp.69-85.
114. Burgess, S., Timpson, N.J., Ebrahim, S. and Davey Smith, G. Mendelian randomisation: where are we now and where are we going? *International Journal of Epidemiology.* 2015, **44**(2), pp.379-388.
115. Diez Roux, A.V. Integrating social and biologic factors in health research: a systems view. *Annals of Epidemiology.* 2007, **17**(7), pp.569-574.
116. Banack, H.R. and Kaufman, J.S. Estimating the Time-Varying Joint Effects of Obesity and Smoking on All-Cause Mortality Using Marginal Structural Models. *American Journal of Epidemiology.* 2016, **183**(2), pp.122-129.
117. Bedard, A., Serra, I., Dumas, O., Basagana, X., Clavel-Chapelon, F., Le Moual, N., Sanchez, M., Siroux, V., Varraso, R. and Garcia-Aymerich, J. Time-Dependent Associations between Body Composition, Physical Activity, and Current Asthma in Women: A Marginal Structural Modeling Analysis. *American Journal of Epidemiology.* 2017, **186**(1), pp.21-28.
118. Byberg, K.K., Eide, G.E., Forman, M.R., Juliusson, P.B. and Oymar, K. Body mass index and physical activity in early childhood are associated with atopic sensitization, atopic dermatitis and asthma in later childhood. *Clinical and Translational Allergy.* 2016, **6 (1) (no pagination)**(33).

119. Danaei, G., Robins, J.M., Young, J.G., Hu, F.B., Manson, J.E. and Hernán, M.A. Weight Loss and Coronary Heart Disease: Sensitivity Analysis for Unmeasured Confounding by Undiagnosed Disease. *Epidemiology.* 2016, **27**(2), pp.302-310.
120. Karlsen, M., Grandjean, P., Weihe, P., Steuerwald, U., Oulhote, Y. and Valvi, D. Early-life exposures to persistent organic pollutants in relation to overweight in preschool children. *Reproductive Toxicology.* 2017, **68**, pp.145-153.
121. Medenwald, D., Loppnow, H., Kluttig, A., Nuding, S., Greiser, K.H., Thiery, J., Tiller, D., Herzog, B., Werdan, K. and Haerting, J. Educational level and chronic inflammation in the elderly - the role of obesity: Results from the population-based CARLA study. *Clinical Obesity.* 2015, **5**(5), pp.256-265.
122. Murphy, C.C., Martin, C.F. and Sandler, R.S. Racial differences in obesity measures and risk of colorectal adenomas in a large screening population. *Nutrition and Cancer.* 2015, **67**(1), pp.98-104.
123. Ahern, A.L., Wheeler, G.M., Aveyard, P., Boyland, E.J., Halford, J.C.G., Mander, A.P., Woolston, J., Thomson, A.M., Tsiountsioura, M., Cole, D., Mead, B.R., Irvine, L., Turner, D., Suhrcke, M., Pimpin, L., Retat, L., Jaccard, A., Webber, L., Cohn, S.R. and Jebb, S.A. Extended and standard duration weight-loss programme referrals for adults in primary care (WRAP): a randomised controlled trial. *The Lancet.* 2017, **389**(10085), pp.2214-2225.
124. Barrientos-Gutierrez, T., Zepeda-Tello, R., Rodrigues, E.R., Colchero-Aragones, A., Rojas-Martonez, R., Lazcano-Ponce, E., Hernandez-Avila, M., Rivera-Dommarco, J. and Meza, R. Expected population weight and diabetes impact of the 1-peso-per-litre tax to sugar sweetened beverages in Mexico. *PLoS ONE.* 2017, **12**(5).
125. Basu, S., Seligman, H. and Winkleby, M. A metabolic-epidemiological microsimulation model to estimate the changes in energy intake and physical activity necessary to meet the Healthy People 2020 obesity objective. *American Journal of Public Health.* 2014, **104**(7), pp.1209-1216.
126. Castilla, I., Mar, J., Valcarcel-Nazco, C., Arrospide, A. and Ramos-Goni, J.M. Cost-Utility Analysis of Gastric Bypass for Severely Obese Patients in Spain. *Obesity Surgery.* 2014, **24**(12), pp.2061-2068.
127. Hoerger, T.J., Zhang, P., Segel, J.E., Kahn, H.S., Barker, L.E. and Couper, S. Cost-effectiveness of bariatric surgery for severely obese adults with diabetes. *Diabetes Care.* 2010, **33**(9), pp.1933-1939.
128. Kristensen, A.H., Flottemesch, T.J., Maciosek, M.V., Jenson, J., Barclay, G., Ashe, M., Sanchez, E.J., Story, M., Teutsch, S.M. and Brownson, R.C. Reducing childhood obesity through U.S. Federal policy: A microsimulation analysis. *American Journal of Preventive Medicine.* 2014, **47**(5), pp.604-612.
129. Wentworth, J.M., Dalziel, K.M., O'Brien, P.E., Burton, P., Shaba, F., Clarke, P.M., Laiteerapong, N. and Brown, W.A. Cost-effectiveness of gastric band surgery for overweight but not obese adults with type 2 diabetes in the U.S. *Journal of Diabetes and its Complications.* 2017, **31**(7), pp.1139-1144.
130. Auchincloss, A.H., Riolo, R.L., Brown, D.G., Cook, J. and Diez Roux, A.V. An agent-based model of income inequalities in diet in the context of residential segregation. *American Journal of Preventive Medicine.* 2011, **40**(3), pp.303-311.
131. El-Sayed, A.M., Seemann, L., Scarborough, P. and Galea, S. Are network-based interventions a useful antiobesity strategy? An application of simulation models for causal inference in epidemiology. *American Journal of Epidemiology.* 2013, **178**(2), pp.287-295.
132. Li, Y., Zhang, D. and Pagan, J.A. Social Norms and the Consumption of Fruits and Vegetables across New York City Neighborhoods. *Journal of Urban Health.* 2016, **93**(2), pp.244-255.

133. Orr, M.G., Galea, S., Riddle, M. and Kaplan, G.A. Reducing racial disparities in obesity: simulating the effects of improved education and social network influence on diet behavior. *Annals of Epidemiology.* 2014, **24**(8), pp.563-569.
134. Wang, Y., Xue, H., Chen, H.J. and Igusa, T. Examining social norm impacts on obesity and eating behaviors among US school children based on agent-based model. *BMC Public Health.* 2014, **14**, p.923.
135. Zhang, D., Giabbanelli, P.J., Arah, O.A. and Zimmerman, F.J. Impact of different policies on unhealthy dietary behaviors in an urban adult population: an agent-based simulation model. *American Journal of Public Health.* 2014, **104**(7), pp.1217-1222.
136. Siebert, U., Alagoz, O., Bayoumi, A.M., Beate, J., Owens, D.K., Cohen, D.J. and Kuntz, K.M. State-Transition Modeling: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force-3. *Value in HEalth.* 2012, **15**, pp.812-820.
137. Baroni, E. and Richiardi, M. *Orcutt's Vision, 50 Years On.* Torino: Laboratorio Riccardo Revelli, 2007.
138. Batty, M. *Cities and complexity: Understanding cities with cellular automata, agent-based models, and fractals.* Cambridge: MIT Press, 2005.
139. Oakes, J.M. Invited Commentary: Rescuing Robinson Crusoe. *American Journal of Epidemiology.* 2008, **168**(1), pp.9-12.
140. Marshall, B.D.L., Paczkowski, M.M., Seemann, L., Tempalski, B., Pouget, E.R., Galea, S. and Friedman, S.R. A Complex Systems Approach to Evaluate HIV Prevention in Metropolitan Areas: Preliminary Implications for Combination Intervention Strategies. *PLoS ONE.* 2012, **7 (9) (no pagination)**(e44833).
141. Crooks, A.T. and Hailegiorgis, A.B. An agent-based modeling approach applied to the spread of cholera. *Environmental Modelling & Software.* 2014, **62**, pp.164-177.
142. Kumar, S., Piper, K., Galloway, D.D., Hadler, J.L. and Grefenstette, J.J. Is population structure sufficient to generate area-level inequalities in influenza rates? An examination using agent-based models. *BMC Public Health.* 2015, **15**, p.947.
143. Neubacher, D., Furian, N. and Vossner, S. An agent-based approach to reveal the effects of age-related contact patterns on epidemic spread. In: *European Simulation and Modelling Conference*, *Leicester, United Kingdom*. 2015.
144. Li, Y., Lawley, M.A., Siscovick, D.S., Zhang, D. and Pagan, J.A. Agent-Based Modeling of Chronic Diseases: A Narrative Review and Future Research Directions. *Preventing Chronic Disease.* 2016, **13**, p.E69.
145. Zaidi, A. and Rake, K. *Dynamic Microsimulation Models: A Review and Some Lessons for SAGE.* London: The London School of Economics, 2001.
146. Siebert, U. The role of decision-analytic models in the prevention, diagnosis and treatment of coronary heart disease. *Zeitschrift für Kardiologie.* 2002, **91**(3), pp.144-151.
147. Lord, F.M. A paradox in the interpretation of group comparisons. *Psychological Bulletin.* 1967, **68**(5), pp.304-305.
148. Kim, Y. and Steiner, P.M. Gain Scores Revisited: A Graphical Models Perspective. *Sociological Methods & Research.* 2019, p.0049124119826155.
149. Thomas, D.R. and Zumbo, B.D. Difference Scores From the Point of View of Reliability and Repeated-Measures ANOVA: In Defense of Difference Scores for Data Analysis. *Educational and Psychological Measurement.* 2011, **72**(1), pp.37-43.
150. Cronbach, L.J. and Furby, L. How we should measure "change": Or should we? *Psychological Bulletin.* 1970, **74**(1), pp.68-80.
151. Shahar, E. and Shahar, D.J. Causal diagrams and change variables. *Journal of Evaluation in Clinical Practice.* 2012, **18**(1), pp.143-148.
152. Senn, S. Change from baseline and analysis of covariance revisited. *Statistics in Medicine.* 2006, **25**(24), pp.4334-4344.
153. Werts, C.E. and Linn, R.L. A general linear model for studying growth. *Psychological Bulletin.* 1970, **73**(1), pp.17-22.

154. Maris, E. Covariance adjustment versus gain scores—revisited. *Psychological Methods.* 1998, **3**(3), pp.309-327.
155. Wainer, H. Adjusting for differential base rates: Lord's Paradox again. *Psychological Bulletin.* 1991, **109**(1), pp.147-151.
156. Wainer, H. and Brown, L.M. Two Statistical Paradoxes in the Interpretation of Group Differences: Illustrated with Medical School Admission and Licensing Data. *The American Statistician.* 2004, **58**(2), pp.117-123.
157. Allison, P.D. Change Scores as Dependent Variables in Regression Analysis. *Sociological Methodology.* 1990, **20**, pp.93-114.
158. Glymour, M.M., Weuve, J., Berkman, L.F., Kawachi, I. and Robins, J.M. When Is Baseline Adjustment Useful in Analyses of Change? An Example with Education and Cognitive Change. *American Journal of Epidemiology.* 2005, **162**(3), pp.267-278.
159. Richiardi, L., Bellocco, R. and Zugna, D. Mediation analysis in epidemiology: methods, interpretation and bias. *International Journal of Epidemiology.* 2013, **42**(5), pp.1511-1519.
160. National Health and Nutrition Examination Survey (NHANES), 2009-2014. Hyattsville, MD: *Department of Health and Human Services, Centers for Disease Control and Prevention (CDC)*. [Online]. 2016.
161. *R: A language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing, 2013.
162. Holland, P.W. and Rubin, D.B. On Lord's Paradox. *ETS Research Report Series.* 1982, **1982**(2), pp.i-41.
163. Senn, S. Rothamsted Statistics meets Lord's Paradox. *Error Statistics Philosophy.* 2018. [Online]. Available from: https://errorstatistics.com/2018/11/11/stephen-senn-rothamsted-statistics-meets-lords-paradox-guest-post/
164. Chiolero, A., Paradis, G., Madeleine, G., Hanley, J.A., Paccaud, F. and Bovet, P. Birth weight, weight change, and blood pressure during childhood and adolescence: A school-based multiple cohort study. *Journal of Hypertension.* 2011, **29**(10), pp.1871-1879.
165. Grijalva-Eternod, C.S., Wells, J.C., Girma, T., Kaestel, P., Admassu, B., Friis, H. and Andersen, G.S. Midupper arm circumference and weight-for-length z scores have different associations with body composition: evidence from a cohort of Ethiopian infants. *American Journal of Clinical Nutrition.* 2015, **102**(3), pp.593-599.
166. Johnson, W. Analytical strategies in human growth research. *American journal of human biology : the official journal of the Human Biology Council.* 2015, **27**(1), pp.69-83.
167. Yesil, G.D., Gishti, O., Felix, J.F., Reiss, I., Ikram, M.K., Steegers, E.A.P., Hofman, A., Jaddoe, V.W.V. and Gaillard, R. Influence of maternal gestational hypertensive disorders on microvasculature in school-age children. *American Journal of Epidemiology.* 2016, **184**(9), pp.605-615.
168. Toemen, L., Gishti, O., Van Osch-Gevers, L., Steegers, E.A.P., Helbing, W.A., Felix, J.F., Reiss, I.K.M., Duijts, L., Gaillard, R. and Jaddoe, V.W.V. Maternal obesity, gestational weight gain and childhood cardiac outcomes: Role of childhood body mass index. *International Journal of Obesity.* 2016, **40**(7), pp.1070-1078.
169. Toemen, L., Gaillard, R., van Osch-Gevers, L., Helbing, W.A., Hofman, A. and Jaddoe, V.W. Tracking of structural and functional cardiac measures from infancy into school-age. *European Journal of Preventive Cardiology.* 2017, p.2047487317715512.
170. Sonnenschein-van der Voort, A.M.M., Gaillard, R., de Jongste, J.C., Hofman, A., Jaddoe, V.W.V. and Duijts, L. Foetal and infant growth patterns, airway resistance and school-age asthma. *Respirology.* 2016, **21**(4), pp.674-682.
171. Lira, P.I.C., Eickmann, S.H., Lima, M.C., Amorim, R.J., Emond, A.M. and Ashworth, A.N.N. Early head growth: relation with IQ at 8 years and determinants in term infants

of low and appropriate birthweight. *Developmental Medicine & Child Neurology.* 2010, **52**(1), pp.40-46.

172. Wills, A.K., Strand, B.H., Glavin, K., Silverwood, R.J. and Hovengen, R. Regression models for linking patterns of growth to a later outcome: infant growth and childhood overweight. *BMC Medical Research Methodology.* 2016, **16**, p.9.
173. Horta, B.L., Gigante, D.P., Osmond, C., Barros, F.C. and Victora, C.G. Intergenerational effect of weight gain in childhood on offspring birthweight. *International Journal of Epidemiology.* 2009, **38**(3), pp.724-732.
174. Richter, L.M., Victora, C.G., Hallal, P.C., Adair, L.S., Bhargava, S.K., Fall, C.H., Lee, N., Martorell, R., Norris, S.A., Sachdev, H.S. and Stein, A.D. Cohort profile: The consortium of health-orientated research in transitioning societies. *International Journal of Epidemiology.* 2012, **41**(3), pp.621-626.
175. Gandhi, M., Ashorn, P., Maleta, K., Teivaanmaki, T., Duan, X. and Cheung, Y.B. Height gain during early childhood is an important predictor of schooling and mathematics ability outcomes. *Acta Paediatrica, International Journal of Paediatrics.* 2011, **100**(8), pp.1113-1118.
176. Gonzalez, D.A., Nazmi, A. and Victora, C.G. Growth from birth to adulthood and abdominal obesity in a Brazilian birth cohort. *International Journal of Obesity.* 2010, **34**(1), pp.195-202.
177. Keijzer-Veen, M.G. Response to Tu and Gilthorpe: Preventing misinterpretation of coefficients in analysis of fetal origins of adult disease. *Journal of Clinical Epidemiology.* 2007, **60**(3), pp.319-320.
178. Hernán, M.A. and Robins, J.M. Instruments for causal inference: An epidemiologist's dream? *Epidemiology.* 2006, **17**(4), pp.360-372.
179. Angrist, J.D. and Imbens, G.W. Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity. *Journal of the American Statistical Association.* 1995, **90**(430), pp.431-442.
180. Hardy, R., Ghosh, A.K., Deanfield, J., Kuh, D. and Hughes, A.D. Birthweight, childhood growth and left ventricular structure at age 60-64 years in a British birth cohort study. *International Journal of Epidemiology.* 2016, **13**, p.13.
181. Sonnenberg, F.A. and Beck, J.R. Markov Models in Medical Decision Making: A Practical Guide. *Medical Decision Making.* 1993, **13**(4), pp.322-338.
182. Rutter, C.M., Zaslavsky, A.M. and Feuer, E.J. Dynamic Microsimulation Models for Health Outcomes: A Review. *Medical Decision Making.* 2010, **31**(1), pp.10-18.
183. Murray, E.J., Robins, J.M., Seage, G.R., Freedberg, K.A. and Hernán, M.A. The Challenges of Parameterizing Direct Effects in Individual-Level Simulation Models. *Medical Decision Making.* 2020, **40**(1), pp.106-111.
184. Belanger, A. and Sabourin, P. *Microsimulation and Population Dynamics: An Introduction to Mogden 12.* Springer International Publishing, 2017.
185. Krijkamp, E.M., Alarid-Escudero, F., Enns, E.A., Jalal, H.J., Hunink, M.G.M. and Pechlivanoglou, P. Microsimulation Modeling for Health Decision Sciences Using R: A Tutorial. *Medical Decision Making.* 2018, **38**(3), pp.400-422.
186. Stensrud, M.J., Young, J.G., Didelez, V., Robins, J.M. and Hernán, M.A. Separable Effects for Causal Inference in the Presence of Competing Events [v2]. *ArXiv e-prints.* [Online]. 2019. [Accessed 11 February, 2020]. Available from: https://ui.adsabs.harvard.edu/abs/2019arXiv190109472S
187. Young, J.G., Stensrud, M.J., Tchetgen Tchetgen, E.J. and Hernán, M.A. A causal framework for classical statistical estimands in failure time settings with competing events [v3]. *ArXiv e-prints.* [Online]. 2018. [Accessed 11 February, 2020]. Available from: https://ui.adsabs.harvard.edu/abs/2018arXiv180606136Y
188. Ryder, N.B. Notes on the Concept of a Population. *American Journal of Sociology.* 1964, **69**(5), pp.447-463.

189. van Imhoff, E. and Post, W. Microsimulation Methods for Population Projection. *Population: An English Selection.* 1998, **10**(1), pp.97-138.
190. Team, R.C. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. 2013. Available from: http://www.R-project.org/
191. Health Survey for England, 1994. 5th Edition ed. *UK Data Service*. [Online]. 2017. Available from: http://doi.org/10.5255/UKDA-SN-3640-2
192. Health Survey for England, 2004. 2nd Edition ed. *UK Data Service*. [Online]. 2010. Available from: http://doi.org/10.5255/UKDA-SN-5439-1
193. Kim, S.B. Explaining Lord's Paradox in Introductory Statistical Theory Courses. *International Journal of Statistics and Probability.* 2018, **7**(4), pp.1-10.
194. Lesko, C.R., Buchanan, A.L., Westreich, D., Edwards, J.K., Hudgens, M.G. and Cole, S.R. Generalizing Study Results: A Potential Outcomes Perspective. *Epidemiology.* 2017, **28**(4), pp.553-561.
195. Pearl, J. and Bareinboim, E. *Transportability across studies: A formal approach.* 2018.
196. Pearl, J. and Bareinboim, E. External Validity: From Do-Calculus to Transportability Across Populations. *Statistical Science.* 2014, **29**(4), pp.579-595.
197. Ogburn, E.L. and VanderWeele, T.J. Causal Diagrams for Interference. *Statistical Science.* 2014, **29**(4), pp.559-578.
198. Ogburn, E.L. and VanderWeele, T.J. Vaccines, contagion, and social networks. *Ann. Appl. Stat.* 2017, **11**(2), pp.919-948.
199. Ogburn, E.L. Challenges to Estimating Contagion Effects from Observational Data. In: Lehmann, S. and Ahn, Y.Y. eds. *Complex Spreading Phenomena in Social Systems. Computational Social Sciences*. Springer, 2018.
200. Ogburn, E.L., Sofrygin, O., Diaz, I. and van der Laan, M.J. Causal inference for social network data. *ArXiv e-prints.* [Online]. 2017. p.arXiv:1705.08527. [Accessed May 01, 2017]. Available from: https://ui.adsabs.harvard.edu/abs/2017arXiv170508527O
201. Venables, W.N. and Ripley, B.D. *Modern Applied Statistics with S.* 4 ed. New York: Springer, 2002.
202. Okumura, Y. *rpsychi: Statistics for psychiatric research*. R. 2012. Available from: https://CRAN.R-project.org/package=rpsychi
203. Gentle, J.E. *Matrix Algebra: Theory, Computations. and Applications in Statistics.* New York: Springer, 2007.
204. Bates, D. and Maechler, M. *Matrix: Sparse and Dense Matrix Classes and Methods*. R. 2018. Available from: https://CRAN.R-project.org/package=Matrix
205. Novomestky, F. *matrixcalc: Collection of functions for matrix calculations*. R. 2012. Available from: https://CRAN.R-project.org/package=matrixcalc
206. Wickham, H., Hester, J. and Chang, W. *devtools: Tools to Make Developing R Packages Easier*. R. 2019. Available from: https://CRAN.R-project.org/package=devtools
207. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis.* New York: Springer-Verlag, 2016.
208. Auguie, B. *gridExtra: Miscellaneous Functions for "Grid" Graphics*. R. 2017. Available from: https://CRAN.R-project.org/package=gridExtra
209. Chang, W. *extrafont: Tools for using fonts*. R. 2014. Available from: https://CRAN.R-project.org/package=extrafont
210. Harrell Jr., F.E. *Hmisc: Harrell Miscellaneous*. R. 2019. Available from: https://CRAN.R-project.org/package=Hmisc
211. Wickham, H. and Henry, L. *tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions*. R. 2019. Available from: https://CRAN.R-project.org/package=tidyr
212. Wickham, H. and Bryan, J. *readxl: Read Excel Files*. R. 2019. Available from: https://CRAN.R-project.org/package=readxl
213. Wickham, H. *stringr: Simple, Consistent Wrappers for Common String Operations*. R. 2019. Available from: https://CRAN.R-project.org/package=stringr

214. Wickham, H. The Split-Apply-Combine Strategy for Data Analysis. *Journal of Statistical Software.* 2011, **40**(1), pp.1-29.
215. Wickham, H. *scales: Scale Functions for Visualization*. R. 2018. Available from: https://CRAN.R-project.org/package=scales
216. Dalton, J.E. and Nutter, B. *HydeNet: Hybrid Bayesian Networks Using R and JAGS*. R. 2019. Available from: https://CRAN.R-project.org/package=HydeNet
217. Dowle, M. and Srinivasan, A. *data.table: Extension of `data.frame`*. R. 2019. Available from: https://CRAN.R-project.org/package=data.table