Understanding uncertainty and bias to improve causal inference in health intervention research

Timothy Royce Watkins

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy
in the School of Public Health, Faculty of Medicine and Health

The University of Sydney

July 2019

Declaration

I, Timothy Royce Watkins, certify that the intellectual content of this thesis is my own work and that all the assistance received in preparing this thesis and all sources of information have been properly acknowledged. It does not contain any material previously published or written by another person and has not been submitted for any degree or other purpose.

(Tim Watkins)

Timothy Royce Watkins

31 December 2018

Abstract

Most research on health interventions aims to find evidence to support better causal inferences about those interventions. However, for decades, a majority of this research has been criticised for inadequate control of bias and overconfident conclusions that do not reflect the uncertainty. Yet, despite the need for improvement, clear signs of progress have not appeared, suggesting the need for new ideas on ways to reduce bias and improve the quality of research.

With the aim of understanding why bias has been difficult to reduce, we first explore the concepts of causal inference, bias and uncertainty as they relate to health intervention research. We propose a useful definition of 'a causal inference' as: 'a conclusion that the evidence available supports either the existence, or the non-existence, of a causal effect'.

We used this definition in a methodological review that compared the statistical methods used in health intervention cohort studies with the strength of causal language expressed in each study's conclusions. Studies that used simple instead of multivariable methods, or did not conduct a sensitivity analysis, were more likely to contain overconfident conclusions and potentially mislead readers. The review also examined how the strength of causal language can be judged, including an attempt to create an automatic rating algorithm that we ultimately deemed cannot succeed.

This review also found that a third of the articles (94/288) used a propensity score method, highlighting the popularity of a method developed specifically for causal inference. On the other hand, 11% of the articles did not adjust for any confounders, relying on methods such as t-tests and chi-squared tests. This suggests that many researchers still lack an understanding of how likely it is that confounding affects their results.

Drawing on knowledge from statistics, philosophy, linguistics, cognitive psychology, and all areas of health research, the central importance of how people think and make decisions is examined in relation to bias in research. This reveals the many hard-wired cognitive biases that, aside from confirmation bias, are mostly unknown to statisticians and researchers in health. This is partly because they mostly occur without conscious awareness, yet everyone is susceptible. But while the existence of biases such as overconfidence bias, anchoring, and failure to account for the base rate have been raised in the health research literature, we examine biases that have not been raised in health, or we discuss them from a different perspective. This includes a tendency of people to accept the first explanation that comes to mind (called take-the-first heuristic); how we tend to believe that other people are more susceptible to cognitive biases than we are (bias blind spot); a tendency to seek arguments that defend our beliefs, rather than seeking the objective truth (myside bias); a bias for causal explanations (various names including the causality heuristic); and our desire to avoid cognitive effort (many names including the 'law of least mental effort').

This knowledge and understanding also suggest methods that might counter these biases and improve the quality of research. This includes any technique that encourages the consideration of alternative explanations of the results. We provide novel arguments for a number of methods that might help, such as the deliberate listing of alternative explanations, but also some novel ideas including a form of adversarial collaboration.

Another method that encourages the researcher to consider alternative explanations is causal diagrams. However, we introduce them in a way that differs from the more formal presentation that is currently the norm, avoiding most of the terminology to focus instead on their use as an intuitive framework, helping the researcher to understand the biases that may lead to different conclusions.

We also present a case study where we analysed the data for a pragmatic randomised controlled trial of a telemonitoring service. Considerable missing data hampered the forming of conclusions; however, this enabled an exploration of methods to better understand, reduce and communicate the uncertainty that remained after the analysis. Methods used included multiple imputation, causal diagrams, a listing of alternative explanations, and the parametric g-formula to handle bias from time-dependent confounding.

Finally, we suggest strategies, resources and tools that may overcome some of the barriers to better control of bias and improvements in causal inference, based on the knowledge and ideas presented in this thesis. This includes a proposed online searchable causal diagram database, to make causal diagrams themselves easier to learn and use.

Acknowledgements

I would firstly like to express to my family my thanks and deepest gratitude for their support over the past 4 years. My wife Michelle and daughter Elise have shown tremendous patience with the time that my PhD studies have taken up. And my parents were there when I needed help in the final year.

My sincere thanks also go to my supervisors Laurent Billot and Anushka Patel for their support and advice. And I want to thank HCF for their scholarship that has allowed me to undertake this doctorate.

Contents

Chapter 1
Causal Inference

1.1 Introduction

A fundamental question of interest to most people is how can we speed up the rate of progress in health intervention research? The appeal is simple: better interventions for health problems sooner. More health researchers might be one answer, and that is perhaps the easiest action to take, but better health research may make a greater difference by providing more reliable and useful answers to our questions.[1,2] And this is where causal inference enters the scene, because progress in research depends heavily on researchers inferring cause and effect associations that are accurate. This applies not only to finding the true causes of disease, but also to finding out if proposed health interventions can cause some people to become healthier. If a researcher infers wrongly that an intervention caused good health outcomes, then progress can be delayed while people use an ineffective intervention hoping for a benefit, or it is further evaluated far more than necessary, wasting valuable resources.[3,4]

The World Health Organisation describes a health intervention as "an act performed for, with or on behalf of a person or population whose purpose is to assess, improve, maintain, promote or modify health, functioning or health conditions".[5] Research that evaluates one or more health interventions often aims to answer one or both of the causal questions: does it work, and is it safe. These can be stated more explicitly as 'does the intervention cause an improvement in the health outcome of interest' and 'does the intervention cause health problems'.

To better understand how health researchers might improve the causal inferences they make, we will first examine the different ways in which a "cause" and a "causal inference" have been defined, because different definitions and conceptual models can reflect the different perspectives someone may take when considering a causal association. And different perspectives can lead to different interpretations of research data. One famous example involved the well-known statistician from last century, Ronald Fisher. Despite the considerable body of evidence on smoking and lung cancer that had accumulated by the late 1950's, Fisher was among a small number of prominent scientists who still believed that only an association had been demonstrated, rather than sufficient evidence for causation.[6] In fact, some statisticians have even claimed that they don't make causal claims, they only estimate associations.[7] Causal inference itself is a cognitive process, influenced by many factors, so understanding how different people think about and resolve causal questions may be an important consideration in our pursuit of better causal inference.

1.2 Causes

It could be argued that one of the first things we learn about as an infant is the existence of cause and effect. At some point a baby might realise, in a very basic sense, that crying will often cause them to be picked up and held.[8] Indeed, there is a large body of evidence from psychology that suggests we have an innate tendency to view the world in terms of cause and effect,[9] and thus are pre-programmed to detect causal associations.

But while we intuitively know what is meant by the phrase 'X caused Y', health-related research and indeed, science in general, is usually more concerned with finding out if 'X causes Y' or 'X is a cause of Y', and these statements are less straightforward because they are aimed at predicting future events. Yet these concepts are also fundamental to our day-to-day experience, as how else could we navigate our physical and social world without some ability to predict the effects of actions.

As something fundamental yet hard to precisely define, causality has been the subject of philosophical debate for thousands of years. Aristotle identified four types of causes of an object: that which it is made out of (e.g. bronze), that into which it is made (e.g. a statue), the reason that it was made (e.g. to commemorate a war), and that which made it (e.g. the sculptor).[10] As such, Aristotle used the word "cause" in the wider sense of an "explanation" or an “answer to a why question”.

Aristotle’s way of thinking more or less held sway until the age of enlightenment in the 17th and 18th centuries.[11] The Scottish philosopher David Hume challenged the idea that causal associations could be known with certainty, since all we ever have are a sequence of perceptions of one thing following another, that we cannot guarantee will be repeated in the future.[12] This somewhat idealist philosophy[13] could be seen in the views of statisticians like Karl Pearson who believed we could never know more about two variables than that they were correlated.[14]

A more pragmatic approach was promoted by the 17th century philosopher John Locke who defined a cause as that which makes something begin to be; and an effect as that which had its beginning from some other thing.[15] With a belief in the value of experimentation,[16] the 19th century philosopher John Stuart Mill devised criteria for identifying a causal relationship. Namely that cause should precede and covary with effect, and that alternative explanations for the relationship between cause and effect are implausible.[15]

There are plenty of other definitions of a cause or causal effect to be found in both the academic philosophy and epidemiology literature, along with those from the other sciences. Definitions found in epidemiology tend to be associated with a detailed framework for causal inference, and these will be discussed in section 1.5. First however, we will examine the concept of uncertainty and the important role it plays in causal inference.

1.3 Uncertainty and causal inference

Following Hume, it became clear that causality can never be established beyond all doubt because, however bizarre, a plausible alternative explanation for observed associations will always be conceivable. Hence, there will always be some uncertainty.

The word *uncertainty* refers simply to “a state of being uncertain”,[17] or a state where something is ‘not able to be relied on; not known or definite’. And when we are uncertain about something, such as whether a drug causes some people to improve in their state of health, this implies an underlying truth exists,[18] we just do not know with certainty what that truth is.

This uncertainty is why we use the term *inference* when talking about causes, because we cannot simply see that one thing causes another, even though for practical purposes we often think that way.[19] Instead, we need to use reason.

The Oxford English Dictionary defines inference as "a conclusion reached on the basis of evidence and reasoning".[17] Inference has been divided into three types: deduction, induction and abduction. As with most words, definitions tend to differ depending on the user and the context,[20] nevertheless, deduction generally refers to inference that logically derives a conclusion from information known or assumed to be true, and thus the conclusion is known or assumed to be true; induction is the process of drawing a more general conclusion from specific information so that, in contrast to deduction, the specific information does not guarantee the truth of the conclusion,[21] in other words, the process of generalising; and abduction, which starts with an observation or set of observations and seeks to find the simplest and most likely explanation.[22] The generation of hypotheses might be called abduction, though it is often called induction.[20]

The principle of falsification was introduced by Karl Popper in 1935 in which he rejected induction as a valid method of inference[23] and proposed that a theory or hypothesis should not be considered scientific unless it can be falsified.[24] He also believed that science progresses only by falsifying hypotheses.[25] While popular with many scientists, it may be an approach better suited to physics than to epidemiology, because the failure to observe a relationship in a health-related research study will always have room for alternative explanations, just as the observation of an effect will have more than one explanation.[26]

The label applied to someone's approach to causal inference is not very important, but the approach will have an influence on the conclusions they reach. For example, those favouring the deductive approach might follow Popper's philosophy and design studies that favour the refutation of hypotheses through deductive means, instead of looking for evidence that supports hypotheses – an inductive approach. It has been suggested that randomised controlled trials follow the deductive approach with blinding, randomisation and controls aimed at refuting hypotheses.[23] While no-one would dispute the value of blinding, randomisation and control in the evaluation of possible causes, nor the value of falsification itself, there are those in epidemiology who do not agree that the primary focus should be on deductive refutation.[20,26,27] The debate continues[28–30] and probably will for some time yet, but it is easy to view causal inference itself as always inductive,[31] at least in the sense that there is always some uncertainty present. And ever since Hume exposed the fallibility of induction, the question of how to progress in science when nothing may be proved has been the subject of debate.[12]

This uncertainty is why *statistical inference* came into being, though as with all terminology, it is used a little differently by different groups of people. At the heart of all definitions, however, is the use of probability theory and other mathematics to derive insights from data, often about a population from which a sample has been observed, and usually through the use of statistical models.[32] As such, while deductive inference plays a role when developing a model,[33] statistical inference can largely be described as an inductive process.[20]

When making a causal inference, the underlying truth might resemble either ‘A causes B’ or ‘A does not cause B’. However, in general, labelling something as a cause of a particular effect does not mean that if the cause is present then the effect will always occur; though extreme examples will exist, such as the incineration of a person will always cause death. And nor does it mean that if the effect is observed, a specific cause will have preceded it — a logical, or deductive fallacy called the Fallacy of Affirming the Consequent — thought by some to be especially common in epidemiology, though also common throughout science. For example, a researcher’s hypothesis H implies a prediction B, he or she observes that B is indeed what has been observed, and concludes that H must therefore be correct.[20] Hence, with reference to health outcomes we should say instead that, in general, when contemplating a question of causality, the underlying truth will either be ‘A causes B for at least some people' or ‘A does not cause B for anybody'.

An *accurate* research conclusion or causal inference, even when highly cautious such as ‘drug A was associated with a higher outcome than drug B’, is one that happens to agree with the underlying truth. But while we can never really know if a conclusion is accurate, the results of future research should give an indication over time.

1.3.1 A definition of a causal inference

Lastly, with uncertainty ever present to some degree, a causal inference can be defined simply as a statement about a causal effect.[34] However, a more precise definition is needed and we propose the following:

> *A causal inference* is conclusion that the evidence available supports either the existence, or the non-existence, of a causal effect.

This definition acknowledges the reality that a decision such as ‘A is not a cause of B’ can have just as much of an impact on people’s behaviour regarding A (e.g. a health intervention) as a decision that ‘A is a cause of B’.

Often accompanying a causal inference, especially when it relates to research, is some sense of the uncertainty associated with that inference. This uncertainty might even be considered as part of a causal inference, though we have opted to treat them as separate concepts because a causal inference is often made with no conscious sense of uncertainty.

1.4 Early frameworks for causal inference in epidemiology

1.4.1 Bradford Hill criteria

We now turn to modern frameworks for causal inference, within which, additional definitions of a cause can be found. One of the best known remains the “Bradford Hill criteria”, so named even though Sir Austin Bradford Hill only called them “viewpoints” when they were published in 1965.[35] The nine items can be briefly summarised as:

1. **Strength:** strength of the association
2. **Consistency:** association is consistently observed (reproducibility)
3. **Specificity:** the effect is associated specifically with this particular cause
4. **Temporality:** the cause precedes the effect
5. **Biological gradient:** stronger effects with increasing dose or exposure
6. **Plausibility:** in terms of current scientific knowledge and theory
7. **Coherence:** with related facts and evidence
8. **Experiment:** study where the exposure is manipulated e.g. RCT
9. **Analogy:** with similar effects and exposures

Hill did not believe these should be considered criteria for causal inference, however, but instead "a useful tool to help us make up our minds on the fundamental question - is there any other way of explaining the set of facts before us, that is equally, or more likely, than the cause and effect association we suspect?".[35] He did not believe that all criteria should be met for valid causal inference and, in truth, the only criteria necessary is that the cause precedes the effect in time.[14] Nevertheless, while time and experience has led to criticism that many of the items do not work in practice,[36] there remains strong support for their use within epidemiology.[37,38]

1.4.2 Sufficient-component cause model

Just over a decade later, Kenneth Rothman defined a cause as "an act or event or a state of nature which initiates or permits, alone or in conjunction with other causes, a sequence of events resulting in an effect".[39] His definition accompanied a conceptual framework for causal inference that he independently introduced into epidemiology, following Mackie in philosophy in 1965[40] and Cayley in 1853.[41,42] It came to be called the 'sufficient-component cause model',[43] 'sufficient-cause framework',[42] 'sufficient-cause model'[44] or 'component-cause model'[45]. It derives from the fact that every event occurring in nature will be caused by many prior events that combine to produce the eventual outcome. And each of these prior events will consist themselves of prior events that caused them to occur. He called the final sequence of *component* causes a *sufficient cause*, with the event not occurring if any of the component causes did not occur beforehand. Hence, a sufficient cause is the complete causal mechanism that produced an event,[8] and when each component cause occurs leading up to the event, the probability of the event occurring increases.[39]

One of the main advantages of this framework, is that it emphasises that few things that we label causes will always be followed by an effect, because if one or more component causes are not present, the effect will not occur. So, when we denote something as a cause, we rarely, if ever mean, that the cause will always be followed by the effect we associate with it. Just as smoking is not always followed by lung cancer and lung cancer sometimes develops in people with no history of smoking.

But while the sufficient-component cause model is conceptually very useful,[46] it could be said that it focuses on the *causes of effects*, which are potentially limitless in number, whereas the potential outcomes framework, which we cover next, focuses on the *effects of causes*,[42] and thus is better suited to the analysis of a single cause, such as an intervention.

1.5 Potential outcomes or counterfactual framework

1.5.1 Development and definition

In the health sciences today, the potential outcomes framework, also called the counterfactual framework, is the most commonly used formalised framework for analysing causal effects.[47] It is based on common ideas about counterfactuals that can be found at least as far back as David Hume.[48] A counterfactual refers to what would have been the case if something in the past had been different, for example, a person with a headache took aspirin and the headache went away (the fact), instead of what would have happened if they had not taken aspirin (the counterfactual).[49] Thus, when the potential cause A is dichotomous, such as taking or not taking aspirin, a counterfactual definition of a causal effect can be stated: if we compare the outcome when A is present to the outcome when A is absent, all else being equal, and the outcomes differ, then A has had a causal effect on the outcome.[50]

One example of this definition is sometimes called the ‘ideal experiment’.[51] We first take an individual and give them the active treatment. We wait and measure the health outcome of interest. Next, we jump in a time machine and go back in time and, without changing anything else, switch the treatment the individual received to the control. We then wait and measure the outcome as before. If the outcomes are different, we know that the only possible cause of the difference was the active treatment.

If, on the other hand, the potential cause is not dichotomous, such as when different amounts of aspirin are being considered, then the counterfactual definition of a causal effect becomes more complicated. For example, simply comparing the outcome when the dose of aspirin is set to level $A = a$ with the outcome when the level $A \neq a$ would lump together all possible outcomes for levels of aspirin both above and below a. When A is essentially continuous, there may also be the problem of deciding how close the level would need to be to a to be considered equal to a.

The potential outcomes framework was first formalised mathematically by Jerzy Neyman in 1923, although this was not widely known until 1990.[52–54] His treatment was limited only to concepts involving randomisation, however, though it was a couple of years before R. A. Fisher proposed randomised experiments.[55] Donald Rubin then extended the model to observational studies in the 1970’s.[56–58]

1.5.2 Mathematical notation

A key aspect of the potential outcomes framework is its mathematical formulation. Using the notation of Hernan and Robins,[59] then for a binary treatment (e.g. drug or placebo) and a binary outcome (e.g. death or survival):

- if Y is a random variable representing the outcome of an individual, and
- A is a random variable representing the treatment an individual received, then
- let $Y^{a=1}$ be the potential outcome variable that would have been observed following the treatment $a = 1$, and
- $Y^{a=0}$ be the potential outcome variable that would have been observed under the treatment $a = 0$

Then, a formal definition of a *causal effect for an individual* can be stated mathematically as:

- the treatment A has a causal effect on an individual's outcome Y if $Y^{a=1} \neq Y^{a=0}$ for an individual

As already mentioned, observing both outcomes for the same individual is not possible (as far as we know), however, the average causal effect in a population of individuals can be estimated if we combine their observed outcomes.

In this case, a formal definition of the *average causal effect in a population* can be stated:[59]

- an average causal effect of treatment A on outcome Y is present if $\Pr[Y^{a=1} = 1] \neq \Pr[Y^{a=0} = 1]$ in the population of interest

This can be contrasted with a definition of association:

- treatment A and outcome Y are associated if $\Pr[Y = 1|A = 1] \neq \Pr[Y = 1|A = 0]$

More generally, for outcomes that are nonbinary as well as those that are binary, the average causal effect can be stated in terms of expected values:[59]

- an average causal effect of treatment A on outcome Y is present if $\mathrm{E}[Y^{a=1}] \neq \mathrm{E}[Y^{a=0}]$ in the population of interest

1.5.3 'Potential outcomes' or 'counterfactual outcomes'

The mathematical framework was first labelled the "randomization model"[60,61] in 1973 and then the "Rubin causal model"[16] in 1986, though not initially by Rubin. In 1990, Rubin referred to the framework as the "potential outcomes with assignment mechanism perspective" after he became aware that Neyman had given the first formal treatment in 1923 where the term 'potential yield' was used.[54] Since then, it has become widely known by various names such as the 'potential outcomes framework', 'potential outcomes model', 'counterfactual framework', or 'counterfactual model', as well as the 'Rubin causal model'.

The terms 'potential outcomes' and 'counterfactual outcomes' are often used interchangeably by authors,[10] suggesting they are viewed as having an equivalent meaning.

Rubin, however, prefers ‘potential outcomes’ because he considers that neither of the potential outcomes are counterfactual until after treatments are assigned, and then only one of the outcomes will be counterfactual.[62] On the other hand, some authors prefer to use ‘counterfactual outcomes’ for the random variables $Y^{a=1}$ and $Y^{a=0}$ because they are viewed, with reference to the naturally occurring outcome Y, as only occurring if an intervention is set to $a = 1$ or $a = 0$. Hence, they are both considered counterfactual.[10]

1.5.4 Assumptions for valid causal inference

Before discussing some criticisms recently directed at the most common way of using the potential outcomes model, we need to briefly introduce the four main assumptions for causal inference that this framework rests upon: consistency, exchangeability conditional on the measured covariates, positivity, and faithfulness.[59]

Exchangeability conditional on the measured covariates means that there is no confounding or selection bias (informative censoring) using the structural definitions of these terms developed in recent decades.[63] Exchangeability will be discussed in more detail in Chapter 2.

Positivity refers to the assumption that there were participants in both the intervention and the control groups with each possible combination of values for the observed confounders.[64] In other words, participants with each unique combination of individual characteristics had a positive probability of receiving either the intervention or the control. This is important because if positivity does not hold, then for some confounder values, no treated and untreated participants can be compared.[63]

Faithfulness is the assumption that no perfect cancellation of effects has occurred in the study, such as might be seen with a high risk medical intervention that saves some patients’ lives buts kills others, leading to the appearance that the intervention has no effect on the outcome.[65]

The *consistency assumption* is often stated: an individual's potential outcome under their observed exposure history is precisely their observed outcome.[66] In the mathematical notation used above,[59]

- if a subject's observed treatment is $A = a$ then his observed outcome Y should be equal to his potential outcome Y^a
- the consistency assumption can also be expressed as $Y = Y^A$

To understand the importance of this assumption, recall that within the potential outcomes framework, a binary treatment A has had a causal effect on outcome Y if, had A been absent with everything else the same, then Y would have been different. For this hypothetical situation, however, we must assume that treatment A was assigned to the individual, as in experimental trials, and not a result of choice, because if everything else had been the same except for A, no reasons would have existed that might have led to a different choice.[67]

Thus, to satisfy the consistency assumption, the outcome observed for each person needs to be the same as the outcome that would have been observed had the intervention been assigned to them, instead of being chosen. In other words, we need to be able to explain how a particular value of the treatment or exposure (e.g. the control) could be hypothetically assigned to a participant exposed to another value (e.g. the treatment).[66] For this to be possible, the intervention being investigated needs to be well defined, because otherwise the causal contrast $Y^{a=1} - Y^{a=0}$ would not itself be well defined.

This requirement for well-defined interventions is easily satisfied in most trials where interventions are assigned to participants, such as in randomised controlled trials. But in observational studies, and randomised trials where the interventions may exhibit some variation, this assumption may not be satisfied, and as a result, the assumptions of exchangeability and positivity also become less plausible.[68]

A brief note about terminology is appropriate here. Robins introduced the *consistency assumption* in the 1990's in relation to his development of structural nested failure time[69] and structural nested mean[70] models, both of which adjust for time-dependent covariates. This assumption has since been widely adopted for models not involving time-dependent variables.

However, a similar assumption is incorporated within Rubin's stable unit treatment value assumption or SUTVA, that he introduced in 1980,[71] which states that "the potential outcomes for any unit do not vary with the treatment assigned to other units, and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes". This assumption clearly incorporates the consistency assumption, as well as the condition: no multiple versions of the same treatment. An additional assumption this includes is no interference between study subjects, where one subject's treatment sometimes affects another subject's outcome.[63] This condition is clearly violated in many different types of studies, with one of the most obvious examples that of vaccination, where an unvaccinated subject might infect another subject, but not if they had been vaccinated. Such examples have encouraged research in recent decades investigating and developing methods for causal inference in the presence of interference.[72]

Because of the overlap of these various assumptions, either the consistency assumption or SUTVA tends to be referred to in practice, but not both. In this thesis, we will use the consistency assumption, as it appears to be more commonly used, and because the no interference assumption may sometimes be relaxed.[72]

1.5.5 Criticisms

Asserting that only well-defined causal effects are worth investigating would seem to be at odds with the more traditional way that epidemiologists look for aetiologic factors.[73] For example, should non-manipulable factors such as sex and race not be regarded as causes?[73] Indeed, there has been growing debate for a number of decades. In recent years, a number of articles on this topic have appeared from prominent epidemiologists,[31,73–81] criticising what some called the "restricted potential outcomes approach (RPOA)",[73] along with replies from those maintaining the need for well-defined causes.[47,82–86] It should be noted, however, that there is also much agreement between the groups, with Robins and Weissman making the following point about an issue that often plagues debates: "the exchanges for and against the counterfactual approach to causation to this point appear to exhibit much mutual misunderstanding about what different players advocate, leading to many 'straw-man' complaints".[87]

The proposed alternatives to a more restrictive interpretation of causes have been called 'inference to the best explanation',[73,88] and 'triangulation'.[73,89]

1.6 Inference to the best explanation and triangulation

Inference to the best explanation (IBE) is the name of a framework developed by the philosopher of science Peter Lipton[88] and centres on ruling out competing hypotheses that could explain the evidence by utilising, in turn, a two-stage mechanism that involves generating hypotheses and then selecting from among them. As Krieger and Davey Smith put it: "IBE is thus driven by theory, substantive knowledge, and evidence, as opposed to being driven solely by logic or by probabilities".[77] But those on the other side of the 'causality wars'[90] seemed to be largely in agreement with IBE as an important way that scientists reason.[47,85,87]

A similar approach, *triangulation*, was also put forward and described as: "one's confidence in a finding increases if different data, investigators, theoretical approaches and methods all converge on that finding",[73] and also "the practice of strengthening causal inferences by integrating results from several different approaches, where each approach has different (and assumed to be largely unrelated) key sources of potential bias"[89]. But as with IBE, this is also seen as an important approach to science from those who otherwise disagree on methodological details.[85,90]

1.7 Other frameworks for causal inference

1.7.1 Structural equation modelling

Structural equation modelling uses two types of variables in equations that represent structural models.[91] The observed variables contained in the data are one type, while the other type of variables are called *latent variables*, which correspond to hypothetical constructs presumed to represent something not directly observable[92] but that influence the measured variables and so might be estimated. For example, the latent constructs might be factors that affect the chance of infection with an influenza virus, such as 'perception of infection risk', 'host susceptibility to virus infection' and 'compliance with preventive behaviours'.[93]

The collection of related techniques that are grouped under the term structural equation modelling (SEM) emerged following work on exploratory factor analysis and path analysis in the early part of the 20th century.[92] The techniques are widely used in the social and behavioural sciences, but are relatively uncommon in the health sciences.

Structural equation modelling appears to have been one of the more controversial approaches when used for causal inference, attracting considerable criticism from statisticians for the strong assumptions required.[94–96] Consequently, causal language has often been avoided, with models usually interpreted as either strictly confirmatory, for testing alternative models, or as tools to discover possible models by repeatedly fitting different models to the data.[92]

Nevertheless, some authors believe this has been unfortunate, with Bollen and Pearl concluding that "the current capabilities of SEMs to formalize and implement causal inference tasks are indispensable".[97]

1.7.2 Decision-Theoretic approach

Philip Dawid has probably argued the loudest against the majority view of counterfactuals, beginning his campaign in the 1970's[98] and persisting still[99]. But unlike the recent concerns detailed above about the restrictive way counterfactuals are used, Dawid argues against their use altogether.[100,101] He believes that making inferences with counterfactuals involves assumptions that can be arbitrary and untestable. However, others would argue that this simply reflects the nature of reality and, in turn, because this makes them sensitive to assumptions, it is a strength rather than a weakness, and making clearly defined assumptions allows them to be tested.[102]

His alternative framework derives from Bayesian decision analysis and thus incorporates a probabilistic view of causality.[103] But while this has many attractions, he is somewhat vague on how these probabilities are to be calculated in practice, other than with randomised trials.[104]

He also notes the potential for confusion given the variety of formal and informal frameworks for causal inference that are available. But although he wishes for "the arrival of a messianic figure who (just as Kolmogorov did for probability theory) will sweep away the confusion and produce a single theory that everyone can accept",[103] this might simply reflect his preference, as a mathematician, for the precision found in mathematical theorems. And though some, such as Pearl, have claimed in their work that "causality has been mathematized",[91] it seems to us unlikely that a problem, still unresolved after centuries of philosophical debate, can ever be settled with mathematics alone.

Nevertheless, while Dawid's particular approach to causal inference has not appealed to many, by providing a counterpoint to conventional views over recent decades he has helped spark debate around important issues in the field of causal inference, and public debate about research and ideas is partly how science progresses.[23,105]

1.7.3 Threats to validity

The final framework that we will briefly cover was developed by Donald Campbell and his colleagues from the 1950's through to the 1970's and has been the most influential approach to causal inference in field settings (non-laboratory research) in psychology and education.[49] Campbell's framework revolves around threats to validity, grouped into statistical conclusion validity, internal validity, construct validity, and external validity.[106] In fact, these terms were invented by Campbell, along with quasi-experimental research designs such as the regression discontinuity design.[49] Campbell was a psychologist, rather than a statistician, and this may be one reason that his causal framework emphasised design over analysis.[107] There can be benefits to such an approach, however, as Campbell's initial focus when designing studies, was to reduce the number of plausible alternative hypotheses that could explain the data.[49]

1.8 Causal inference from different perspectives

By understanding the concepts surrounding causes and causal inference, and the different ways that people think about and identify causes, our ability to improve our own causal inferences may be enhanced. More importantly, it may also lead to new ideas on how we might improve the causal inferences made by health researchers generally.

Chapter 2
Concepts and classifications of bias in health research

2.1 Bias and causal inference uncertainty

Improvements in healthcare, such as a new intervention, often hinge on multiple research studies returning findings that are true. Operating against better healthcare is research that clouds the truth by delivering findings that are false. Such research is sufficiently biased to produce conclusions that are wrong. But while this cannot be entirely avoided, the damage to scientific progress will be even greater if the uncertainty conveyed with the result is also untrue because most of the time this uncertainty is underestimated;[108,109] leading to conclusions that are not only false, but overconfident in the accuracy of their claim.

For research asking a causal question, such as an intervention study, the uncertainty of a result relates to doubt about its accuracy. P-values and confidence intervals provide numerical estimates of uncertainty; however, these only reflect a random component that depends on factors like sample size and sample variability. In statistics, the difference between an estimate of a parameter, such as a causal effect, and the true value is called the *error*, which is traditionally split into a *random* and a *systematic* component.[8] The systematic component is the net effect of the sources of bias that have influenced the calculation of the estimate. If the opposite of random error is *precision* and the opposite of systematic error is *validity*, then the *accuracy* of an estimate can be defined by its validity and precision.[8]

We can likewise say that an accurate causal inference is one that agrees with the underlying truth in three ways:

1. The inference of whether the causal effect exists or not is true.
    - ▹ And if we believe that most hypothesised causal effects cannot equal exactly zero,[110–115] † then we will assume a true causal effect is strong enough to be detected and is meaningful.
2. The inferred direction of the causal effect is true. That is, if an association between A and B is detected and is causal, then the inference made that either 'A causes B' or 'B causes A', is true.
3. The magnitude of the causal effect is also true. However, just as a measurement of length is always to a finite number of decimal places, this last component of an accurate causal inference can only ever be accurate to some approximation. Likewise, there is going to be a point where increased precision would have no practical value. And in health research, where estimated causal effects are usually population averages, the true magnitude will depend on the population it belongs to, suggesting that an effect size more precise than a rough measure of strength is unlikely to be useful, and potentially misleading in its accuracy.

If the third component includes the sign (positive or negative), then we know that the first component is 'it exists', and that the second is given by the sign. Nevertheless, uncertainty always implies some doubt about all three of these components, even though in some cases,

[†] For example, most health interventions, such as a drug or even a placebo, will have some effect on everyone even if that effect is extremely small. If the population is large enough, this effect will inevitably be the tiny 'straw that broke the camel's back' in some people and hence, will be a cause of the outcome in that case. In other words, for possible health causes, the underlying truth is rarely, if ever, 'A does not cause B in anyone', or 'the effect does not exist', where zero effect means 0.0000.... Likewise, when two interventions are compared, there could not possibly be zero difference between their average effects, though the difference might be very small. In both cases, the important question is whether the difference is clinically meaningful, rather than asking if an effect or difference exists. This view also poses problems for the traditional null hypothesis as it assumes a null, or zero, effect size.

doubt about the first and second will be very small (e.g. smoking increases the risk of lung cancer).

The difficulty in estimating the uncertainty lies, not surprisingly, in its uncertain nature. The random component, often expressed as confidence intervals though other forms exist such as Bayesian credible intervals, is the easiest to quantify because it can be modelled using the laws of probability. The non-random component, however, derives from sources of bias that were either not measured accurately, not measured at all, or are not even known, and so is mostly estimated through judgement based on sensitivity analyses, prior knowledge of other studies, relevant experience, and the plausibility of different types of bias. Quantitative bias techniques also exist and will be briefly discussed in Chapter 4.

To reduce the uncertainty surrounding a research conclusion, sources of bias need to be identified and their influence removed. However, history suggests that if a type of bias is unknown to the researcher, they are unlikely to detect it. While the basic concept of confounding bias can be traced at least back to the 18th century,[116] and some types of selection bias to the 19th,[117] it wasn't until the second half of the 20th century[117] that a great number of additional types[118] of bias were progressively identified. Hence, many types of bias were not revealed to exist until decades after susceptible studies were first run, so unless a researcher has either prior knowledge of the bias, or they somehow become alerted to its possibility, the existence of that bias will not only distort the results, but remain hidden, encouraging the researcher to feel overconfident that the result is accurate.

The identification and control of bias underpins the truth of causal inferences from research, so to gain an understanding of bias, we next examine what is meant by the word itself, and how that meaning has evolved over time. But in mathematical statistics, the approach to bias became more implicit than explicit: ignoring possible biases with terms like 'objective methods' and 'test assumptions'; made worse by the dominant practice of null-hypothesis significance testing. The use of causal diagrams, discussed in the last section of this chapter, is a different approach that aims is to make potential sources of bias as explicit as possible. By also making the goal of causal inference explicit—a task avoided by statisticians for many years—causal diagrams can help researchers avoid more bias and thus make better causal inferences.

2.2 Use and meaning of the word bias

2.2.1 Evolution of the word bias in English

The issue of bias is approached somewhat differently in different research settings, and this is partly linked to the evolution of the word's use in English.

The word *bias* first appeared in written English in the 16th century, derived from the French word *biais*,[119] and it may have first been applied to the way a bowl in lawn bowls moves away at an angle from the straight line it was propelled along.[120] But it was soon also used in the modern social sense of an "inclination or prejudice for or against one person or group, especially in a way considered to be unfair".[119,120]

At least as early as 1827, the word bias can be seen in relation to mortality statistics, in a warning to view other statistical estimates with caution unless they come from someone without "bias for, a particular party, or who possess so rare a degree of candour, as to enable them to state facts without partiality or concealment".[121] This use corresponds to a number of biases we are familiar with today, such as confirmation bias. Another example, also referring to a type of cognitive bias, comes from an 1885 *Science* article in which it is suggested that people's "natural bias in favour of round numbers" had resulted in census reports containing "many more persons … recorded as being just 20 or just 50 years old than were as being 19 or 49".[122]

By the turn of the century, however, *bias* was also starting to be used in the burgeoning field of statistics, such as this line from *Elements of Statistics (1901)* by Arthur Bowley: "in calculating averages give all your care to making the items free from bias".[123] The first giant of the statistics profession, Karl Pearson, similarly used it to describe a dice experiment where "the results show a bias from the theoretical results".[124]

From describing numbers, this versatile word was then utilised by another giant of statistics, Ronald Fisher, to not only describe measurement error, but also for when equations gave an "unbiased estimate" of a statistic, such as variance.[55] This additional use of the word was perhaps why the statistician John Wishart suggested, in 1939, that "… difficulties might arise because of the ambiguity of language. Consider, for example, the various meanings that might be attached to the word 'bias'".[125] Nevertheless, around the same time, the abstract term *unbiased estimator* was introduced into mathematical statistics.[126] This was, and remains, a much narrower use of the word bias, however, as it relates only to idealised settings with a definable "true" value of a parameter. The following is a standard definition:

> A point estimator $\hat{\theta}$ is said to be an *unbiased estimator* of θ if $E(\hat{\theta}) = \theta$ for every possible value of θ. … That is, $\hat{\theta}$ is unbiased if its probability (i.e., sampling) distribution is always "centred" at the true value of the parameter.[127]

Converting this to non-mathematical language: for example, assuming the parameter of interest is the mean height of all males in a specified population, then the *mean* height of a random sample of males from that population $(\hat{\theta})$ is an *unbiased estimator* of the true mean height of males in the population. In this case, *unbiased* simply indicates that the average value $(E(\hat{\theta}))$ of all the mean heights calculated from all possible samples taken from the population (the sampling distribution), is equal to the population's true mean male height (θ).

This was, and still is, an important theoretical concept in mathematical statistics. However, estimates produced from an unbiased estimator assume that no unmeasured confounding, selection bias or measurement error exists. Hence, as Greenland and Pearce note: "no available estimator can be shown to be unbiased or consistent under realistic epidemiologic conditions".[128] Nevertheless, this use of the word 'unbiased' is routinely included in introductory statistics courses, and so at times, may have encouraged the widespread overconfident belief of researchers that the estimates produced by their analysis really are unbiased.

Finally, many early advances in science were reported only in languages such as French or German, and they appear to have used words with a similar meaning to *bias*. Though not in English, communication between scientists of that era might nevertheless have led to some influence of these concepts on the modern meaning of *bias*. For example, concepts similar to confirmation bias and social-desirability bias can be found in an 1825 book on probability theory by the French mathematician Pierre-Simon de Laplace.[129] And in 1835, Pierre-Charles-

Alexandre Louis, a French physician investigating the efficacy of bleeding patients, identified confounding bias as a threat to inference.[130]

2.2.2 Modern use of the word bias

Word meaning changes over time and few words mean the same now as when they first appeared in the language.[131] We cannot know for certain how the word *bias* was first used in English, but over time it has acquired a variety of senses by which it is commonly used. Word meaning usually depends on context, however, even native speakers might interpret a word a little differently.[132] But with this in mind, from an epidemiological point of view, Porta et al. provides a good summary of the various meanings that have come to be associated with *bias* in health-related research:

> A systematic deviation of results or inferences from truth. Processes leading to such deviation. An error in the conception and design of a study - or in the collection, analysis, interpretation, reporting, publication, or review of data - leading to results or conclusions that are systematically (as opposed to randomly) different from truth.[118]

There are also words and terms with meanings that closely relate to bias. For example, an inference that is free from bias might be described as valid or accurate; while a biased inference might be called a statistical artefact or spurious. Bias is also called, and sometimes defined as *systematic error* (as opposed to *random error*). In common usage, *valid* will be in reference to a logically sound or reasonable argument, but not necessarily a true argument, if it is based on assumptions that are false. The related terms 'internal validity' and 'external validity' do refer to truth, however, and have become popular, perhaps because they succinctly capture two related concepts we can easily picture as inside and outside a study.

It is not uncommon to reserve the word *bias* for when there is a lack of internal validity, but not for when there is a lack of external validity, which is also known as *generalizability*.[133] And we will follow this convention as well, partly because internal validity biases relate directly to causal inference in health intervention studies, and hence, are the only ones we will be examining in this thesis. However, it is terminology associated with the many classifications of bias, such as *selection bias, confounding, measurement error*, and all of the individual types of bias, that seem to cause the most confusion. Yet, as psycholinguist Steven Pinker notes: “when it comes to correct English, there’s no one in charge”,[131] so familiarity with multiple meanings is, to some extent, necessary. He does go on to say, however, that tacit conventions about word meaning and use emerge over time, even though this implicit consensus “… can change over the years in a process as unplanned and uncontrollable as the vagaries of fashion”.[131]

2.3 The approach to bias in mathematical statistics

Emerging in the latter part of the 19th century, the discipline of statistics was initially led by Francis Galton (1822-1911), who developed ideas around regression and correlation,[134] followed by his protégé, Karl Pearson (1857-1936), who was the first to incorporate probability distributions into the analysis of data with chi-squared goodness-of-fit tests.[135] Pearson also had strong views about the concept of causality and how it should be treated in science; these he discussed in detail in *The Grammar of Science*[136] with lines such as “science for the past is a description, for the future a belief; it is not, and has never been, an explanation”. As a consequence, he focused on developing methods to find associations, or correlations, in data.

Pearson’s contributions added momentum to the shift in science towards the abstract, where reality is often described in terms of probability distributions, parameters and degrees of freedom; instead of the natural categories we use to understand ourselves and our world.[135] It may not be surprising then, that scientists have struggled to fully understand how to use or interpret the results of statistical methods over the last 80 or more years.

This mathematical perspective also led to a contrasting approach to bias. As mentioned in 2.2.1, mathematical statistics makes use of the word *bias* only to refer to an *unbiased estimator* or the *unbiased estimate* that the estimator produces. The danger of this approach is that statisticians working in mathematical statistics might lack an adequate understanding of the problems that scientists are trying to solve with statistics.[137] Practising statisticians who work with subject matter experts need to assume the scientific meaning of the word bias when constructing statistical models. However, the definition of the word bias in one currently available statistics dictionary is simply: “bias. See estimator”.[126]

In the 1920s and ’30s, three statisticians developed the theory and methods that remain the basis for most statistical analyses carried out today: Ronald A. Fisher, Jerzy Neyman and Egon Pearson (son of Karl Pearson).[8] But it was Fisher who developed the bulk of it, including the distinction between a population and sample, the modern conception of a statistical model, analysis of variance, p-values, significance tests, and maximum likelihood estimation.[135] His most important innovation, however, was randomised experiments, published in his 1935 book *The Design of Experiments*.[138] And in just over a decade it would lead to the first randomised controlled trial; designed by Austin Bradford Hill.[139] This approach to study design ensures that the bias known as ‘confounding by indication’ cannot occur, though it does not protect against any of the other biases that observational studies can fall prey to.[140] However, confounding by indication is a common source of bias when randomised treatment allocation is not used.[141]

The methods that Fisher developed were fundamentally aimed at providing “objective” means by which conclusions could be formed from data.[142] In other words, conclusions that were independent of personal biases.[143] These methods included his version of significance testing which incorporated a suggestion to use $p < 0.05$ as a cutoff to decide significance, a custom that has lasted more than 90 years. Nevertheless, Fisher believed that the interpretation of the p-value should be made each time by the researcher, and he included $p < 0.01$ as another rule that he sometimes used. His thoughts on the appropriate use of such decision rules were not clear in his writing, however.[144]

This apparent arbitrariness in decision making led Neyman and Egon Pearson to propose “hypothesis tests” so that further rules could be imposed on, and restrict, decisions.[145] By predefining the Type 1 error rate (probability of rejecting the null hypothesis when it is really true) and the Type 2 error rate (probability of accepting the null hypothesis when it is really false), along with a null and an alternative hypothesis, the researcher would know whether to accept or reject the null hypothesis at the end of the experiment, thereby limiting personal biases.[145] In fact, Neyman and Pearson believed that “as far as a particular hypothesis is concerned, no test based upon the theory of probability can by itself provide any valuable evidence of the truth or falsehood of that hypothesis".[146] What they wanted, instead, was a rule that would tell them how they should behave regarding the null and alternative hypotheses.[147]

Fisher did not favour this predefined hypothesis testing approach,[148] but scientists soon adopted what was, perhaps, the easiest option of combining Fisher's suggestion to use p < 0.05 as a decision rule, misinterpreting it as the "observed Type 1 error", while mostly ignoring the Type 2 error rate.[147] And with a null and alternative hypothesis defined beforehand they could then accept or reject the null hypothesis at the end of the experiment. This model for scientific experiments and studies has been the dominant method for making inferences ever since, yet an implicit assumption when using tests that produce p-values and confidence intervals is that non-random errors, meaning sources of bias, do not exist for the problem at hand, regardless of the regression model or estimation technique used.[130] One implication is that plausible sources of bias may be overlooked by researchers who are not sufficiently aware that the non-existence of non-random sources of bias is only an assumption.[149]

2.4 Reasons to classify types of bias

To reduce the chance of bias affecting their research findings, researchers need to recognise how such bias might occur, and to do this at each stage of their study (e.g. design, implementation, analysis, interpretation, publication). But it would be unrealistic to expect anyone to learn and reliably recall every type of bias that has ever been catalogued. To assist and provide a basis for understanding the nature of bias, a range of conceptual tools have been proposed and developed over time, such as bias classification schemes and risk of bias checklists, and they continue to evolve. Causal diagrams are also increasingly being used to help identify and reduce the effects of bias in research studies,[59] and they provide one way to avoid the confusion that stems from ambiguous terminology, such as 'selection bias'.[150] First, however, we will consider the potential benefits of a widely accepted classification system.

2.4 Reasons to classify types of bias

There are several distinct advantages to having a classification system of bias, and one that is common to many people. The first, already suggested above, is that it will assist in learning about and recalling specific types of bias that a study may be susceptible to. In 2010, Chavalarias and Ioannidis[151] identified 235 separate bias terms in a review of PubMed titles and abstracts, with 40 terms in over 100 papers each. They also noted that "the wide diversity in this nomenclature makes categorisation difficult". Nevertheless, categorising specific types of bias, such as 'choice of treatment influenced by a participant's health', into a vague sort of category hierarchy, is something we do naturally over time, anyway. There is evidence to suggest that much, or perhaps all, of our thinking involves a constant flow of conceptual categories that are formed over time through analogy-making, with new concepts understood by relating them to existing concepts through their similarities, and in time, similar concepts are grouped, or 'chunked',[152] into new information units that we can process mentally as a single category. And all of this depends on our personal experience.[153] So, without awareness of an existing classification system, each person with enough experience involving bias identification will end up mentally grouping specific biases in a haphazard and not always helpful way. For example, someone with limited experience in clinical trials might group under a 'doctor bias' category, without thinking deliberately about it, 'the bias caused by doctors ignoring random treatment allocation' along with 'the bias caused by doctors treating the intervention patients differently to control patients in an unblinded trial'. However, the 'doctor bias' concept would not help in recognising this type of bias if the next trial they were involved in saw nurses assigning the intervention.

Learning about specific types of bias through a well-designed classification system would facilitate the formation of a more useful and enduring mental hierarchy that (1) reduced the complexity of many types of bias, (2) highlighted the similarities and differences, and as such, (3) showed how each bias could be easily related conceptually.[154] A coherent classification system would also guide the formation of checklists used to determine the risk of bias in studies in a systematic review. Existing examples of this include Cochrane Risk of Bias tool for randomised trials,[155,156] the Quality of Cohort studies tool (Q-Coh),[157] and the Risk Of Bias in Non-randomized Studies of Interventions [ROBINS-I] tool.[158]

On the other hand, while risk of bias judgements are important for evidence synthesis, or even for judging a single study, it would clearly be better if there were fewer studies at high risk of bias in the first place. To this end, checklists of some kind may help. They have had widespread success in the aviation industry (e.g. preflight checklists) as well as product manufacturing,[159] and in recent decades they have seen increasing use in medicine, initially in nursing, and now also with doctors, especially in critical care settings.[160] By assisting with memory recall and establishing a minimum standard of bias identification and control, checklists derived from a bias classification system might prove a feasible method of improving the quality of research.

The other obvious benefit of a shared classification system is ease of communication. It has been noted by many that the variation in bias terminology and classification can create confusion in communication and understanding.[150,161–165] In 1992,[166] the goal of developing a widely accepted bias classification scheme was called the 'holy grail of epidemiologic research'. It seems an unlikely prospect, now, yet remains a worthy goal, and each contribution may bring a larger consensus one step closer. On the epidemiological side, people like David Sackett, Oli Miettinen, and many others have developed classifications of bias in research. But before any of these were considered, statisticians examined the problem from a mathematical perspective, and their work had an enormous impact on scientific research, including all the areas related to health.

2.5 Early concepts of bias in health research

One of the themes of this chapter is the variation in meaning of important terminology, like the words *bias* and *valid* in Chapter 2. An additional word that is used in a number of ways is *epidemiology*. It originally referred only to epidemics, however, its meaning has expanded over the last 90 years[118] so that definitions at times imply all research that relates to human health, including experimental studies like clinical trials,[167] though usually the definition will refer to groups or populations of people, rather than individuals. On the other hand, use of the term 'epidemiological studies' is more likely to refer to observational studies, but not experimental studies,"[167] while 'clinical trials' always refer to experiments where (at least some) study participants are deliberately given a treatment. This confusion of terminology can lead to confusing definitions, as evidenced by the Wikipedia page for 'epidemiology', where the first paragraph states "Epidemiology is the study and analysis of the distribution (who, when, and where) and determinants of health and disease conditions in defined populations."[168] However, the first sentence of the third paragraph starts with "Major areas of epidemiological study include disease causation, …", and ends with "… and comparisons of treatment effects such as in clinical trials". Hence, to avoid ambiguity in this chapter, we will use 'health research' as an umbrella term for all research relating to health, while separating epidemiology and clinical trials. Partly, this is because bias has come to be classified somewhat differently in these two areas.

Research on human health has a long history, but before the development of study designs and statistical methods that could aid analysis, it was difficult to avoid the cognitive biases that affect the perception of cause and effect, such as confirmation bias. A clear example is the fact that bleeding people with an illness, also known as bloodletting, survived as a standard therapeutic treatment from ancient times until late into the 19th century. Many factors would have contributed to its ineffectiveness not being discovered, including well-established traditions among physicians; ill people preferring to be treated rather than left alone; and the effect of confirmation bias where physicians would have focused their attention on those who improved or recovered following treatment, thus confirming their belief, unaware that they would have improved or recovered without treatment.[169] Bloodletting was still commonly used when the physician Pierre-Charles-Alexandre Louis (1787–1872) assessed the treatment by comparing the average number of deaths and time to death, or time to recovery, between those who were bled and those not bled, for patients with typhoid fever, pneumonia, and angina tonsillaris.[139] While not the first physician to compare patient outcomes by group rather than individually, he was nevertheless the most prominent to show a preference for average number statistics over clinical judgement.[139]

With the growth of statistical theory in the early 20th century, an obvious application was the study of health and disease. Karl Pearson had an interest in promoting the new statistical methods to the medical profession and occasionally contributed to *The Lancet* and the *British Medical Journal*.[170] One physician, Major Greenwood (not a military title)(1880-1949), became a statistician in 1910 after training under Pearson, and in 1924 published an article in The Lancet titled "Is the statistical method of any value in medical research?". While it clearly promoted the use of statistics, the final paragraph contained:

> When I first took an interest in these matters, more than 20 years ago, there was some tendency to treat the statistician or biometrician as a pariah, and he acquired the virtues and vices of a minority, a certain courage and a certain trick of over-emphasis - they always characterise a fighting minority. Now, statistics and statisticians are perfectly respectable; there may even be a risk of putting the claims of the statistical expert too high. ... The statistician must be the equal not the predominant partner.[171]

Sander Greenland recently labelled this "a prescient warning against inference dominated by statisticians".[86] This would also apply to statistics throughout health research, whether a statistician is involved or not, and points to the evident overconfidence that many researchers feel about the accuracy of their results. This issue will be examined in more detail in Chapter 4.

2.6 Clinical trials

Around the same time that Fisher was revolutionising statistics, in 1927 Greenwood was appointed as Professor in Epidemiology and Vital Statistics at the newly created London School of Hygiene and Tropical Medicine (LSHTM).[139] One of his students was Austin Bradford Hill (1897-1991) (known as Tony to his family and friends, he included his middle name Bradford to be distinguished from the physiologist A. V. Hill),[172] who would take over Greenwood's post when he retired in 1945.[170] In 1946, influenced by Fisher's work on randomised experiments, he designed what is considered the first properly designed randomised controlled trial (RCT); it aimed to assess the efficacy of streptomycin as a treatment for tuberculosis.[173] Although placebos were not used, allocation of streptomycin and bedrest, or just bedrest, was random and contained in sealed envelopes to preserve the randomisation. Within a decade, concerns by clinicians about withholding treatments from control group patients had given way to concerns about the claims from drug companies, with a wave of new medications entering the market in the 1950s, cementing the place of RCTs in health research.[174]

2.6 Clinical trials

The masking or blinding of a treatment with, for example, placebos or sham procedures for patients, and some form of deliberate ignorance for investigators to allow blind assessment, sometimes literally a blindfold, can occasionally be found in research studies over the last few hundred years.[175] Following World War II, however, its value as an addition to the new RCT methodology was soon realised, enabling a further reduction in the bias that patients and investigators could subconsciously impart to the data; in the case of investigators, this was sometimes called 'experimenter bias'.[176]

Beginning in the 1960s, governments took advantage of these developments in experimental design and began to require pharmaceutical companies to conduct clinical trials, both randomised and blinded, as the only way to show sufficient proof of efficacy and safety before regulatory approval would be granted.[139] This sparked a boom in the number of RCTs that added to the available evidence, yet often the evidence for a particular intervention is not consistent. This led to the rise of meta-analyses and systematic reviews in the 1980s as the only way for the medical community to make reasonably informed decisions.[177] It also led to many statisticians specialising in clinical trials, both in private companies and in academia, with the focus of concerns about bias tending to be different to bias concerns in observational epidemiology. Not surprisingly, the terminology has also evolved differently, and this is most evident in the way bias has been classified in the various 'risk of bias' assessment tools that have been developed over the last 20 years.

Through the 1980s and '90s, many scales and checklists were published that could help researchers judge the methodological quality of RCTs.[178] This task is clearly important for anyone conducting a systematic review, with or without a meta-analysis, but the limited utility of summary scores from the use of a scale was well recognised by the end of the century.[179] This drove the development of more comprehensive "risk of bias" tools, such as the Cochrane Collaboration's Risk of Bias tool for randomised trials (2008), [155,180] and the ROBINS-I tool for non-randomised intervention studies (2016).[158]

The Cochrane tool highlights the different way that categories of bias have developed since the 1960's in comparison to observational epidemiology, with the classification system in Table 1 given as part of their tool.[180]

Table 2.1 Bias domains in the Cochrane Collaboration's Risk of Bias tool[155,180]

Bias domain	Brief description	Examples
Selection bias	Systematic differences between baseline characteristics of groups compared	Inadequate generation or concealment of allocation sequence
Performance bias	Systematic differences between groups in the care, treatment or exposures, other than the intervention	No blinding of participants or trial staff to treatment allocation
Detection bias	Systematic differences between groups in how participant outcomes are determined	No blinding of outcome assessment
Attrition bias	Systematic differences between groups in completeness of outcome data resulting from participant withdrawals or exclusions	Inadequate procedures to retain participants or measure outcome
Reporting bias	Systematic differences between reported and unreported findings	Selective outcome reporting
Other biases	Sources of bias relevant in specific trial designs or circumstances	Carry-over in cross-over trials; recruitment bias in cluster randomized trials; contamination where experimental and control interventions get mixed

On the other hand, the ROBINS-I tool for non-randomised intervention studies (including cohort, case-control and quasi-randomised studies) uses domains of bias that are closer to the common tripartite classification (confounding, selection bias, measurement or information bias) used in epidemiology, with 7 domains of bias, grouped by the stage of research:[155]

Pre-intervention

1. Bias due to confounding
2. Bias in selection of participants into the study

At intervention

3. Bias in classification of interventions

Post-intervention

4. Bias due to deviations from intended interventions
5. Bias due to missing data
6. Bias in measurement of outcomes
7. Bias in selection of the reported result

2.7 Classifications of bias in epidemiology

Outside clinical trials, health research usually falls under the heading of epidemiology, although in this sense it has been called ‘traditional epidemiology’.[181] The discipline mostly relies on observational (non-randomised) study designs, such as case-control and cohort designs, to assess the distribution, potential causes of disease and other health-related states like injury, as well as interventions. When assessing health interventions, similar designs are used, and aim to provide evidence that is either additional to or not feasible to obtain with an RCT. This includes research on long term efficacy, rare side-effects, and efficacy and safety in a large and diverse clinical population.[158] In this thesis, we have mostly restricted the scope to health intervention research.

2.7.1 A common classification of bias

At the beginning of the 1950s, epidemiology was in the process of expanding from a discipline long associated with communicable diseases, like typhus, malaria, tuberculosis, and many others, to one that would also take on noncommunicable diseases, a relatively new field, with targets such as lung cancer and cardiovascular disease.[182] But the considerable research that was spurred by the tobacco-lung cancer debate was also followed by criticism from statisticians about their research and analysis methods. They expressed scepticism about results derived from case-control studies that contained few protections against selection bias and no agreed upon methods for analysing the data.[182]

However, the criticism had a positive effect, stimulating numerous developments in statistical design and analysis.[183] Nevertheless, many epidemiologists remained resistant to formal methods.[86] Efforts to get epidemiology onto a firmer methodological foundation started to gain momentum in the 1970s and '80s, led by people such as Olli Miettinen and Kenneth Rothman at Harvard University, and joined by Greenland, Morgenstern, Kleinbaum, Kupper, and others.[182] By the end of the 1980s, epidemiology had transformed from a 'classical' to a 'modern' phase, where epidemiologists were more likely to have PhDs instead of medical degrees, and most would have some training in statistics.[183] In the process, the discipline became much more mathematical, resulting in for example, the introduction of 'cumulative incidence' to better distinguish 'risk' from 'rate'; methods for matching in cohort and case-control studies; case-control designs were split into three types based on how the controls were sampled; and distinct types of bias were more carefully defined or identified, with their similarities and differences better explained.[183]

These methodological developments were generally aimed at reducing the chance of bias, and some are discussed in sections 2.8.2 and 2.8.3 on confounding and selection bias. Out of this process came the idea of classifying biases into confounding, selection bias and measurement bias, first mentioned in an article by Kleinbaum et al. in 1981,[184] but based on ideas developed by Miettinen.[185]

2.7.2 Confounding

The word *confounding* has been used in two primary ways by groups that are distinct, yet often closely related:

1. epidemiologists, who use the word in its oldest and most commonly used sense that describes a mixing together of separate causal effects with the effect of interest[8]
2. mathematical statisticians, for whom confounding relates to a concept called non-collapsibility, where an association is non-collapsible if the summary measure of

association (e.g. odds ratio) changes when conditioning on, compared to not conditioning on, a potential confounder[186]

The result is that when the word confounding is used, miscommunication can easily occur if the meaning of the word is assumed incorrectly.

Confounding may have been the earliest type of bias to be identified, with the concept appearing in a variety of 18th and 19th century treatises, sometimes beginning with the English philosopher John Stuart Mill (1806-1873),[187] although it was observed earlier, and was sometimes used as a criticism of another's study.[116] This early concept can be broadly defined as the non-comparability of groups,[188] or as the British statistician G. Udny Yule (1871-1951) described it in 1903:[189] a "fictitious association caused by mixing records". In other words, a mixing of the effect of one factor on an outcome with the effects of other factors on that outcome,[8] and it can also be roughly understood in the same fashion as the well-known idiom 'like comparing apples and oranges'.

Use of the word *confounding* did not appear in health research until 1970,[116] but appears to derive from its use by Ronald Fisher[116] who included a long chapter with the title 'Confounding' in his 1935 book *The Design of Experiments*.[138] But contrary to the meaning implied by the word now, Fisher described an experimental design that could take advantage of 'confounding'. One example he used involved small agricultural land plots containing fertiliser made up with differing amounts of each ingredient. Called a factorial design, it meant that more than one comparison was possible. In this example, the experimental units are the different amounts, sometimes zero, of each ingredient and the measured outcome is the amount of corn produced from each plot. If certain interactions between experimental units, that is, ingredient combinations, were not of interest in the analysis, the precision of the main effect, such as the ideal level of one of the ingredients, could be increased by eliminating some high-order interactions; that is, by deliberately introducing 'confounding'. The book's influence did not come from this, however.[188]

The word *confounding* appeared next in an influential 1959 methodologic paper in sociology by Leslie Kish.[188] Meanwhile, the concept of confounding, which at this time consisted of the two criteria (a) the confounder must cause the outcome, and (b) the confounder must be associated with the exposure under study, was discussed in occasional health research articles, such as the 1959 landmark paper on smoking and lung cancer by Jerome Cornfield et al..[190] The word confounding finally appeared in the epidemiology literature in 1970 with an article on matching by Olli Miettinen,[191] who has said he got the word from Fisher.[192] It then appeared in a few influential articles and books through the 1970s, including Kenneth Rothman (1975)[193] and David Sackett (1979).[194] With Greenland and Neutra (1980)[195] and Miettinen and Cook (1981),[196] a third requirement for a *confounder* was added: the confounder must not be a mediator on the causal pathway between the exposure and outcome.[116]

Perhaps the most important development was the 1986 article by Sander Greenland and James Robins titled ‘Identifiability, exchangeability, and epidemiological confounding’.[197] Using the potential outcomes framework, they drew a connection between epidemiological confounding; the term *identifiability* from mathematical statistics, which relates to whether the parameters in a statistical model can be identified from the available data, which depends on no unmeasured confounding;[198] and *exchangeability*[†] from Bayesian statistics, which means the same data would be obtained if the intervention group participants received the control treatment and the control group received the intervention;[197] in other words, if the participants in each group are *exchangeable*, it means they are sufficiently identical that the same data would be expected if they were, in fact, exchanged. Greenland and Robins also discussed *collapsibility*-based definitions of confounding which state that, if after stratification the effect measure (e.g. odds ratio), in each stratum is the same and also equals the crude effect measure, then the effect measure is said to be collapsible and the crude effect measure is unconfounded. They agreed with Miettinen and Cook[196] that a collapsibility-based definition is not ideal because it depends on the chosen measure of effect, and they give the example that a cohort study might find the risk difference collapsible but the odds ratio not collapsible. They conclude that a comparability-based

[†] (not to be confused with an 'exchangeable working correlation' used in generalized estimating equations)

definition of confounding is preferred over collapsibility-based ones, and should relate in some way to assumptions about exchangeability.[197]

Finally, it is worth noting that some authors prefer to keep confounding and bias as separate concepts. For example, in "Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and Elaboration" (2007):[199]

> Bias is a systematic deviation of a study's result from a true value. … Bias and confounding are not synonymous. Bias arises from flawed information or subject selection so that a wrong association is found. Confounding produces relations that are factually right, but that cannot be interpreted causally …

But this interpretation is not common in non-experimental epidemiology.

2.7.3 Selection bias

Like confounding, some types of selection bias were identified prior to the 20th century. For example, the British statistician and public health proponent William Augustus Guy (1810-1885), who would go on to serve as president of the Statistical Society of London,[139] tested the possibility that self-selection of workers might have biased an association between occupation and 'pulmonary consumption'.[117] An early example in the 20th century was demonstrated by Joseph Berkson (1899-1982), an American statistician who, in 1946,[200] used algebraic analysis to show the theoretical possibility of what came to be known as Berkson's bias, though he only relates it to hospital-based case-control studies.[183] However, it is now thought unlikely to have had much effect on the results of epidemiological studies,[201] though the ensuing controversy it generated may have helped drive the development of more general selection bias theories.[183] And a quick search on Google Scholar suggests it is still prompting ideas.[202]

In 1977, Greenland published 'Response and Follow-Up Bias in Cohort Studies' in which he states that, at that time, selection bias was a well-known problem in case-control studies, perhaps because of Berkson's warning, yet selection bias was less well known as a possibility from loss to follow-up in cohort studies.[203] A subsequent paper by Kleinbaum, Morgenstern and Kupper in 1981 offers a definition of selection bias as "a distortion in the estimate of effect resulting from the manner in which subjects are selected into the study population".[184] This is quite general, however, and suggests that many sources of bias more commonly thought of as confounding would instead be labelled selection bias; such as confounding by indication where, for example, a doctor 'selects' the patients that are prescribed the treatment, based on their symptoms or health history, which in turn has an influence on the probability of the outcome occurring.[118]

Confounding and selection bias have been distinguished by Rothman et al. (2008)[8] with:

> … differential selection that occurs before exposure and disease leads to confounding … In contrast, selection bias as usually described in epidemiology … arises from selection affected by the exposure under study …

Examples of selection bias that are common across epidemiology include differential loss to follow-up; missing data from reluctance of participants to provide detailed information; and self-selection or volunteer bias. On the other hand, healthy worker bias, can be classed as selection bias or confounding, depending on the definition of the bias and the classification system.[204]

2.7.4 Measurement bias

Under the heading of *measurement bias*, also known as *information bias* or *measurement error*, we find errors in the measurement or recording of information about participants, including their baseline characteristics, exposure status and outcome data.[8] A bias will exist when these errors differ between comparison groups. For errors in discrete data, such as the recording of sex or disease status, the term *misclassification* is often used, sometimes divided into *differential misclassification* and *nondifferential misclassification*, where the first refers to misclassification that is more likely for one of the study groups, and the second refers to equal likelihood of misclassification for each group.[118]

The way in which misclassification errors can produce a bias seems intuitive, and this may have been why it was the first type of measurement bias discussed theoretically in the epidemiological literature;[182] in a 1954 article by the American statistician Irvin Bross.[205] Further types of measurement bias were discussed soon after as epidemiology and biostatistics rapidly evolved.[182,183] Other examples include recall bias, response or self-report bias, and a bias sometimes called observer-expectancy bias, detection bias, or ascertainment bias, that can be reduced with blinded outcome assessment.

2.8 Classifications that did not catch on

Substantial change rarely occurs unless the existing way of doing things is challenged, and we saw this in the 1970s with the campaign to introduce more methodological rigour into epidemiology. But at the same time, certainty can never exist about the best way forward, and no two people will view a problem from the same perspective. This means that different people will come up with different solutions, and this is what has happened and continues to happen with the classification of bias in epidemiology.

One of the earliest and best known classifications of bias was put together by David Sackett (1979)[194] and based, in part, on earlier work by Murphy (1976)[206] and Feinstein (1967).[207] He lists 35 biases and groups them by the stage of research they occur in:

1. reading-up on the field
2. specifying and selecting the study sample
3. executing the experimental manoeuvre (or exposure)
4. measuring exposures and outcomes
5. analysing the data
6. interpreting the analysis
7. publishing the results

Sackett presented the list at a symposium on case-control methodology, however, he included biases specific to cohort studies as well. In fact, it is one of the most comprehensive taxonomies produced for epidemiology in terms of the areas it covers. For example, although he specifically tried to avoid the inclusion of ‘biases of rhetoric’ (p.51), which he thought were not appropriate for the symposium aimed at the design of case-control studies, he nevertheless includes a few such biases in the first stage, ‘reading-up on the field’, such as ‘The all's well literature bias’, and the ‘One-sided reference bias’. He also included biases that relate to the analysis and interpretation of results, usually not combined into a single taxonomy.

Although many of the names he gave to biases have not survived, such as his original publication has inspired a new initiative based at the University of Oxford's Centre for Evidence-Based Medicine: The Catalogue of Bias Collaboration and the associated Catalogue of Bias website (catalogofbias.org).[208]

In the 1980's, the main development was the growing popularity of dividing bias is in epidemiology into confounding, selection bias and information or measurement bias. One prominent academic who disagreed with this classification was Alvan Feinstein, well known for occasional disagreements with Miettinen, Rothman, Greenland and many others,[209] who exerted part of his influence as co-editor of the Journal of Clinical Epidemiology from 1982 until his death in 2001; though it was called the Journal of Chronic Diseases until 1988. Feinstein preferred to classify biases into:[210]

1. *susceptibility bias* (the same as confounding by indication)
2. *performance bias* (different treatment or phenomena experienced by groups)
3. *detection bias* (different methods of outcome measurement)
4. *transfer bias* (differential loss to follow-up)

And he thought that "Instead, the customary approach is to use vague terms, such as "information bias", "selection bias", and "confounders".[211]

A few years later, Choi and Noseworthy (1992)[212] extended the now common three category framework to "include subclassification according to the type of study design: cross-sectional, case-control, retrospective cohort, and prospective cohort". This can be seen in some later taxonomies of bias, as well such as Delgado-Rodríguez and Llorca (2004).[185]

That same year, Steineck and Ahlbom (1992)[213] published "A definition of bias founded on the concept of the study base" which utilised Miettinen's idea of the 'study base', a concept that he recently said could be "rather subtle".[214] Steineck and Ahlbom described the study base as "a specific slice of person-time; it is from the study base that the data are collected",[213] while Kass,[166] in a commentary on Steineck and Ahlbom's paper, described the study base as "the source population of individuals to be enrolled in an epidemiologic study". Steineck and Ahlbom used this concept to classify biases into one of three distinct stages of a study: definition of the study base, data collection "on disease events and person-time among the exposed and unexposed from the study base", and analysis of the data. It didn't catch on, and Kass suggests that a potential drawback of Steineck and Ahlbom's approach was the use of new terminology for familiar biases, such as "analysis deviance" instead of "specification bias" or 'misspecification'.

In a different approach by Maclure and Schneeweiss (2001),[215] instead of defining a different classification system of bias – they used confounding, information bias and selection bias – they presented an alternative model to help us understand how biases might influence our perception of causal effects. As such, it is similar to an alternative classification system. They used the analogy of a telescope that contained lenses and filters, the "episcope", through which an epidemiologist observes possible causal effects in a population. Each lens or filter is where certain biases act. They then combine, as if within a telescope, to distort our perception of a possible causal effect. Eleven layers of lenses and filters were described:

1. The causal effect, if it exists
2. Random confounding
3. Correlated causes producing non-random confounding
4. Making of and recording of diagnoses
5. Recording of exposures
6. Missing data and data aggregation errors
7. Hypothesis generation and forming of cohorts
8. Selection of cases and controls and loss to follow-up
9. Interpretation of results
10. Judgments of journals after paper submission (publication bias)
11. Biases in reviews and meta-analyses

They also used causal diagrams to describe specific examples of biases within each layer.

2.8 Classifications that did not catch on

Using text-mining and the PubMed database, Chavalarias and Ioannidis (2010)[151] searched for "235 bias terms and 103 other terms that appear commonly in articles dealing with bias", while noting that:

> New terms have been coined, cumulatively creating an extensive dictionary of bias nomenclature. Some biases are relevant to a wide spectrum of research designs, studies, and settings, whereas others are specific to special situations.
>
> The wide diversity in this nomenclature makes categorization difficult.

This helps explain why they chose a different strategy that avoided a classification scheme; instead they identified clusters of bias terms that were organised and displayed in network visualisation maps and in tables. The bias terms came from all areas of biomedical research with many specific to certain areas e.g. "codon usage bias". One thing they found that is relevant to this discussion is that the terms publication bias, confounding, selection bias, and response bias (also called self-report bias, a type of measurement bias) have been increasing noticeably in the literature over the past few decades. However, one unexplained curiosity is that the term 'performance bias' was not mentioned.

Three final classifications warrant mentioning. One was proposed by Weisberg (2011),[165] with 20 sources of bias grouped into 5 categories: Sampling (e.g. participation voluntary); Assignment (e.g. subject can influence assignment); Adherence (e.g. requirements onerous for subjects); Exposure ascertainment (e.g. inaccurate exposure reported or recorded); Outcome measurement (e.g. inaccurate outcome reported or recorded). Another, by Howe et al. (2015),[216] explained how biases normally classified under confounding, selection or measurement bias could instead be characterised as missing data problems. The final system takes, in a sense, an approach that advocates tightening existing definitions, rather than suggesting something entirely new. Schwartz et al. (2015)[217] with "Toward a Clarification of the Taxonomy of "Bias" in Epidemiology Textbooks", expressed a desire for epidemiology textbooks to all use exactly the same "consistent taxonomy" of bias, and they go on to suggest one that is based on the standard three categories of confounding, selection bias and information bias, as well as random error. However, history does not present many examples of disparate terminology being successfully merged into one, so this does not seem likely to be a productive exercise.

2.8 Classifications that did not catch on

As to why the three-part classification system became dominant in epidemiology, we can think of a few possible reasons:

1. It was one of the first coherent systems to be proposed - the critical period appears to be the late 1970s and the 1980s, when a lot of new terminology was being introduced into health research
2. A related reason is the popularity of textbooks written by proponents of this system, especially "Modern Epidemiology" by Kenneth Rothman in 1986[218] (with Sander Greenland as co-author for later editions), which had become the most cited epidemiology textbook by 2006[219]
3. Its simplicity in terms of only three classifications, though this was at the expense of leaving out biases relating to data analysis, and the interpretation and communication of the results which, for example, Sackett[194] had included

Finally, a possible explanation for the different terminology associated with clinical trials, at least in regards to risk of bias assessments, revolves around the many disagreements between Feinstein, who was often highly critical of observational study designs,[220–223] and epidemiologists or biostatisticians such as Miettinen, Rothman and Greenland. Terms now used in clinical trials such as "performance bias" and "detection bias" appear to originate with Feinstein, whereas the term confounding came from Miettinen (though Fisher first proposed it) and this word became dominant in epidemiology, where selection bias is used for the same concept in clinical trials.

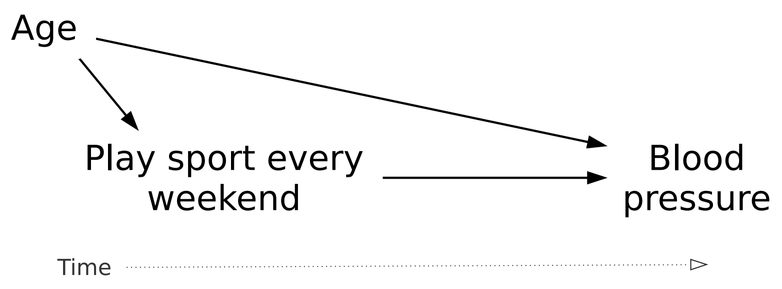
This brief review of the history and current status of bias classification in health research suggests at least two things. First, it does not appear likely that a consensus would ever be reached on a common system, not only across all of health research, but even just in observational epidemiology which Schwartz et al.[217] showed did not use consistent definitions across the field. And second, it seems likely that new classification systems will continue to be suggested.

Chapter 3
Causal Diagrams

3.1 What causal diagrams are

A causal diagram is a visual model of the cause and effect relationships between variables in a system of interest.[224] Such a system might comprise the variables that are causally related to an activity, such as playing sport every weekend, and an outcome it may affect, such as blood pressure. For the research question 'does playing sport every weekend reduce the chance of high blood pressure', imagine that we analysed a sample of patient blood pressure measurements, where all patients, regardless of age, were asked if they played sport every weekend. A simplified system containing only three variables is shown in Figure 3.1, and describes how confounding might occur in this example. In this case, while playing sport might decrease the chance of high blood pressure, age may confound the observed relationship because older people are less likely to play weekend sport but more likely to have high blood pressure.

Figure 3.1 Simple causal diagram that describes possible confounding



3.1 What causal diagrams are

Put simply, causal diagrams can make it easier to draw realistic causal inferences.[59] They can help by stimulating the identification of more potential confounders and sources of selection bias than might otherwise have been considered; and they can help to illuminate the set of assumptions that are made when inferring a result from the statistical analysis.

The causal diagram in Figure 3.1 is also an example of a *directed acyclic graph*, or DAG, by far the most common type of causal diagram used in health research. In this case, the word 'graph' refers to its meaning from mathematical graph theory: a set of points where some points are connected by lines;[134] instead of meaning a chart or plot as commonly used in data analysis.[142]

A *directed graph* is one in which the connecting lines represent a direction from one point to another, and a *directed acyclic graph* is a directed graph where it is not possible to move from one point to another, following the directed lines (usually drawn as arrows), and arrive back at the original point. In other words, one cannot follow the arrows along a path that forms a closed loop or cycle.[224] This is necessary for a causal model so that past events can cause future events but future events cannot affect past events.[225] It is also a common convention for a DAG to be drawn where time flows to the right.[59] This may enhance both the drawing and interpretation of a DAG because it enables a causal story[226] to be constructed that agrees with English and other language speakers' intuition that time flows from left to right.[227] And the dominant view in cognitive science is that people understand the world largely by constructing causal narratives or stories.[228–230]

Unlike most introductions to causal diagrams in epidemiology that include some of the formal language and procedures, in this thesis we have instead attempted an alternative approach that avoids the mathematical terminology of DAGs unless it will hinder an initial understanding. We suspect that most of the concepts can be understood using words in common English, and with fewer new words to keep stored in working memory, an ease of understanding will hopefully be promoted.[131] In Chapter 4, we expand on the influence that cognitive ease has on the decisions people make, such as whether to continue learning about causal diagrams. Once the core concepts have been understood and can be retrieved from long-term memory, the more formal terms such as nodes, edges, vertices, d-separation and back-door criterion[231] can easily be associated with those concepts.

3.2 Brief history

The geneticist Sewall Wright, in 1921, was the first to use directed graphs to represent probabilistic cause and effect relationships among a set of variables.[232] He developed path diagrams and path analysis,[233] which later went on to be used in the social sciences in methods such as structural equation modelling in the 1970s.[97] Path diagrams also led to probabilistic DAGs known as Bayesian networks in the 1980s, with artificial intelligence researcher Judea Pearl one of the leading developers.[91] And soon after, causal path diagrams and probabilistic DAGs were merged[234] by Spirtes, Glymour and Scheines (1993)[235] and Pearl (1995, 2000)[236,237] into a formal theory of causal diagrams, before its introduction into epidemiology in 1999 by Greenland, Pearl and Robins.[224] At the same time, a concerted effort by Pearl and others fought against the longstanding prejudice in statistics over causality.[234]

Pearl, especially with his book *Causality: Models, Reasoning, and Inference* in 2000,[237] developed a detailed structural theory of causation that he claims incorporates and unifies other approaches to causation, namely causal graphs, structural equation modelling, and potential outcomes.[238] It is a mathematical theory and includes a new operator he called the $do(\cdot)$ operator that is to be interpreted as an intervention in the underlying model.[237] The word ‘structural’ is in reference to the causal structure underlying effects in a research study, as represented in a causal DAG,[239] and Pearl defines a structural causal model as one that represents the causal relationships underlying a dataset.[91] As such, it represents any assumptions we might make in the analysis of that data.[*] Each structural causal model is related to a graphical model, usually a DAG,[226] but it is mainly his development of DAGs that have earned widespread application.

Nevertheless, some prominent statisticians still regard causal diagrams as inferior to other options. For example, Donald Rubin states that while these “graphical approaches seem to be a clear advance with respect to causal inference over older, less subtle graphical approaches”, he nevertheless feels that “the framework is inherently less revealing than the

[*] ‘data’ is used here in the modern sense as a mass noun rather than the plural of datum

potential outcomes framework because it tends to bury essential scientific and design issues".[240]

Despite such views, however, over the last two decades the use of causal diagrams has grown, and they have even been called the "flagship of the new methods",[241] though perhaps claims that "there must be few epidemiologists who do not use directed acyclic graphs"[242] are more applicable to some universities than others, given that many epidemiological articles do not yet mention them. Nevertheless, numerous researchers and statisticians are now promoting their use,[59,86,150,241,243–253] so continued growth does seem likely.

3.3 Structural classification of bias

Of the classifications of bias examined in the previous chapter, two apparent problems are that:

1. the same terminology is often used with different meanings, such as selection bias
2. the same type of bias is often known by different names, for example, see Table 3.1

This can lead to both misunderstandings in communication between researchers and confusion of students in epidemiology and biostatistics.[217]

Table 3.1 Bias terms in clinical trials and epidemiology adapted from Mansournia et al.[150]

Cochrane Bias Domain	Epidemiologic Term
Selection bias	confounding or selection bias
Performance bias	Biased direct effect or confounding
Detection bias	Measurement bias
Attrition bias	Selection bias
Reporting bias	Non-structural bias that cannot be represented in causal diagrams

Language is full of ambiguity,[132] however, which possibly evolved because of our need to communicate with the least effort needed and rely heavily on context instead.[254] Thus, ambiguity cannot be eliminated. But in a series of papers from 2002 to 2009,[255–257] Miguel Hernán and colleagues took an entirely different approach by defining types of bias using causal diagrams. They did not avoid terminology, but they were able to give precise definitions for the standard epidemiological terms of confounding, selection bias, and measurement bias; calling it the "structural classification of bias".[256]

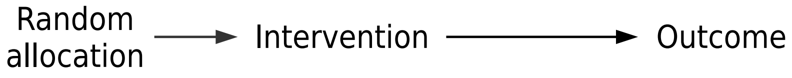
Before defining the types of bias, we need to understand how to use a causal diagram once the variables and arrows have all been added. In a DAG, the arrows represent the belief that one variable causes another, and in a DAG with many variables, a causal pathway can be traced by following the arrows from one variable to another, and this can indicate how one variable might influence another further down the causal pathway. An association, on the other hand, does not have a direction, and in a DAG, an association will exist between two variables if a path can be traced along some arrows, regardless of the direction of the arrows.[59]

In terms of the structural definition of bias, an association between two variables in a study can be explained by one of three possible causal structures. With an intervention and an outcome as the two variables of interest, these are:[256]

1. Cause and effect: The intervention caused changes in the outcome, or the outcome caused changes in the intervention, on average, in the study population

    ↳ For example, a randomised trial with a true causal effect (Figure 3.2)

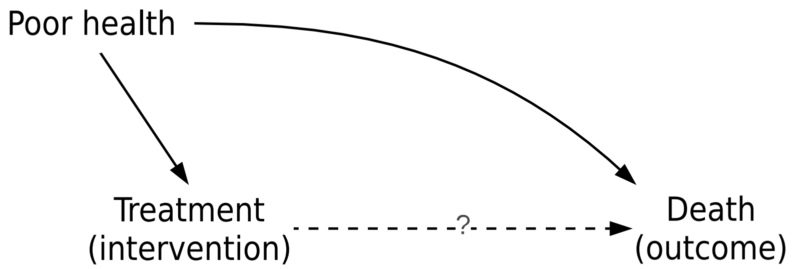
Figure 3.2 Cause and effect in a randomised trial



2. A shared cause: A third variable, a confounder, caused either the receiving of the intervention, or the type of the intervention received, and also caused changes in the outcome

↳ For example, Figure 3.3 depicts an observational study where poor health makes it more likely that a study patient was given a particular intervention, for example, an expensive treatment drug, but poor health also makes it more likely that the patient will die, producing an association between receiving the intervention and the outcome (which may, in this case, cancel out an association produced by the intervention causing a reduction in the chance of death)

Figure 3.3 Confounding produced by a common cause



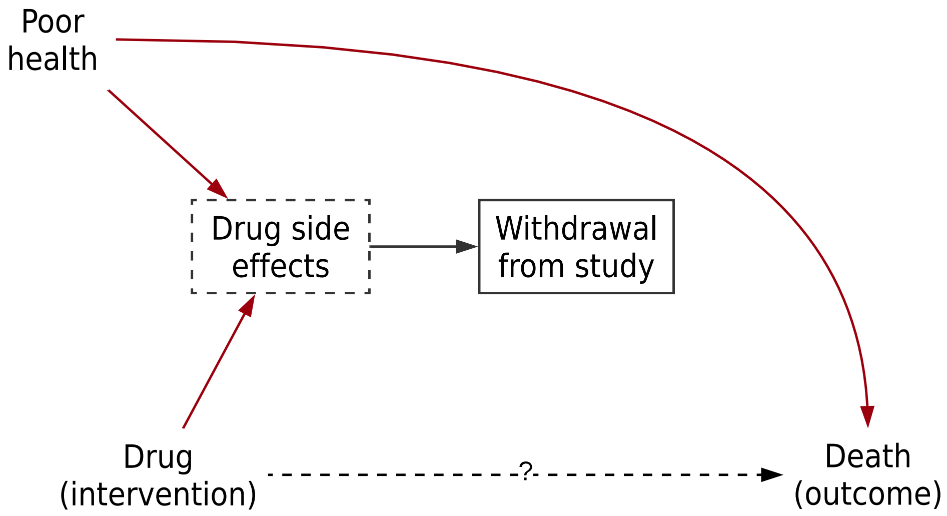
3. A shared effect: A third variable that was *conditioned on*[†] was affected by both the intervention and the outcome; that is, a third variable, called a collider, was affected by either receiving the intervention, or by the type of the intervention received, and the collider was also affected by the chance of experiencing the outcome; called selection bias or collider bias

    ↳ For example, in a randomised controlled trial depicted in Figure 3.4, patients with poor health are more likely to die (the outcome), and receiving the treatment drug instead of the placebo (the intervention) was more likely to produce side

[†] The term ‘conditioned on’ or ‘conditional on’ derives from probability theory and intuitively means that the data or the results of the analysis depend on information contained by the variable(s) conditioned on. This might occur by restricting the data to a specific value of a variable, such as including only patients who did not withdraw from a study, or it might occur by adjusting the results of the analysis to remove the effect of (‘condition on’) confounding variables, usually by including the variables in a regression model or stratifying. Conditioning on a variable can also be described as narrowing the scope of the discussion to those situations where the variable is a given value; in other words, where the variable is held constant. [231].

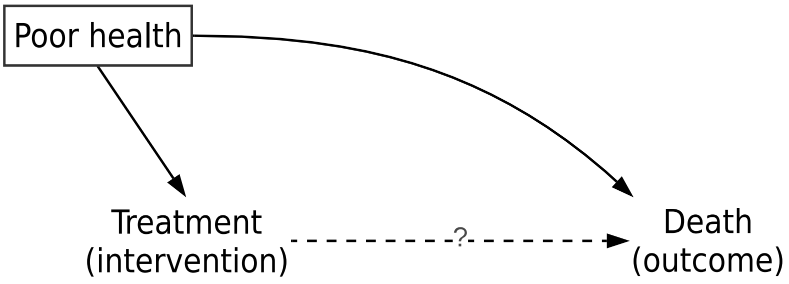
effects (the shared effect) that led to withdrawal from the study, which is the same as conditioning on patients not withdrawing from the study

Figure 3.4 Selection bias created by conditioning on a common effect (collider bias)



In Figure 3.3, the association between the intervention and the outcome can be blocked by conditioning on the confounder, often achieved by stratifying or including the confounder in a regression model. A common practice with causal diagrams is to place a border around variables that are conditioned on, such as in Figure 3.5; and also done in Figure 3.4, where the results of the study are conditioned on patients remaining in the study, hence a border is around the variable 'Withdrawal from study'. But in this case, the effect on bias is the same as conditioning on whether the patients got side effects, and this is why it is called collider bias, because the arrows 'collide' at the collider. With this example, however, the selection bias from dropout can be removed by conditioning on poor health, thus blocking the associational pathway highlighted in red. With the structural classification of bias, both selection bias and confounding result in a lack of exchangeability, or non-comparability, with statistical adjustment achieved using the same type of methods for both types of bias.[256]

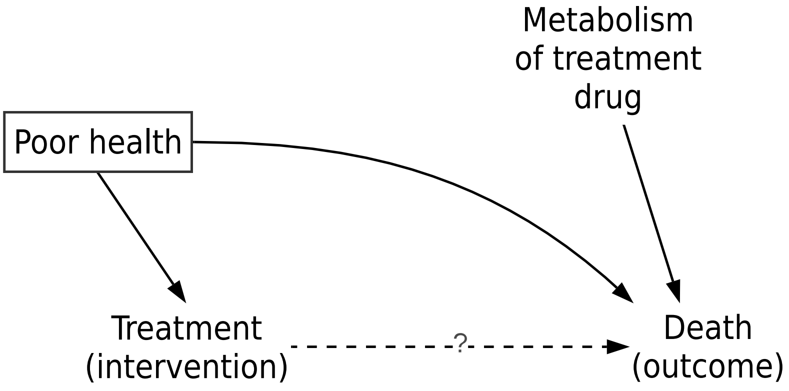
Figure 3.5 DAG with confounding removed by conditioning on the common cause



The remaining type of bias is measurement bias, and Hernán and Cole (2009)[257] identified 4 general types using causal diagrams. However, because there is no apparent confusion of terminology regarding measurement bias, we won't explore this type of bias any further.

Finally, there is sometimes confusion about the difference between confounding and effect modification,[39] so an effect modifier was added to the causal diagram in Figure 3.6. A fundamental difference is that confounding is a bias that we aim to either prevent by design or remove by conditioning, whereas effect modification is a property of the causal effect being studied and ideally, we would like to estimate and describe it.[8]

Figure 3.6 DAG with confounding and the addition of an effect modifier



In the example in Figure 3.6, poor health is a suspected confounder of the relationship between taking the treatment drug and the chance of dying. However, it is also suspected that the causal effect of the drug will vary depending on how quickly the drug is metabolised and that is determined by each patient's genotype, though not in a way that can be tested.

Hence, the drug’s metabolism in each patient does not affect their chance of receiving the treatment.

Effect modification is especially important for the generalizability of any findings, because if the intervention only works, or is only safe for some people, then such effect modifiers need to be identified. Hence, another term for effect modification is effect heterogeneity.[258] An intervention is also likely to work better for some individuals than for others, potentially leading to different decisions on whether to use it if information were available to be able to predict someone’s outcome.

It is important to note, however, that causal diagrams are limited in how well they can portray effect modification, where we cannot usually distinguish between multiple possible modifications of the effect.[59] And in general, it is not possible to show how variables might interact using causal diagrams, though some work has been done to suggest exceptions may exist.[259] There have also been proposals to modify causal diagrams so that interactions could be displayed, but this would mean they would no longer be directed acyclic graphs.[260]

The main advantage of using the structural classification system to define biases like confounding and selection bias is that, although terminology still plays a role, the use of a causal diagram to guide decisions about the study design, analysis or interpretation, means that the terminology a researcher uses for these biases should not affect such decisions. In this way, the problem of ambiguity can be avoided. But even if a researcher does not use causal diagrams, this classification system might provide the rigorous, formal definitions of confounding and selection bias that will appeal to some researchers, especially those unhappy with the uncertainty that can surround whether a bias should be called confounding or selection bias.[256]

3.4 Constructing a causal diagram

Non-DAGs

Although standard DAGs are by far the most common type of causal diagram in use, they may not always be the best choice. But while various alternatives have been developed, such as chain event graphs,[261] compartmental model diagrams,[262] diagram-based analysis of causal systems,[263] graphical chain models,[262] and single world intervention graphs, simple conceptual causal models can also be constructed[264] without regard to the rules that go with DAGs, to help understand the possible causal paths between the variables in a study.

DAGs

The basic actions needed to construct a causal DAG are:

1. Add variables for the exposure/intervention and the outcome
2. Add all other variables for which data was collected or is expected to be
3. Add the potential confounders collected in the study or expected to be
4. The causes of any one variable currently in the diagram may be included, but causes of two or more variables must be included for it to be considered a causal DAG[256]
    - This includes suspected unknown common causes of two or more variables, in which case a symbol such as U might serve as a label
5. Draw an arrow between any variables thought likely to be causally associated that indicates the direction of the causal relationship
6. If the study is longitudinal and a prior value of the outcome Y affects the exposure X, which then affects the following Y, each instance of the exposure and each measurement of the outcome must be shown as separate variables, for example: $X_0 \rightarrow Y_0 \rightarrow X_1 \rightarrow Y_1$
7. Do not draw an arrow between two variables if available knowledge and the plausibility of potential mechanisms suggests it is unlikely one may cause the other

  - This also means that our research conclusions rest, in part, on our assumption that no causal relationship exists between them

Software

A possibly neglected issue in the promotion of causal diagrams has been the availability of software and published guidance on the choices that are available. A number of software packages have been developed over the years to facilitate the drawing and analysing of causal diagrams. One of the first was TETRAD in 1986,[265] becoming the TETRAD Project in 1998,[266] but it was aimed primarily at structural equation modelling. It has since been expanded and is available at www.phil.cmu.edu/projects/tetrad/, however, it is still not really aimed at most types of health research.

The only software package specifically designed to create DAGs that has been made known to health research through publications in epidemiology journals is DAGitty,[267] available at www.dagitty.net and also as the R package 'dagitty'.[268] As such, to our knowledge, it is the only package that has been mentioned whenever the software used to create a DAG is listed in an article. And while it is being improved from time to time, it is non-commercial software with very few programmers, so progress is slow, and its limited features and interface full of what to many, is technical jargon, may act to discourage some researchers from getting started with causal diagrams.

Alternatives to DAGitty are mostly diagramming software packages like Microsoft Visio (visio.microsoft.com), LucidChart (www.lucidchart.com) and Gliffy (www.gliffy.com). However, while easy to use, they do not offer features that are specific to DAGs.

3.5 Uses of causal diagrams

The widespread use of diagrams to convey abstract information shows it is generally accepted that diagrams can assist in the understanding of abstract concepts, at least sometimes.[269] Research in cognitive science has suggested that diagrams can make it easier to find the information relevant to a concept,[270] such as the causal paths between variables that might lead to selection bias in a study. Diagrams can also help when considering alternative possibilities by making all the possibilities explicit,[271,272] such as when a researcher is forming conclusions at the end of a study, based partly on alternative explanations for the results.

Causal diagrams, which in most cases are DAGs, provide an intuitive framework that can help researchers conceive of and understand the biases that might influence a study, and can make communicating more difficult concepts easier than explaining solely with words.[59] This makes DAGs a useful tool to enhance the communicating of concepts relating to bias, whether teaching basic concepts[59,150,253,273] or publishing the results of methodological research.[274–277] This is especially the case with the structural classification of bias, covered in the previous section, but DAGs have also been used to explain more specific types of bias, such as different types of time-dependent confounding,[278] missing data biases,[244,279–281] and possible explanations for apparent paradoxes such as Simpson's paradox,[282] the birth weight paradox,[283] and the obesity paradox.[284–286]

It is now well established that an analysis of observational data should take into consideration not only the study design, but also substantial background subject-matter knowledge if the goal is to obtain evidence regarding a causal association.[255,287] Otherwise, important uncontrolled confounding might not be considered when making inferences, or variables might be included in a model that instead of reducing bias, increases it via collider bias. Also, by constructing a causal DAG that aims to adequately represent background causal knowledge, a researcher or statistician might be prompted to include variables that otherwise would not have been considered.

This means that if a DAG is constructed during the planning stage of a study, potential confounders that otherwise might not have been considered, can instead be either controlled by modifying the design, or else have data collected on that variable so it can be used to adjust the analysis.[224] The DAG can also be used to communicate this understanding to fellow investigators or study staff, or to ask for feedback from subject matter experts.[59]

Once a study's data has been collected, a DAG can be useful in identifying previously unconsidered sources of bias, such as from missing data,[244] loss to follow-up[279] or time-dependent confounding.[288] And this can help plan the analysis with the most appropriate methodology.[47]

It is also possible to use a DAG to identify a minimally sufficient set of variables that is needed to control for confounding in the analysis.[224] This would exclude variables such as intermediates on the causal pathway between the exposure and the outcome. The program DAGitty was recently criticised, however, because it can calculate such a set automatically. This may potentially mislead a researcher into thinking they could successfully control for confounding by adjusting for the variables DAGitty chose, even though important confounders were not included in the DAG.[76]

Finally, a DAG can help with the interpretation and communication of the results. By making the assumptions on which causal inferences rest more explicit, such as the possibility of confounding from sources that were not controlled, conclusions by researchers might be more likely to be adequately cautious. The DAG can, and should, also be included with any published report, to help communicate the sources of bias identified, how they were controlled in the design and the analysis, and the assumptions and associated uncertainty that remains following the analysis. Unfortunately, it is still not uncommon to find articles that merely mention that a DAG was used to help select the model covariates, without providing the DAG itself.

Chapter 4
Understanding the biases in health intervention research

4.1 Introduction

Two questions are very important when considering the quality of health research, yet are hard to answer accurately: How often are research study findings sufficiently biased that there are consequences for human health and, has progress been made over recent decades in reducing the level of bias in health research? Articles criticising the standard of health research can be found in any era (for example),[121,149,289–295] and many factors will contribute to the volume of such criticism, such as changing expectations of research quality. But these articles are important for motivating improvement, and while expectations might now be higher than in the past, there is an abundance of evidence to suggest that, despite regular educational efforts by a variety of researchers and statisticians, improvements in research quality appear to have been somewhat limited. Methodological advances might have helped experts, but misuse of those methods by many less familiar with the details could still lead to biased results. For example, logistic regression became popular in health research in the 1980s[296] but problems with its use since then have been well documented.[297–299]

To motivate change, researchers and statisticians must first be aware that an important problem exists. To address this issue, the current level of bias in health research is examined next in section 4.2, with the evidence strongly suggesting that not only are the findings of many studies likely to be biased, but that improvements over time have been modest, at best. And this is despite researchers now receiving considerably more training in statistics than many of their predecessors.[300]

Possible reasons for this lack of improvement are discussed in sections 4.5 – 4.10, with the focus primarily on the science of cognition and how our brain possesses many energy-conserving and time-saving features that, while mostly helpful, can also lead to errors in judgement and behaviour that can introduce bias into a research project. Some cognitive biases are well known, such as confirmation bias, but many are known only in fields such as cognitive psychology and behavioural economics. However, the topic has had a well-established presence in the area of medical decision making since the 1990s,[301–306] and some physician training programmes now include education on the role of cognitive biases in diagnostic errors and poor treatment.[307] Application of this knowledge to research environments has been limited, but in recent years a number of articles have raised the issue in relation to scientific research generally,[308,309] or health research in particular.[109,310–314]

One recommendation we make is the use of causal diagrams to make it easier for researchers to identify, and appropriately control for, potential sources of bias. These diagrams will be described, along with the associated structural classification system of bias, that seeks to avoid some of the problems with terminology highlighted in Chapter 3, such as the confusing number of ways the term 'selection bias' is used. The chapter will conclude with examples of biases commonly encountered in health intervention research; described with the use of causal diagrams.

4.2 Evidence of bias

One of the problems with detecting bias is that we can never know with certainty what the true result or inference should be, and this means that allegations or suggestions of bias can be easily rejected by the authors of studies alleged to be biased, and in our experience mostly are. This makes it more difficult to determine whether the results from the initial study are indeed biased, even though many would suspect the original authors may be guilty of a conflict of interest akin to 'myside bias'.[315] Evidence can still be gathered, however, and inferences formed about the degree of bias that might exist in an area of research.

4.2.1 Continuum from unintentional bias to fraud

Our primary focus is bias in health intervention research that is not deliberately created by researchers. Hence, cases of fraud, such as the deliberate fabrication of data or statistics, will not be included here, partly because clear cases of scientific misconduct are most likely uncommon.[316,317] However, there are many questionable behaviours that lie on a continuum between scientific fraud and unintentional bias.[316–320] These include

- presenting a relationship found to be statistically significant as being the main hypothesised target of a study, when it was really just one of many possible relationships tested, a practice known by many names, such as 'data dredging',[321] 'data trawling',[322] 'P-hacking',[323] and 'significance questing'[324]
- not publishing a study's results because they contradicted one's previously published findings, or in the case of a commercial interest, not publishing results that might harm those interests, such as results showing little difference between a pharmaceutical company's drug and a placebo; this is called publication bias[325]
- concealing a conflict of interest, such as a source of funding that has a financial interest in a particular outcome of the study, even if the researcher does not believe it influenced their behaviour[326]

In some cases, although the behaviour is deliberate, it might be so common as to be standard practice in their field,[327] which in the eyes of many will make it acceptable behaviour.[328] Nevertheless, the consequence of questionable behaviour by researchers is an increased chance that their study will be biased.

4.2.2 Randomised versus non-randomised study results

In the 1970s and 1980s, a number of reviews compared the results of non-randomised with those of randomised trials testing the same interventions and outcomes.[329–331] They found that non-random selection was associated with results more likely to favour the treatment and with larger effects. Under the assumption that RCT effect sizes were likely to be closer to the true effect size, these results suggested that non-randomised studies, which include observational studies, were more susceptible to confounding caused by their treatment allocation procedures. Similar comparisons of randomised and non-randomised studies since then have yielded mixed results, however, with some finding differences[332–334] and some not.[335,336] Comparisons where the non-randomised trials used propensity score methods have been likewise mixed, with differences sometimes found[337,338] and sometimes not.[339]

Randomised controlled trials have been called the "gold standard" of cause and effect research since 1982[340,341] because randomisation of treatment allocation greatly reduces the chance of substantial confounding, assuming there are a sufficient number of participants and concealment of treatment allocation is used. It also facilitates valid interpretation of inferential statistics like p-values and confidence intervals;[342] and randomisation is essential for blinding of participants, investigators and outcome assessors sufficient to prevent biases like observer bias, response bias and placebo effects.[343,344] However, use of the term "gold standard" can sometimes sound like religious dogma, implying a perfection that does not exist.[341,345,346] In reality, RCTs investigating the same intervention often report contradictory results,[347] yet when the results from an RCT are compared with a non-randomised trial or an observational study, the RCT's results are often assumed to be the correct ones. Many feel that when comparing results from different studies, the individual quality of each should be considered as important as the strength of their underlying research design.[347–352]

While we should probably avoid automatically favouring an RCT's result over those of a contradictory observational study result, the fact that they disagree highlights that either one is biased, or in fact, they do not test the same intervention or outcome. Some specific examples from the last two decades include:

Hormone replacement therapy and risk of coronary heart disease

Observational studies in the early 1990s concluded that postmenopausal hormone replacement therapy (HRT) led to a reduction in the risk of coronary heart disease.[353,354] Later randomised controlled trials, however, found no beneficial effect of HRT on cardiovascular disease,[355–358] leading to numerous post-mortems of what went wrong[359–362] and much criticism of observational epidemiology,[363] with prominent article such as "The scandal of poor epidemiological research",[364] and newspaper headlines like "Do We Really Know What Makes Us Healthy?".[365]

Antioxidant vitamin supplements

Findings that oxidative stress has a role in many diseases such as cancer, cardiovascular disease, and neurodegenerative diseases[366] have led many people to take antioxidant vitamin supplements such as β-carotene, vitamin C and vitamin E to try to prevent these diseases.[367] Early observational studies suggested they could provide a protective effect against these diseases,[368,369] but many RCTs since then have found either no effect,[370] or an increased risk of disease.[367,371,372] Articles with titles like "Epidemiology—is it time to call it a day?"[373] followed, along with other similar commentaries.[364,374,375] Yet a visit to any store selling vitamins will quickly reveal the continued popularity of taking these supplements. An explanation for this will be explored below in the section on causal thinking.

Statins

While the efficacy and safety of statins has been well established,[376] there has been plenty of controversy surrounding adverse events,[377] and the effect of statins on non-cardiovascular diseases.[378] This controversy is partly due to the conflicting results of studies, and especially between observational studies and RCTs, with links between statins and adverse events much more common in observational studies than RCTs.[379] This difference may be due to 'nocebo' effects,[380] which are adverse symptoms experienced during an unblinded trial that the participant mistakenly attributes to the treatment. Similarly, suggestions from observational studies that statins might prevent some cancers were not backed up in RCTs,[381] with selection bias and immortal time bias possible explanations.[382]

4.2.3 Reviews, commentary and further evidence of bias

The ongoing frustration with research quality, especially as it relates to the conduct and interpretation of statistical analyses, is well summarised in the opening lines from "*Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations*" (2016)[294] by the prominent statisticians and epidemiologists Sander Greenland, Stephen Senn, Ken Rothman, John Carlin, Charles Poole, Steve Goodman, and Doug Altman:

4.2 Evidence of bias

> Misinterpretation and abuse of statistical tests … remain rampant.
> … correct use and interpretation … seems to tax the patience of working scientists. This high cognitive demand has led to … interpretations that are simply wrong … yet these misinterpretations dominate much of the scientific literature.

This article coincided with the unusual step taken by the American Statistical Association of releasing a "Statement on Statistical Significance and P-Values",[383] a response to the increasing concerns expressed in the literature over recent years about a "reproducibility crisis"[384] in all areas of science, including health research.[385] And one of the main concerns is the continuing oversimplification of scientific reasoning encouraged by the use of "null-hypothesis significance testing", where the standard binary cutoff of $p < 0.05$ is used to decide whether an effect might be real or not. In terms of causal inference, it can:

- lead to confounders being dropped from models, such as with stepwise regression;[128]
- encourage the perception by many researchers, including statisticians, that a single study can tell us whether an effect is real or not[386]
- strengthens the natural human tendency toward overconfidence in the accuracy of our inferences[313]

Compared to articles criticising the use of null-hypothesis significance testing, very few have been published defending the practice,[149] although it may have limited utility for some research tasks.[387]

To a large extent, the above article on misinterpretation and misuse of statistics, mirrors those that have appeared regularly for decades. A small sample of titles can be seen in Table 4.1. These commentaries, and the many others that have been published, all suggest that a sizable proportion of health intervention research studies have been analysed and interpreted poorly, greatly increasing the chance that the results are biased.

Further evidence comes from reviews investigating conflicting results in health research (Table 4.2). When results from difference studies conflict, it suggests that at least one of the studies must be biased.

4.2 Evidence of bias

Randomised controlled trials are also susceptible to bias, though not to confounding by indication if the randomisation was done properly and concealed before allocation. Some articles that found evidence of bias in RCTs are listed in Table 4.3.

Table 4.1 Articles criticising the misuse of statistics from each decade of the last 80 years

Year	Article title
1942	"Tests of Significance Considered as Evidence"[289]
1959	"Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance - Or Vice Versa"[388]
1960	"The Fallacy of the Null-Hypothesis Significance Test"[389]
1966	"Statistical Evaluation of Medical Journal Manuscripts"[290]
1979	"Some Problems of Statistics and Everyday Life"[295]
1982	"Statistics in Medical Journals"[390]
1985	"The Religion of Statistics as Practiced in Medical Journals"[291]
1990	"How Trustworthy is Epidemiologic Research?"[391]
1994	"The Scandal of Poor Medical Research"[292]
2005	"Why Most Published Research Findings Are False"[392]
2018	"Medical Research - Still a Scandal"[393]

Table 4.2 Reviews investigating conflicting results in health research

Year	Article title
2005	"Contradicted and Initially Stronger Effects in Highly Cited Clinical Research"[293]
2007	"How Quickly Do Systematic Reviews Go Out of Date? A Survival Analysis"[394]
2011	"The Frequency of Medical Reversal"[395]
2013	"Pioglitazone and Bladder Cancer: Two Studies, Same Database, Two Answers"[396]
2013	"A Decade of Reversal: An Analysis of 146 Contradicted Medical Practices"[397]
2015	"Eggs and Beyond: Is Dietary Cholesterol No Longer Important?"[398]
2016	"A Corpus of Potentially Contradictory Research Claims from Cardiovascular Research Abstracts"[399]
2018	"Association Between Risk-of-Bias Assessments and Results of Randomized Trials in Cochrane Reviews: The ROBES Meta-Epidemiologic Study"[400]

Table 4.3 Articles with evidence of bias in randomised controlled trials

Year	Article title
1995	"Empirical Evidence of Bias: Dimensions of Methodological Quality Associated with Estimates of Treatment Effects in Controlled Trials"[401]
2005	"Identifying Outcome Reporting Bias in Randomised Trials on PubMed: Review of Publications and Survey of Authors"[402]
2008	"Empirical Evidence of Bias in Treatment Effect Estimates in Controlled Trials with Different Interventions and Outcomes: Meta-Epidemiological Study"[403]
2012	"Observer Bias in Randomised Clinical Trials with Binary Outcomes: Systematic Review of Trials with Both Blinded and Non-Blinded Outcome Assessors"[404]
2013	"Volunteer Bias in Recruitment, Retention, and Blood Sample Donation in a Randomised Controlled Trial Involving Mothers and Their Children at Six Months and Two Years: A Longitudinal Analysis"[405]
2014	"Bias Due to Lack of Patient Blinding in Clinical Trials. A Systematic Review of Trials Randomizing Patients to Blind and Non-Blind Sub-Studies"[406]
2014	"Comparison of Anticipated and Actual Control Group Outcomes in Randomised Trials in Paediatric Oncology Provides Evidence that Historically Controlled Studies are Biased in Favour of the Novel Treatment"[407]
2015	"Data Interpretation in Analgesic Clinical Trials with Statistically Nonsignificant Primary Analyses: An ACTTION Systematic Review"[408]
2016	"Empirical Evidence of Study Design Biases in Randomized Trials: Systematic Review of Meta-Epidemiological Studies"[409]
2017	"Congruence Between Patient Characteristics and Interventions May Partly Explain Medication Adherence Intervention Effectiveness: An Analysis of 190 Randomized Controlled Trials from a Cochrane Systematic Review"[410]
2017	"Cherry-Picking by Trialists and Meta-Analysts Can Drive Conclusions about Intervention Efficacy"[411]
2017	"Simple Randomization Did Not Protect Against Bias in Smaller Trials"[412]
2018	"A Review of Cluster Randomized Trials Found Statistical Evidence of Selection Bias"[413]

Finally, evidence of bias is also suggested by articles (Table 4.4) identifying problems with methodologies, errors, reporting biases, and also by retractions, where the implication is that many more articles containing errors or poor judgement in methodology, as well as deliberate fraud, would be retracted if those problems were discovered.[317,414,415]

Table 4.4 More articles with evidence of bias from the last 5 years

Year	Article title
2013	"Why Has the Number of Scientific Retractions Increased?"[416]
2015	"Biased and Inadequate Citation of Prior Research in Reports of Cardiovascular Trials is a Continuing Source of Waste in Research"[417]
2017	"Indirect Evidence of Reporting Biases was Found in a Survey of Medical Research Studies"[418]
2017	"Top Ten Errors of Statistical Analysis in Observational Studies for Cancer Research"[419]
2017	"The Distribution of P-Values in Medical Research Articles Suggested Selective Reporting Associated with Statistical Significance"[420]
2017	"Survival Biases Lead to Flawed Conclusions in Observational Treatment Studies of Influenza Patients"[421]
2018	"High and Unclear Risk of Bias Assessments are Predominant in Diagnostic Accuracy Studies Included in Cochrane Reviews"[422]
2018	"Interpretation of Epidemiologic Studies Very Often Lacked Adequate Consideration of Confounding"[423]
2018	"Kaplan-Meier Survival Analysis Overestimates Cumulative Incidence of Health-Related Events in Competing Risk Settings: A Meta-Analysis"[424]
2018	"Three Risk of Bias Tools Lead to Opposite Conclusions in Observational Research Synthesis"[425]

4.3 Consequences of bias

A memorable quote comes from an article by Andrew Vickers (2005):[426]

> A mistake in the operating room can threaten the life of one patient; a mistake in statistical analysis or interpretation can lead to hundreds of early deaths. So it is perhaps odd that, while we allow a doctor to conduct surgery only after years of training, we give SPSS to almost anyone.

Unfortunately, while this statement is quite plausible and, in some cases probably true, except for major studies where the results are likely to influence treatment, for most researchers analysing data, the link between their results being biased and people dying, probably contains too many steps and too much uncertainty to be a real enough threat to their conscience or reputation, such that it would change how carefully they conducted an analysis. Other, more immediate concerns, such as getting more publications, and getting them faster; providing results to colleagues in a timely fashion; or with the answers they are expecting or hoping for; may tend to drown out the less obvious consequences of their decisions.

More generally, research with biased conclusions, regardless of the sources of bias, and depending on how influential each study turns out to be, might:

- lead to new research that wastes money and the efforts of dedicated researchers if, had the original study's conclusions been closer to the truth, more productive research would have been done instead
- slow the availability of better health interventions through, for example, contradicting similar research, thus increasing the uncertainty over the efficacy of the intervention
- mislead clinicians and patients about the intervention most likely to help in their particular case
- lead to some people receiving care or using an intervention that makes their health worse than it would otherwise be
- contribute to a systematic review coming to the wrong conclusion

The Australian government provided $877 million to health research in the 2017 grant funding round through the National Health and Medical Research Council (NHMRC).[427] Funding from state governments, non-government entities such as charities, pharmaceutical and medical device companies, and private donations, would add considerably more each year to health research in Australia. Yet, in a widely publicised 2009 article in The Lancet,[3] it was estimated that as much as 85% of research investment might be wasted because of correctable errors in the production and reporting of research evidence. If the detection of bias in research continues to accelerate as it has over the last 10 years, the public's enthusiasm for funding health research might diminish.[428,429]

4.4 Efforts to reduce bias

Most of the articles listed in section 3.4, either by criticising current practice or by reporting evidence of bias, aim to promote better research practices and thus reduce the number of articles providing biased evidence. Some of these, along with other articles, give explicit recommendations to encourage widespread improvement, including:

- Better and more thorough training in statistics[294]
- More statisticians and greater use of their expertise[2,292,430]
- Better peer review, possibly including a statistician[431]
- Open peer review[432]
- Pre-registration of trials or study protocols[433,434]
- Reporting guidelines[199,435]
- Independent attempts at replicating study findings[436]
- Blind analyses[437]
- Lowering the p-value threshold for statistical significance[438]

And ongoing campaigns include

- Evidence-Based Medicine (EBM)[439]
- Strengthening Analytical Thinking for Observational Studies (STRATOS) initiative[440]

A frustrating problem

While it is likely that some improvement in the quality of research will have occurred over the last 40 years, progress has clearly been much slower than many would have expected given the efforts that have taken place. Many factors are likely at play, including bias-related methodological articles being swamped by the vast number of articles being published,[441] but this seems likely to be a minor cause. In the next section, we explore some insights from cognitive science in an attempt to explain why progress has been slow and in so doing, look for additional measures that might help.

4.5 Decisions in research and bias

To a large extent, a bias in the results or conclusions of a study will exist because of the decisions made by the researchers. These may include:

Decisions in the planning stage:

- which potential confounders to measure
- planned actions to reduce baseline measurement error and missing data
- how to measure the outcome if continuous, such as blood sugar or a pain rating,
    - for example, a continuous scale; ordered categories; or a judgement of responder/non-responder

Decisions during the conduct of the study:

- actions taken to increase the accuracy of measurement, reduce missing outcome data and encourage measurement of non-compliance

Decisions in the analysis of data stage:

- the choice of one or more of the frameworks discussed in Chapter 1 to help guide the analysis methodology
- whether to use a particular common bias classification system to help determine potential sources of bias, including in consultation with others involved in the project
- the researcher's preference for analysis methodology, such as:
    - common frequentist procedures with p-values, with or without null-hypothesis significance testing,
    - or methods within a Bayesian framework
- definitions chosen for different parameters that will depend, in part, on the study design and the availability of data, such as the start and end dates for time at risk of the outcome, or the eligibility window for the start and end dates that covariate data needs to have been collected in order to be included in the analysis dataset[442]

- their final choice of statistical model to estimate the effects of an intervention, from what might seem to be a bewildering variety of options[443,444]
- their choice of statistical software, which can range from simple tests on certain websites to programs such as SAS, Stata, R or SPSS

Decisions when interpreting the results:

- deciding how to interpret statistics such as p-values,
    - for example, using a 0.05 cutoff for 'statistical significance'
- deliberate efforts to think of alternative explanations of the results
- the use of tools to aid consideration of alternative explanations such as causal diagrams

Decisions when communicating the results:

- choice of words to either:
    - unambiguously convey the level of uncertainty remaining in the results
    - or use 'spin' to covey increased confidence in the accuracy of the results

Factors such as the researchers' level of knowledge, experience and understanding of statistical methods, all have some bearing on the way they investigate and try to find answers to causal questions. With the number of potential options, no two researchers are likely to take exactly the same approach to causal inference and so, not surprisingly, the answers to the same causal questions vary, as seen in section 4.2. But only one answer is true, and this suggests that some approaches are better than others. Some of the analysis decisions that can introduce bias include:

- the use of an inappropriate model, such as a linear model for a non-linear relationship
- using an inferior method for handling missing data, such as simply excluding non-complete cases
- using criteria to include or exclude covariates from a model that are not based on background subject-matter knowledge, such as a stepwise algorithm

- not using a method of checking for mistakes, such as another researcher or statistician checking the code or independently duplicating the analysis
- not conducting a sensitivity analysis to see if decisions in the analysis might have biased the result, or greatly underestimated the uncertainty

In general, however, with certainty about the best choices to make not possible, no research philosophy can be proven as best, though debate will continue. Nevertheless, some guidance can be obtained from mathematical and logical arguments, as well as evidence over time revealing which approaches produce results and conclusions that are less often contradicted by considerable later research. The approach that researchers take to the design, conduct, analysis, interpretation and finally, communication of their research, can end with very different, sometimes opposite research conclusions. It all comes down to avoiding the many sources of bias that otherwise result in biased causal inference, and this means making better decisions.

4.6 Insights available from cognitive science

To explore why a researcher or statistician might make a decision that leads to biased results and conclusions, we turn now to the science of decision making; an area that has grown over the last 60 years in fields such as cognitive psychology, behavioural economics and clinical decision making.[445] However, it is important to recognise that some notion of the thinking biases that affect decisions seems to be intuitively understood by almost everyone, with some terms commonly used for this concept including *human nature* and *human fallibility*.[138,328,446,447] A longstanding implicit understanding of this human susceptibility to bias in the research arena is revealed by the generally stated goal of finding *objective* methods of inference in the field of statistics,[143] and also by the strongly recommended technique of *blinding* in clinical trials.[175,344]

The burgeoning science of judgement and decision making reached a widespread audience, perhaps for the first time, in 1974 in the journal *Science*, with the article “Judgment under uncertainty: heuristics and biases”.[448] It was written by two Israeli psychologists, Amos

Tversky (1937-1996) and Daniel Kahneman (1934-), and according to Google Scholar on 3 Sep 2018, it has now been cited 46,715 times. Building on the work summarised there, Kahneman and Tversky went on to be highly influential in the field of decision making, and not only in cognitive psychology. Their work led to the establishment of behavioural economics,[449] in turn leading to the Nobel Prize in Economics for Kahneman in 2002[450] and Richard Thaler in 2017[451], as well as the recent popular books by Nassim Nicholas Taleb *Fooled by Randomness* (2004)[452] and *The Black Swan* (2010)[453].

Outside psychology and economics, however, findings from the decision sciences have so far had a much smaller impact, although interest is growing. One area that has made use of these ideas is clinical decision making. In fact, one of the first publications in cognitive science was by Paul Meehl with his 1954 book *Clinical Versus Statistical Prediction*.[454] It reported studies that suggested linear models of relevant predictor variables performed better at clinical prediction than experts; in this case, mostly clinical psychologists. Later studies extended this to medical decision making, with results suggesting that, at least in some cases, clinical intuition performed less well than a probabilistic analysis.[455–457] Much research followed that looked at the influence of cognitive biases in medical decision making (for example[301–305,328,458–464]).

In the rest of health research, however, concepts relating to cognitive bias have mostly been discussed without reference to research in the cognitive sciences.[446,465] For example, the recently created Bias Catalogue website (catalogofbias.org),[208] developed by a collaboration headed by the Centre for Evidence-Based Medicine at the University of Oxford; it describes 38 biases in detail, of which 16 are essentially cognitive biases, including 'Confirmation bias',[466] 'Positive results bias',[467] and 'Biases of rhetoric'.[468] On the other hand, an increasing number of articles have appeared in the last two decades that have focused on findings in cognitive science and their relevance to decision making in research,[20,109,308,313,319,469,470] including articles aimed at statisticians.[113,129,471,472]

4.7 Models of decision making

The general goal of science is to better understand some aspect of reality, often so we can exert some control over it.[473,474] But because our understanding can never be complete, we need some level of abstraction in the form of a model, the aim of which is a similar yet simpler representation of reality.[473] If sufficiently accurate, models can be very useful, though it is important to avoid model reification,[313] where we think as if the model was indeed reality, such as thinking the true value of a treatment effect really does lie within the bounds of an estimated confidence interval.[313] Yet models are more than just potentially useful tools; they are the only means by which we can understand the natural world, including the processes in our brain we call thinking. Hence, to better understand how decisions are made during research, we first need a model that describes how people make decisions. And even though many have been developed, debated and extended over the last 40 years in cognitive psychology,[475] an idea common to most models is that our decision making processes can be usefully classed into two broad types, hence the name *dual-process models*.[476] These two types of thinking processes are:

1. Type 1 processes and decisions act like automatic mental rules of thumb.[477] They are fast, effortless and mostly occur below our conscious awareness.[478] They include hard-wired heuristics as well as acquired skills.[228,479]

2. Type 2 processes are conscious, deliberate, relatively slow, and often require some effort.[475] They use logic and statistics,[477] but still involve the automatic Type 1 processes, which cannot be turned off. This type of decision-making process, being conscious, is how we perceive ourselves making decisions.[228]

Table 4.5 compares the Type 1 and Type 2 process characteristics that are frequently associated with dual-process models in the cognitive psychology literature.[475,478]

The underlying reality in our brain is, of course, far more complex, and there are researchers in the minority who feel that dual-process models are too vague,[480] and as a result, do not yield precise, testable predictions.[481] But for our purposes, the concept of dual-process thinking, and the heuristics and cognitive biases that relate to this model, can help us understand why causal inference in health intervention research is remaining defiantly resistant to our efforts to improve it.

Table 4.5 Characteristics frequently associated with Type 1 and Type 2 processes

Adapted from Evans (2008)[478] and Evans and Stanovich (2013)[475]

Type 1 processes (sometimes called System 1)	**Type 2 processes (sometimes called System 2)**
Automatic	Controlled
Nonconscious	Conscious
Low effort	High effort
Fast	Slow
Context dependent	Abstract
Pragmatic	Logical
Parallel processing	Serial processing
Autonomous	Involves mental simulations
Heuristic	Rational
Intuitive	Analytic
Impulsive	Reflective
Can produce biased responses	Can inhibit biased responses
Evolved early	Evolved late
Similar to animal cognition	Much more distinct in humans
Responds to basic emotions	Complex involvement of emotions
Universal	Heritable
Independent of cognitive ability	Correlated with cognitive ability
Independent of working memory	Limited by working memory capacity

4.8 Heuristics, cognitive effort and learned expertise

Heuristics can be thought of as automatic, and often subconscious, decision rules that allow for the fast decisions our evolutionary ancestors needed to make.[475] For example, decisions that prepare ourselves for possible danger, such as turning automatically toward a sudden unexpected sound,[228] or decisions important to social goals, such as automatically imitating the behaviour of the majority unless it conflicts with another goal.[482,483] But while physical danger is now a less frequent need for many, our need for fast decision making has not changed, so heuristics are considered helpful most of the time.[480]

Heuristic thought processes and decisions also use less energy than decisions that require effort,[484] and the brain requires at least 20% of the energy consumed by our body.[485] This helps explain why we evolved so that thinking that requires effort is often perceived as a mildly unpleasant experience, a feeling that leads to frequent avoidance of heavy thinking tasks unless a goal is sufficient to motivate the effort,[228,486–488] though the subjective effort needed at any time is subject to factors like the amount of sleep the night before,[489,490] the time of day,[491,492] stimulant drugs such as caffeine,[493,494] age,[492] and many others. Examples of avoidance behaviour include browsing news websites instead of replying to an important email; or avoiding tasks when analysing data if they are somewhat unfamiliar, such as checking model assumptions when such checking has not been done for a long time. This leads to one proposed built-in heuristic called the *law of least mental effort*[487] (see Table 4.6), though also known by other names, including *avoidance of cognitive demand*,[487] *cognitive miser*,[495,496] the *principal of least effort*,[486] and "lazy System 2" in Daniel Kahneman's widely read book *Thinking, Fast and Slow* (2011).[228] This heuristic helps guide our decisions about the cognitive tasks we undertake so that only those tasks sufficiently important to us will be attempted.

In psychology, cognitive effort or ease of cognition is often called *fluency*, and another proposed heuristic is called the *fluency heuristic*[477,497] (Table 4.6). This is similar to the *availability heuristic*[498] (Table 4.6), one of the three general-purpose heuristics, along with *representativeness* and *anchoring*, discussed in Tversky and Kahneman's famous 1974 *Science* paper.[448] The influence of anecdotal evidence on inferences in clinical research, at least for some clinician researchers, may relate in part to the increased ease of recalling the many occasions where a treatment had a particular outcome, compared to it being less easy to recall the occasions where the treatment outcome was different, thus leading to a belief that the anecdotal evidence they have observed is more common than it really is.[499,500]

Another related heuristic has been called the *take-the-first heuristic*[477] (Table 4.6). For example, after reviewing the results of their study, a researcher might consider the question "why did we find $p < 0.05$?" and find that the easiest answer that comes to mind is "because the treatment caused better outcomes than the control".

Table 4.6 Some heuristics relating to cognitive effort that have been proposed

Heuristic	Description
Law of least mental effort[487]	The tendency or urge to avoid cognitive effort unless there is sufficient motivation to make the effort required
Fluency heuristic[477,497]	If we consider a question and two alternative answers come to mind, then the one that is retrieved faster, which also means that it came to mind more easily, is the answer that we tend to give more weight to
Availability heuristic[448,498]	The frequency or probability of an event is judged by the ease with which relevant instances or associations come to mind; for example, judgements about how much work we did on a joint project compared to other members of the team
Take-the-first heuristic[477]	A tendency to accept the first (and easiest) answer that comes to mind

The word 'heuristic' is mostly reserved (in cognitive science) for models of built-in or hard-wired mental processes, assumed to exist by way of evolution,[480,501] though for some heuristics, especially social ones, it may be difficult to determine how much is innate and how much a learned skill. But where intuition does develop with learned expertise or the acquirement of skills, subconscious mental processes appear to work in a very similar way, at least from the perspective of our subjective experience.[479] And as with heuristics, a learned skill can include much activity that takes little to no cognitive effort to perform. As we become familiar with a task, we can often not only do it faster, we can do it with less mental effort.[502] Examples that most people can relate to include driving a car down a familiar road in light traffic, or in the case of a statistician, carrying out routine tasks when analysing data.[479]

4.9 Causal thinking

People appear to believe that almost all events are caused by previous events, and this may explain a general reluctance to consider phrases like "randomness", "random error", or "chance" as an explanation for a correlation between variables.[229] There is also strong evidence to suggest that people have a built-in preference for causal explanations.[228–230,503–506] This has variously been referred to as a "causality heuristic",[507–509] "causal intuition",[50,91,228,510–512] "causal illusion",[19,513] "causal thinking",[514,515] and a variety of other terms.[9,230] It may be experienced as a need to explain events we see as important,[229] so we can understand the effect our actions might have, and the actions of others. In order to respond appropriately to our environment, it needs to make sense to us.

It follows that we understand the world in terms of causes, not associations. Indeed, it may be that the only way we can understand a non-causal association between real events, that is, a correlation between two variables that is not causal, is by thinking of other causes that could produce the association. This could be a common cause, in other words, confounding, as suggested by the Common Cause Principle in philosophy,[516,517] or it could be through a mechanism that produces collider bias.[518]

4.9 Causal thinking

Others, particularly those trained in mathematical statistics, might explain a ‘chance correlation’ more abstractly as arising from the way subjects were sampled, or even more simply as ‘due to chance’. But while ‘chance’ may be the only explanation currently available, it is probably not the best possible explanation, or as Greenland (1988)[519] put it:

> … labelling a result as due to "chance" or "random" variation is analogous to diagnosing an illness as "idiopathic," in that it is just a way of making ignorance sound like technical explanation.

Evidence also suggests that we prefer explanations of events if they contain plausible causal mechanisms. We automatically construct a causal sequence, or causal story, by running mental simulations, and sometimes we will compare different explanations by comparing their simulations.[230] And taking into account the availability and take-the-first heuristics outlined in section 4.8, it seems reasonable to suppose that if a causal mechanism does not come easily to mind, it instead becomes easy to ignore, as if it did not exist. This may partly explain why many researchers find it easy to believe they have found convincing evidence, even though many alternative explanations were not considered. And it may also partly explain why many researchers believe a high p-value is best explained by “the null hypothesis is true”, whereas in reality, this interpretation is not valid.[294,520]

Likewise, ‘nocebo’ effects,[380380] where adverse symptoms in an unblinded trial are mistakenly blamed on the treatment, may instead result from the need for a coherent causal story or mechanism to explain their symptoms. And for many, the most obvious and plausible cause is the treatment.

In a similar fashion, the placebo ‘effect’ has often been attributed solely to psychogenic or psychosomatic mechanisms,[521] with the term ‘psychological factors’ commonly used,[522] and perhaps this is because a psychological mechanism is the most plausible and easiest causal explanation that comes to mind. In reality, however, many factors lead to placebo group members improving, including regression to the mean,[523] spontaneous improvement (natural course of the disease),[524] additional treatment sought by the patient,[525] methodological problems of the study,[524] and others, though including psychogenic and psychosomatic causes in some cases.

Finally, in section 4.2.2, one of the examples discussing incongruent RCT and observational study results concerned antioxidant vitamin supplements. Of note is that, despite the lack of evidence for any benefit after many trials, including RCTs, and possibly even harm, the popularity of antioxidant supplements has continued.[526] An explanation suggested by Ghezzi (2017)[527] is the appeal of the simple causal story that goes something like: a widely accepted theory[366] in science is that oxidative stress is a major cause of disease and aging → one way the body combats reactive oxygen species (ROS) and free radicals is by producing natural antioxidants → therefore supplements of antioxidants should help reduce the 'bad' free radicals.

However, the underlying biology now looks to be far more complicated than the oxidative stress theory entails, and this may explain why antioxidant supplements have so far failed to demonstrate robust health benefits.[527] If this absence of a meaningful effect is maintained, then from the point of view of preferring causal explanations that include mechanisms, we may find that without establishing a plausible and simple causal mechanism that can explain why antioxidant vitamins do not work, the belief that they are beneficial might be hard to change.

4.10 Cognitive biases

Heuristics serve us well most of the time, but not always, and when they lead to errors in a systematic fashion (i.e., not just random mistakes), cognitive biases and illusions result.[228] Table 4.5 lists some attributes of Type 1 processes, which heuristics use, with some that suggest a susceptibility to bias, such as 'automatic', 'fast', 'impulsive', and 'responds to basic emotions'.

However, we do not decide to think this way, and we are mostly unaware of it when it happens.[478] This makes cognitive biases hard to avoid, and everyone is susceptible[308,471] Even higher intelligence, or cognitive ability, will offer only some protection, and only for some biases.[528]

We can sometimes reverse or prevent cognitive biases by monitoring ourselves, but we cannot do this constantly; and it gets harder when we are tired, mentally fatigued,[529] or … relaxed and happy.[530]

Some examples of the cognitive biases that can influence beliefs and decisions in a research project are described below. Like the statistical biases in Chapter 3, the definitions of different biases sometimes overlap, and all will tend to evolve over time. But where statistical biases can often be given precise mathematical definitions, cognitive biases and heuristics are less precise models of our thinking processes, often developed by different researchers with different perspectives. Yet they need only be useful to be worthwhile; by helping us to understand observed behaviour and by accurately predicting behavioural responses in certain situations. And they can help us to better understand why measures such as training in statistics has not greatly improved the quality of causal inference in health research. This improved understanding can suggest which current measures are worth pursuing, and also lead to new ideas that we can try.

Cognitive biases relating to how we view our own

Naïve realism

A tendency to believe that we see the world objectively, or 'as it really is', referring both to physical reality and also to social and political issues. And we expect other reasonable people will perceive the same reality. Hence, if people disagree with us, it must have something to do with them rather than the issue, because we are objective. For example, they might be biased, lacking cognitive ability, not as informed as us, or thinking irrationally (make no sense) [320,531]

Bias blind spot

Related to naïve realism, we tend to believe that other people are more susceptible to cognitive biases than we are. [531,532]

Cognitive biases that relate to scientific reasoning

Motivated reasoning

People are motivated to use reasoning to reach accurate conclusions, but they are also motivated to reach particular conclusions, especially ones they already believe, or opinions they have previously expressed to others. To achieve both goals, people gather and evaluate evidence using strategies that feel appropriate, but which are also more likely to reach the desired conclusion. This leads, in turn, to biased beliefs that nevertheless seem objective.[533,534]

Confirmation bias

A commonly cited‡ definition is by Nickerson (1998):[169] "the seeking or interpreting of evidence in ways that are partial to existing beliefs, expectations, or a hypothesis in hand". And one of the ways it is sometimes countered in research is by using blind analysis.[437] As an example, Cox and Popken (2008)[535] note the tendency of some researchers to prematurely adopt a causal conclusion after inadequate observations. They then seek confirming evidence, or p<0.05, and ignore or underweight disconfirming evidence.[535]

Myside bias

Closely related to confirmation bias is myside bias, defined by Mercier (2017)[536] as: "A tendency to find arguments that defend our beliefs, whether they are supportive (if we agree with something) or refutational (if we disagree with something)". Mercier goes on to argue that confirmation bias is a flawed concept, because people do not seek information to confirm a belief that already exists. Instead, people tend to search for arguments that will defend their position; in other words, they have a myside bias.

To quote the epidemiologist Michael Marmot (1976):[537] "any scientist should begin a scientific paper with the phrase: 'Ladies and gentlemen, these are the opinions on which I base my facts … '"; and (paraphrasing the philosopher of science Thomas Kuhn): "scientists

‡ (3,845 citations in Google Scholar on 24 November 2018)

choose paradigms or research programs in which to work and then attempt to defend their positions". Deliberately provocative, but memorable.

Argumentative theory of reasoning

Closely related to myside bias is a theory of reasoning by Mercier and Sperber (2011),[538] where reasoning can be defined as a cognitive process that produces and evaluates reasons.[315] Rather than holding to the common view of reasoning as a means to make better decisions and improve our knowledge, they suggest that the main function is to argue; that is, to find reasons that will convince others and to assess others' reasons so we can either develop our counterargument, or instead change our mind, though only when appropriate. In their review of the research on reasoning and decision making, they conclude that the evidence strongly suggests that when people reason to produce arguments, they are biased and lazy, yet they are more objective and critical when evaluating the arguments of others. But Mercier and Heintz (2014)[315] also note that, while scientists display the same biases as lay people, science as a community has developed traditions and institutions that encourage, to some extent, the exchange and critical evaluation of each other's ideas, where argumentation plays a leading role.

Overconfidence bias

Known as overconfidence; three forms have been identified:[539]

1. Overestimation of our actual performance, either in the past or in the future. A well-established finding in psychology is that people typically feel that they performed better at a task than they really did.[540] An overestimate regarding the future is described by a well-known example called the planning fallacy, in which people tend to overestimate how fast they will get a complex project done, or in the way it is usually expressed: they underestimate how long it will take. Examples include writing a draft manuscript, buying a new car, preparing to teach a new course, or just about any other complex task sufficiently different to past projects that there will be many uncertainties that need to be considered. And unfortunately, being aware of past failures to accurately predict the time it will take does not seem to help.[541]

2. Overplacement of our relative contribution, performance, or skills. Also called the “better-than-average” effect. For example, “82% of people say they are in the top 30% of safe drivers”.[542]
3. Overprecision, relating to overconfidence in the accuracy of our numerical estimates.[540]

Some evidence of likely overconfidence bias can be found in the recent paper by Hemkens et al. (2018) “Interpretation of epidemiologic studies very often lacked adequate consideration of confounding”.[423] Their conclusion contained:

> … Even when confounding bias is mentioned, authors are typically confident that it is rather irrelevant to their findings and they rarely call for cautious interpretation. …

Dunning-Kruger effect

A twist on the overconfidence bias is that people lacking relevant skill or knowledge tend to overestimate how well they performed in a task, relative to those with greater expertise, who were more accurate in their self-assessments.[543–545] This disparity may result because such people lack the knowledge to recognise their errors, and hence suggests that researchers lacking statistical expertise will tend to be more overconfident about how well they performed their analysis than researchers with greater statistical expertise, who are more aware of the difficulties and errors of judgement that can bias results and inferences.

Cognitive biases where heuristics dominate

Ambiguity or uncertainty aversion

A tendency to choose the option with a known probability over one with an unknown probability, even though the chosen option could easily be the worse option if more information about the alternative were known. The same concept can be seen in the English proverb: "Better the devil you know than the devil you don't".[546] It is possibly explained by a “fear of negative evaluation” brought on by expecting that, if the unknown probability option were chosen, we would not be able to justify that choice to other people.[547]

Default effect

Making an option a default increases the chance that it is chosen, partly because considering other options involves cognitive effort.[548,549] It can be an effective intervention to promote better choices,[550] but can sometimes deter better options being chosen; for example, options in software for regression models[551,552] A similar bias is status quo bias[553]

Observer bias

Outcome assessments systematically deviate from the truth because the observers are influenced during assessments by their expectations or by other factors[404] Also called ascertainment bias or detection bias,[150] blinding is used in RCTs to control for this bias.[404]

Curse of knowledge

Occurs when someone with substantial background knowledge of a subject does not realise that those they are communicating with do not have the same background knowledge or depth of understanding. In other words, it is often difficult to imagine what other people don't know that you know. It can occur in verbal or written communication with examples including technical jargon; uncommon words; examples from the history of the subject; or the deep understanding of a complex concept that comes from greater experience thinking about it.[131] This may partly explain the lack of success of some statistical training which is often taught by people with advanced mathematical training and experience to people with limited mathematical training and experience.

Groupthink

Where a leader, or the 'ingroup' of a group, encourages consensus instead of seeking alternative viewpoints, leading to poor decisions going unchallenged. People in the group feel pressure to conform and become marginalised if they disagree. Tends to be driven by overconfidence and reputational concerns.[554]

Halo effect

If a person is considered to have good or bad characteristics, like their social status, physical attractiveness, or publication record, then anything new that relates to them, regardless of whether it relates to their known characteristics, like an opinion they express, or work they have just finished, is assumed to be similarly good or bad.[555]

4.11 Getting the most out of research

Understanding that we each rely on heuristics when we plan research, analyse data, interpret the results and communicate them, mostly without conscious awareness that we do so, may lead to better research by providing motivation to overcome some of the current obstacles. One obstacle is the high cognitive demand required for valid causal inference, or inference that is reasonably adjusted for controllable biases; with the remaining uncertainty understood and properly conveyed. But tools and techniques exist that can reduce the effort required to achieve these goals. Causal diagrams can trigger more information gathering than would otherwise occur, making it easier to identify variables that should be considered.[86] This, in turn, can stimulate consideration of alternative explanations, which can overcome the biases of some other heuristics, like the bias to find causal explanations, myside bias, and overconfidence bias.

Another strategy might be the deliberate creation of lists, such as one containing alternative explanations for the associations observed or the lack of associations expected. This may lead to a better sense of the uncertainty that remains.

To make causal diagrams easier to use, some possibilities include using variable names instead of letters, because abstract symbols require greater use of working memory;[556] including causal mechanisms adjacent to arrows if the mechanisms are not very obvious, and indicating the expected or observed effect that a change in one variable has on another, similar to the 'signed DAGs' of VanderWeele and Robins (2010).[557] If valid causal inference can be made easier to achieve, it will be more likely to occur.

Chapter 5
A review of the statistical methods and related tools recently used in health intervention cohort studies

5.1 Introduction

In Chapters 1-4, we examined concepts and factors commonly involved when causal inferences are made generally, as well as more specifically in health intervention research. An understanding of these elements can help the statistician or researcher make better decisions during the research process. While an important task of methodological research is to identify bias that research might fall victim to, another is to identify the practices currently being used to avoid those biases, to better inform efforts to improve the quality of research. In this chapter and the next, we present the results of a methodological review of health intervention cohort studies with the dual aims of describing the analysis methodology used to avoid bias, and whether their use affects the strength of causal inference, as expressed in study conclusions. This chapter will focus on the first aim, while Chapter 6 will assess the relationship between the methods used and the strength of causal inference communicated.

A research project can be divided into five stages: the design, collection of data, data analysis, interpretation of results, and communication of the research. It is perhaps the data analysis where decisions made by a statistician or researcher can have the most influence on beliefs held about an intervention after the research has concluded.

Health interventions, such as drugs, tests, and exercise programs, work when they cause an improvement in health. But as discussed in Chapter 1, even when we feel confident that an intervention will work for some, causality can never be established beyond doubt because, however bizarre or unlikely, a plausible alternative explanation for observed associations will always be conceivable. Hence, with uncertainty ever present to some degree, a causal inference can be broadly defined as a conclusion that the evidence available supports either the existence, or the non-existence, of a causal effect.

Research that compares one intervention with another, or with a placebo or usual care, such as randomised controlled trials and many cohort studies, all ask the causal question: does intervention X produce better outcomes on average than no intervention, or intervention Y? Research conclusions that are communicated when such studies are published will generally address this causal question, and hence, can be considered a type of causal inference.

As outlined in section 4.5, many decisions are made in the course of analysing data that can potentially alter the result in a meaningful way, like concluding that an intervention is effective instead of ineffective, or that it alters the outcome by a large amount, on average, instead of a small amount. These decisions include the choice of statistical framework (e.g., frequentist or Bayesian),[558] the types of models and estimation methods employed,[559–564] the choice of software,[565–567] and whether to perform a sensitivity analysis.[568] The range of methods, and the variations within methods, that are available for each research design is considerable and can lead to different conclusions,[442–444] with new or improved methods becoming available on a regular basis.[440]

Some of the methods developed or introduced into health research in recent decades and associated explicitly with causal inference include propensity score methods, instrumental variables and marginal structural models, and they can sometimes remove bias more successfully than traditional regression modelling.[569–575] Other techniques that can assess potential bias and improve causal inferences are causal diagrams (usually directed acyclic graphs),[59] sensitivity analysis[576,577] and a more recent variant of sensitivity analysis sometimes called quantitative bias analysis.[578–580] However, the proportion of studies that use each method type is not clear, with reviews of statistical methodology more common for randomised trials than for observational studies.[440] With observational studies at a greater risk of bias than randomised trials, we decided to focus this review on cohort studies, the type of design that methods using propensity scores are most often used.[581] Evidence suggests that the use of some of the newer methods may be increasing,[581–588] so the review in this chapter aims to provide an update, or some insight, on how often specific statistical methods and related tools are used to analyse data leading to causal inferences in recent, mid to large scale, health intervention cohort studies.

5.2 Methods

5.2.1 Study criteria and selection

The review commenced in mid-2015 with the broad aim of cataloguing the statistical and related methods used in recently published health intervention cohort studies, excluding small studies likely to be underpowered. To enable the review to be carried out successfully by one person, trial and error was used to develop specific criteria with an approximate target of retrieving 1,000 to 2,000 articles for initial screening. In tandem, a PubMed (www.ncbi.nlm.nih.gov/pubmed) query was developed to retrieve the list of articles, with trial and error used to identify PubMed search terms that increased the chance of retrieving studies meeting possible criteria, such as a specific date range, while decreasing the chance of other studies, reviews or commentaries that would need to be manually excluded.

The original definition of 'health intervention' that we used in 2015 was:

> Any type of treatment, preventive care, or test that a person could take or undergo to improve health or to help with a particular problem. Health care interventions include drugs (either prescription drugs or drugs that can be bought without a prescription), foods, supplements (such as vitamins), vaccinations, screening tests (to rule out a certain disease), exercises (to improve fitness), hospital treatment, and certain kinds of care (such as physical therapy).

This definition was obtained from the website http://effectivehealthcare.ahrq.gov although we note it is now no longer available. It is similar to the following definition of a 'health intervention': "an act performed for, with or on behalf of a person or population whose purpose is to assess, improve, maintain, promote or modify health, functioning or health conditions".[55]

The final full criteria for each included article was as follows:

1. assessed a health intervention that was deliberately prescribed or utilised, with the sole aim of improving the health of human study subjects
2. measured some aspect of human health as the outcome
3. was a cohort study
4. was comparative, with at least two interventions compared or an intervention compared to no intervention, and at least two separate patient groups
5. at least 200 people were included in the final analysis
   - ↳ while a somewhat arbitrary number, our focus was on methods to control for confounding and selection bias, so we wanted a majority of the studies to be ones where random error would not be the greater concern
6. was published in a 2014 volume/issue of a journal (excluding 'Articles in Press'/'Early View' etc) that was
   1. in any 2013 Journal Citation Reports (JCR) clinical medicine category
   2. with a 2013 JCR impact factor of at least 4.000
   3. in the top 10 journals by impact factor of any clinical medicine category

5.2.2 Screening and data extraction

Articles returned by PubMed were manually screened using EndNote,[589] often by reading the title, though sometimes with a need to view the abstract or full text. Excluded articles were categorised by the first reason apparent for exclusion. A custom database and data entry application was developed using Microsoft SQL Server[590] and Microsoft Visual Studio,[591] where each article's information was recorded and viewed separately in a form, and data such as methods used could easily and accurately be counted afterwards using SQL database queries. The information extracted from each article included reference details, author country, interventions and their type, outcome and their type, characteristics and size of the study population, cohort study type (retrospective or prospective), statistical and related methods used for the primary analysis, language used to describe the methodology, and statistical software used because the use of statistical methods depends heavily on software capabilities and ease of use.

To offset potential errors with one person extracting the data, a full-text search program (FileLocator Pro[592]) was used to search for approximately 100 keywords, using regular expressions, that might identify statistical method related information. If a word was found in an article, this was highlighted in the data entry application, ensuring that the keyword was not overlooked in the text. Sometimes, however, a keyword found in an article might only be mentioned in places like the Discussion or References, and not relate to the analysis methodology of that study. But when a particular keyword was found to relate to the primary analysis, for example, the word 'logistic' was found in an article that used logistic regression in the analysis, then this positive match was recorded for later use (see section 5.2.3).

Given the number and variety of statistical methods available, no widely agreed-upon way of grouping or classifying methods exists. Nevertheless, the methods extracted from the articles were grouped together under commonly used headings. An exploratory analysis followed that looked at different study features and their possible relationship with the methods used to reduce biased results.

5.2.3 Automated full-text search to assess secular changes

When all the reviews had been completed, search keywords found to be related to the primary analysis in 90% or more of the articles the word was found in, were designated a 'reliable' word for use in automated full-text search of a wider selection of articles from later years. To look for changes in the use of methods since 2014, the same PubMed query was submitted again in May 2017 for the publication years 2014 and 2015, and again in Oct 2018 for the publication years 2016 and 2017. Once articles had been retrieved, and without manual screening, relevant keywords that were considered 'reliable' or, on the other hand, were rarely or never found, but nevertheless related to a method of analysis, were again used in an automatic full-text search using regular expressions (see Appendix A.2 for a list of expressions used). The results were then compared between the years.

We used Stata 15.1[593] for the construction of graphs and conducting of chi-squared tests.

5.3 Results

5.3.1 Sample selection

The final PubMed query that was used to identify articles is shown below, though with the part containing journal titles considerably abbreviated (the full query is contained in Appendix A.1):

```
2014[dp] AND humans[mh]
AND
(cohort[tiab] OR cohorts[tiab] OR cohort studies[mh] OR cross-over studies[mh] OR
follow-up[tiab] OR follow-up studies[mh] OR followup[tiab] OR longitudinal[tiab]
OR observational studies[tiab] OR observational study[pt] OR observational
study[tiab])
AND
(before and after[tiab] OR comparative study[pt] OR compared[tiab] OR
comparison[tiab] OR comparative[tiab] OR versus[tiab])
AND
("Acta Derm Venereol"[ta] OR "Acta Neuropathol"[ta] ...)
NOT
(2013[ppdat] OR 2015[ppdat] OR case series[tiab] OR cross-sectional studies[mh] OR
diagnosis[sh] OR economics[sh] OR genetics[sh] OR meta-analysis[pt] OR
prevalence[mh] OR randomised[tiab] OR randomized[tiab] OR randomized controlled
trial[pt] OR randomly[tiab] OR review[pt] OR systematic[sb])
```

In June 2015, an initial sample of 1,871 references was retrieved from the PubMed website using the query above. Table 5.1 lists the JCR Journal Categories contained in the query and the associated number of journals and references returned by PubMed. While not restricted to any clinical area, one consequence of using JCR impact factors to target the most widely read studies was that some medical fields were likely to be represented more so than others.

Subsequent screening of these articles led to a final sample of 288 studies (Figure 5.1), followed by a detailed full-text review.

5.3 Results

Figure 5.1 Flow diagram of the selection process

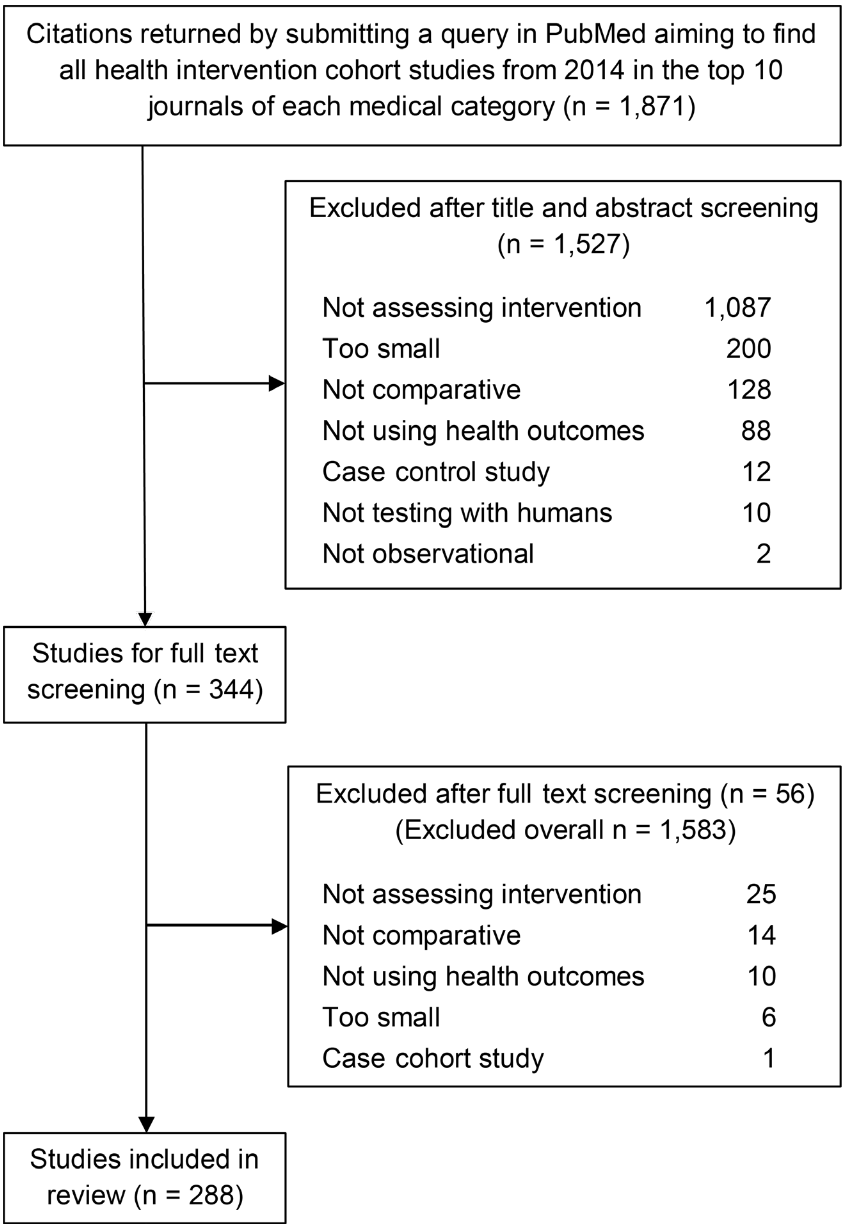

Citations returned by submitting a query in PubMed aiming to find all health intervention cohort studies from 2014 in the top 10 journals of each medical category (n = 1,871)
Excluded after title and abstract screening (n = 1,527)
Not assessing intervention 1,087
Too small 200
Not comparative 128
Not using health outcomes 88
Case control study 12
Not testing with humans 10
Not observational 2
Studies for full text screening (n = 344)
Excluded after full text screening (n = 56) (Excluded overall n = 1,583)
Not assessing intervention 25
Not comparative 14
Not using health outcomes 10
Too small 6
Case cohort study 1
Studies included in review (n = 288)

Table 5.1 JCR Journal Categories

Of 247 journals in the PubMed search, 26 were in 2 categories and 2 journals in 3 categories

	Number of articles*		Number of journals*		
JCR (2013) Journal Category	Returned by PubMed	After screening	In PubMed query	After PubMed	After screening
Allergy	13	1 (8%)	4	3	1
Anesthesiology	29	5 (17%)	3	3	2
Cardiac & Cardiovascular Systems	101	17 (17%)	10	9	6
Clinical Neurology	49	3 (6%)	10	4	2
Critical Care Medicine	115	15 (13%)	5	5	4
Dentistry, Oral Surgery & Medicine	3	2 (67%)	2	1	1
Dermatology	27	2 (7%)	7	4	2
Emergency Medicine	10	3 (30%)	1	1	1
Endocrinology & Metabolism	31	3 (10%)	10	2	1
Gastroenterology & Hepatology	104	17 (16%)	10	9	7
Geriatrics & Gerontology	66	6 (9%)	6	3	2
Health Care Sciences & Services	15	0	5	2	0
Hematology	51	10 (20%)	10	8	4
Immunology	29	6 (21%)	10	2	2
Infectious Diseases	118	15 (13%)	10	8	6
Medical Informatics	6	0	1	1	0
Medicine, General & Internal	97	28 (29%)	10	7	5
Medicine, Research & Experimental	4	0	10	3	0
Nutrition & Dietetics	84	2 (2%)	10	3	1
Obstetrics & Gynecology	121	35 (29%)	4	3	3
Oncology	35	10 (29%)	10	5	5
Ophthalmology	31	2 (6%)	5	3	1
Orthopedics	80	9 (11%)	3	3	2
Pathology	1	0	10	1	0
Pediatrics	48	10 (21%)	3	1	1
Peripheral Vascular Disease	130	19 (15%)	10	7	4
Pharmacology & Pharmacy	2	1 (50%)	10	2	1
Primary Health Care	1	1 (100%)	1	1	1
Psychiatry	25	1 (4%)	10	8	1
Public, Environ. & Occup. Health	63	2 (3%)	10	6	2
Radiology, Nuc. Med. & Med. Imag.	80	8 (10%)	10	9	4
Rehabilitation	1	1 (100%)	1	1	1
Respiratory System	93	13 (14%)	8	7	4
Rheumatology	81	13 (16%)	8	5	4
Sport Sciences	57	5 (9%)	6	3	2
Substance Abuse	9	0	2	2	0
Surgery	233	42 (18%)	10	10	8
Toxicology	11	0	10	3	0
Transplantation	37	5 (14%)	3	2	2
Tropical Medicine	5	1 (20%)	1	1	1
Urology & Nephrology	69	16 (23%)	8	6	4

* Some journals and hence references in more than one category

5.3.2 Handling of missing data

Of the 288 studies, Table 5.2 lists the number of articles that reported specific methods for handling missing data, with the remaining articles not reporting how they dealt with this problem.

Table 5.2 Missing data methods that were reported

Method to handle missing data (as described in article)		Articles
Multiple imputation		21
↳ multiple imputation	15	
multiple imputation using flexible additive imputation models	1	
multiple imputation using Markov chain Monte Carlo method	1	
multiple imputation using sequential regression models	1	
multiple imputation using the chained equations method	2	
multiple imputation via prediction mean matching	1	
Excluded people with missing data		22
Imputation using last observation carried forward		2
Imputation using linear interpolation		1
Imputation using means, medians and/or modes		3
Mid-point imputation		1
Missing indicator method		1

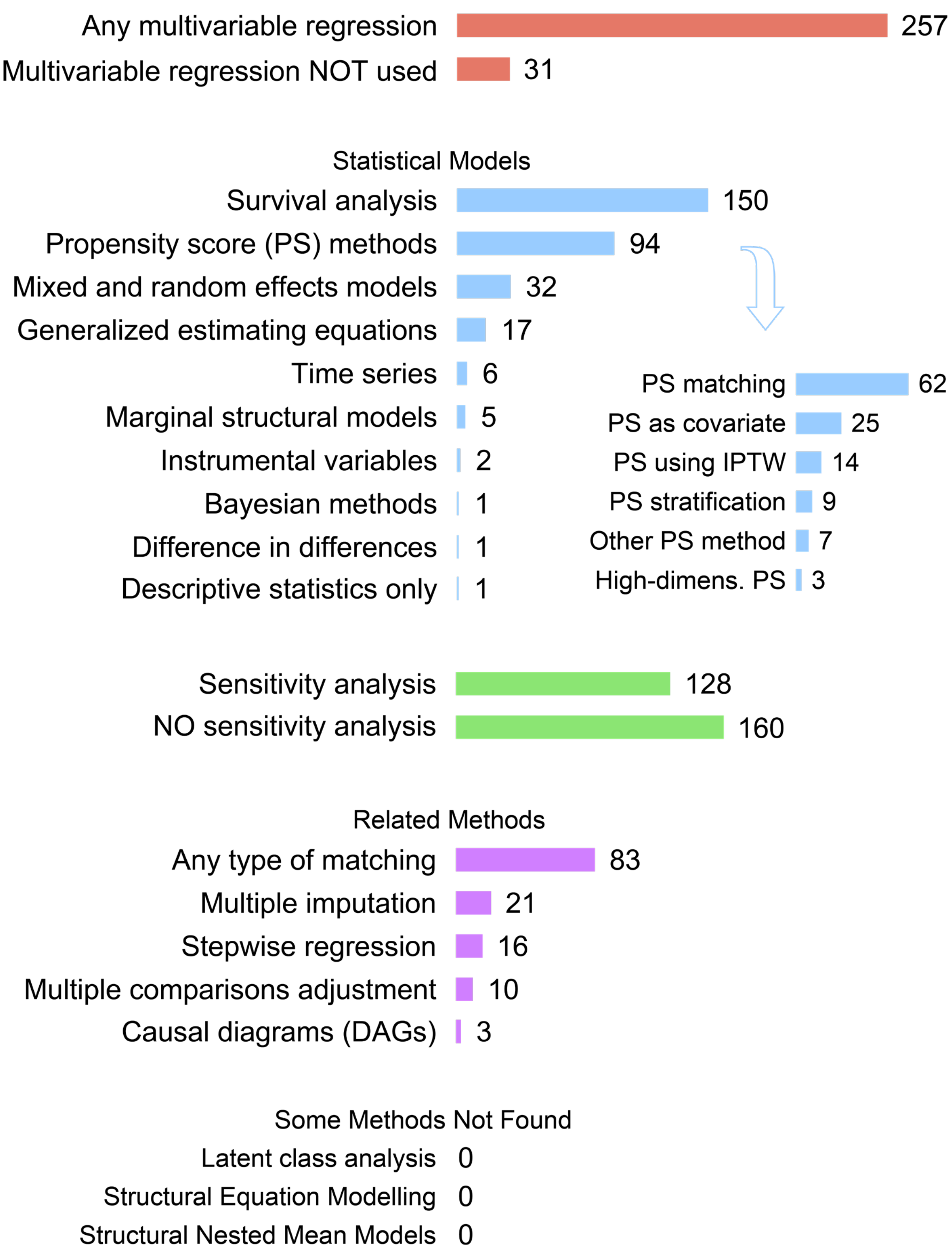
5.3.3 Statistical methods used

There was considerable variation in the types of statistical methods used, however, familiar categories could still be used to group them together. A large majority of articles used at least one multivariable regression model (Figure 5.2). Note that articles often used more than one method and all methods were included in multiple categories.

Aside from propensity score methods, found in 94 (33%) studies, use of methods associated explicitly with causal inference in the literature was uncommon, with 5 using marginal structural models, 3 using causal diagrams and 2 using instrumental variables (Figure 5.2).

Figure 5.2 Number of studies using each method type

All articles and methods were in more than one category; total studies = 288



More generally, of the more advanced regression methods often used implicitly for causal inference, various forms of survival analysis dominated, used in just over half of the articles, and mixed and random effects models were used in twice as many articles as generalised estimating equations. The specific methods reported and grouped in Figure 5.2 as 'Any multivariable regression', 'Multivariable regression NOT used', and 'Propensity score (PS) methods' are listed in Appendix A.3. Also listed are the multivariable methods in articles that used, or did not use, a 'Propensity score method' or a 'Sensitivity analysis', to provide more detail on the overlap of these categories with 'Any multivariable regression'.

Based on the data collected, three broad comparative groupings identified are:

1. Use or non-use of a multivariable regression method

    ↳ adjustment for confounding and selection bias in observational studies cannot be done with methods such as t-tests or chi-squared tests, and stratification is generally limited to a small number of confounders

2. Use or non-use of a propensity score method

    ↳ because there is an explicit association of these methods with causal inference in the literature, and there has also been a rapid rise in the popularity of these methods in the last two decades that some believe may give users a false sense of security about their control of bias[594–596]

    ↳ all articles in this group also used multivariable regression

3. Use or non-use of a sensitivity analysis

    ↳ while the use of this term will vary, any form of sensitivity analysis, if done properly, is likely to reduce the chance of a biased result or interpretation

    ↳ most articles in this group also used multivariable regression (see Table 5.3)

Table 5.3 Articles using multivariable methods and a sensitivity analysis

	No Multivariable regression	Multivariable regression	Total
No Sensitivity analysis	28	132	160 (56%)
Sensitivity analysis	3	125	128 (44%)
Total	31 (11%)	257 (89%)	288

All 3 articles that claimed to conduct a ‘sensitivity analysis’ (Table 5.3), yet did not use a multivariable method, were vaccine studies.

5.3.4 Author location, journal and study size

Comparing the location of the authors with the three method groupings singled out above (Table 5.4), and ignoring the relatively small number of articles from Asia-based authors and the heterogenous other continent (e.g. Africa or Australia) or multiple continent locations, the most obvious feature is that propensity score methods appear to be more commonly used by authors in the United States and Canada than in European countries.

Table 5.4 Article numbers by author location and methods used

Percentages relate to row N and values referred to in the text are highlighted in magenta

Author Base	N (288)	Multivariable regression	Propensity score method	Sensitivity analysis
North America	132	123 (93%)	52 (39%)	64 (48%)
Europe	82	69 (84%)	20 (24%)	36 (44%)
Asia	26	24 (92%)	10 (38%)	5 (19%)
Other or Multiple	48	41 (85%)	12 (25%)	23 (48%)

Comparing the use of methods between JCR journal categories (Table 5.5), all three of the categories with the lowest use of multivariable methods are also three of the four journal categories with the smallest mean sample size (Table 5.6). The most obvious feature of propensity score use is the much higher proportion of cardiac and cardiovascular systems journal articles that used these methods, though the small N makes a chance result very plausible. In the last column sensitivity analysis was most commonly performed in Medicine, General & Internal as well as Urology & Nephrology journal articles, while Obstetrics & Gynecology had the lowest proportion of articles.

Table 5.5 Article numbers by journal category and methods used

Percentages relate to row N and values referred to in the text are highlighted in magenta

JCR Journal Category	N (329†)	Multivariable regression	Propensity score method	Sensitivity analysis
Cardiac & Cardiovascular Systems	17	17 (100%)	13 (76%)	10 (59%)
Critical Care Medicine	15	15 (100%)	7 (47%)	8 (53%)
Gastroenterology & Hepatology	17	14 (82%)	5 (29%)	6 (35%)
Infectious Diseases	15	14 (93%)	5 (33%)	8 (53%)
Medicine, General & Internal	28	26 (93%)	15 (54%)	24 (86%)
Obstetrics & Gynecology	35	25 (71%)	0	6 (17%)
Other categories	125	116 (93%)	38 (30%)	48 (38%)
Peripheral Vascular Disease	19	19 (100%)	9 (47%)	7 (37%)
Surgery	42	35 (83%)	11 (26%)	15 (36%)
Urology & Nephrology	16	16 (100%)	8 (50%)	13 (81%)

† Some journals were in multiple categories

Consistent with this observation, studies containing less than 2,000 subjects had a lower proportion of studies using any of these three methods of analysis (Table 5.7).

Additionally, Table 5.8 suggests that no meaningful difference exists between North American and European studies in terms of the general distribution of study sizes.

Table 5.6 Sample size statistics by journal category

		Sample Size			
JCR Journal Category	**N**	**Mean**	**Median**	**Min**	**Max**
Cardiac & Cardiovascular Systems	17	82,767	5,203	380	725,680
Critical Care Medicine	15	56,544	3,163	402	471,062
Gastroenterology & Hepatology	17	39,200	835	204	550,696
Infectious Diseases	15	318,272	3,990	200	4,231,923
Medicine, General & Internal	28	337,176	37,730	1,838	5,104,594
Obstetrics & Gynecology	35	85,191	3,159	252	1,047,644
Other categories	125	142,079	4,120	207	10,912,834
Peripheral Vascular Disease	19	109,678	4,989	281	985,569
Surgery	42	22,745	1,687	215	371,527
Urology & Nephrology	16	32,578	7,402	361	183,842
Total	*329*†				

† Some journals were in multiple categories

Table 5.7 Study size and methods used

Percentages relate to row N and values referred to in the text are highlighted in magenta

Study Total Subjects	N (288)	Multivariable regression	Propensity score method	Sensitivity analysis
200 - 1,999	111	92 (83%)	25 (23%)	30 (27%)
2,000 - 19,999	93	87 (94%)	38 (41%)	47 (51%)
20,000 - 10,912,834	84	78 (93%)	31 (37%)	51 (61%)

Table 5.8 Study size and author location

Percentages relate to column N

Study Total Subjects	N (288)	North America	Europe	Asia	Multiple	Other
200 - 1,999	111	49 (37%)	31 (38%)	12 (46%)	17 (39%)	2 (50%)
2,000 - 19,999	93	43 (33%)	25 (30%)	10 (38%)	13 (30%)	2 (50%)
20,000 - 10,912,834	84	40 (30%)	26 (32%)	4 (15%)	14 (32%)	0

5.3.5 Type of comparison group and intervention type

While the use of an inactive comparison group (a group receiving no intervention, or usual care interventions common to all groups) (Table 5.9) was found to be a little more likely for studies using propensity score methods, the difference is small and may not be meaningful.

Putting aside the small numbers using radiation therapy, interventions classed as assisted reproductive technology (interventions designed to achieve pregnancy), and vaccine studies, were both less likely to have used multivariable regression, and likewise for propensity score methods and sensitivity analyses, although numbers were small. At the other end of the scale, studies investigating drugs or a mix of interventions (e.g., drug with surgery compared with drug alone or surgery alone) were more likely to have used multivariable regression, propensity score methods and a sensitivity analysis. And while surgery studies were near the middle in their use of multivariable methods, they were relatively less likely to have used propensity score methods.

Table 5.9 Comparison group type, type of intervention and methods used

Percentages relate to row N and values referred to in the text are highlighted in magenta

	N (288)	Multivariable regression	Propensity score method	Sensitivity analysis
Comparison group type				
Active intervention	118	104 (88%)	31 (26%)	50 (42%)
No intervention/Usual care	170	153 (90%)	63 (37%)	78 (46%)
Intervention Type				
Assisted reproductive tech.	19	12 (63%)	0	2 (11%)
Drug	120	113 (94%)	55 (46%)	64 (53%)
Mix	15	14 (93%)	7 (47%)	8 (53%)
Other	56	50 (89%)	17 (30%)	25 (45%)
Radiation therapy	6	6 (100%)	1 (17%)	4 (67%)
Surgery	60	53 (88%)	12 (20%)	19 (32%)
Vaccine	12	9 (75%)	2 (17%)	6 (50%)

To get a more detailed picture of the connections between the data, the type of intervention investigated was compared with author location (Table 5.10) and this suggests that the reason a higher proportion of total drug research takes place in North America is simply because more health intervention cohort studies took place there, at least in this sample.

Table 5.10 Type of intervention and author location

Percentages relate to row N and values referred to in the text are highlighted in magenta

Intervention Type	N	North America	Europe	Asia	Other or Multiple
Assisted reprod. tech.	19	3 (16%)	10 (53%)	3 (16%)	3 (16%)
Drug	120	52 (43%)	33 (28%)	13 (11%)	22 (18%)
Mix	15	6 (40%)	5 (33%)	2 (13%)	2 (13%)
Other	56	34 (61%)	10 (18%)	3 (5%)	9 (16%)
Radiation therapy	6	3 (50%)	1 (17%)	0	2 (33%)
Surgery	60	30 (50%)	16 (27%)	4 (7%)	10 (17%)
Vaccine	12	4 (33%)	7 (58%)	1 (8%)	0
Total	288	132 (46%)	82 (28%)	26 (9%)	48 (17%)

5.3.6 Statistical software use

The software packages R and SAS each had the highest proportion of studies also using multivariable regression, propensity score methods or sensitivity analyses (Table 5.11). Stata had the third highest proportion using each of the three methods, while the other major package SPSS, had the lowest proportion of the major packages.

Table 5.11 Statistical software and methods used

Percentages relate to row N and values referred to in the text are highlighted in magenta

Software	N (333†)	Multivariable regression	Propensity score method	Sensitivity analysis
Not Specified	42	32 (76%)	12 (29%)	17 (40%)
Other‡	20	15 (75%)	4 (20%)	7 (35%)
R[597]	35	35 (100%)	17 (49%)	18 (51%)
SAS[598]	110	109 (99%)	50 (45%)	66 (60%)
SPSS[599]	70	59 (84%)	13 (19%)	20 (29%)
Stata[600]	56	51 (91%)	21 (38%)	28 (50%)

† Some articles used more than one software package;

‡ Included: JMP,[601] Microsoft Excel,[602] GraphPad Prism,[603] and 14 others;

Table 5.12 Software by author location

Percentages relate to row N and values referred to in the text are highlighted in magenta

Author continent	N	SAS	SPSS	Stata	R	Other	Not spec.
North America	152	63 (41%)	15 (10%)	25 (16%)	17 (11%)	9 (6%)	23 (15%)
Europe	98	14 (14%)	34 (35%)	23 (23%)	12 (12%)	6 (6%)	9 (9%)
Asia	33	13 (39%)	12 (36%)	3 (9%)	3 (9%)	1 (3%)	1 (3%)
Other / Multiple	50	20 (40%)	9 (18%)	5 (10%)	3 (6%)	4 (8%)	9 (18%)
Total	333†	110 (33%)	70 (21%)	56 (17%)	35 (11%)	20 (6%)	42 (13%)

† Some articles used more than one software package

Among the cohort studies in this sample, SAS was much more commonly used in North American studies than in European ones, whereas the opposite was true of SPSS and Stata.

The type of software used by journal category can be seen in Appendix A.4.

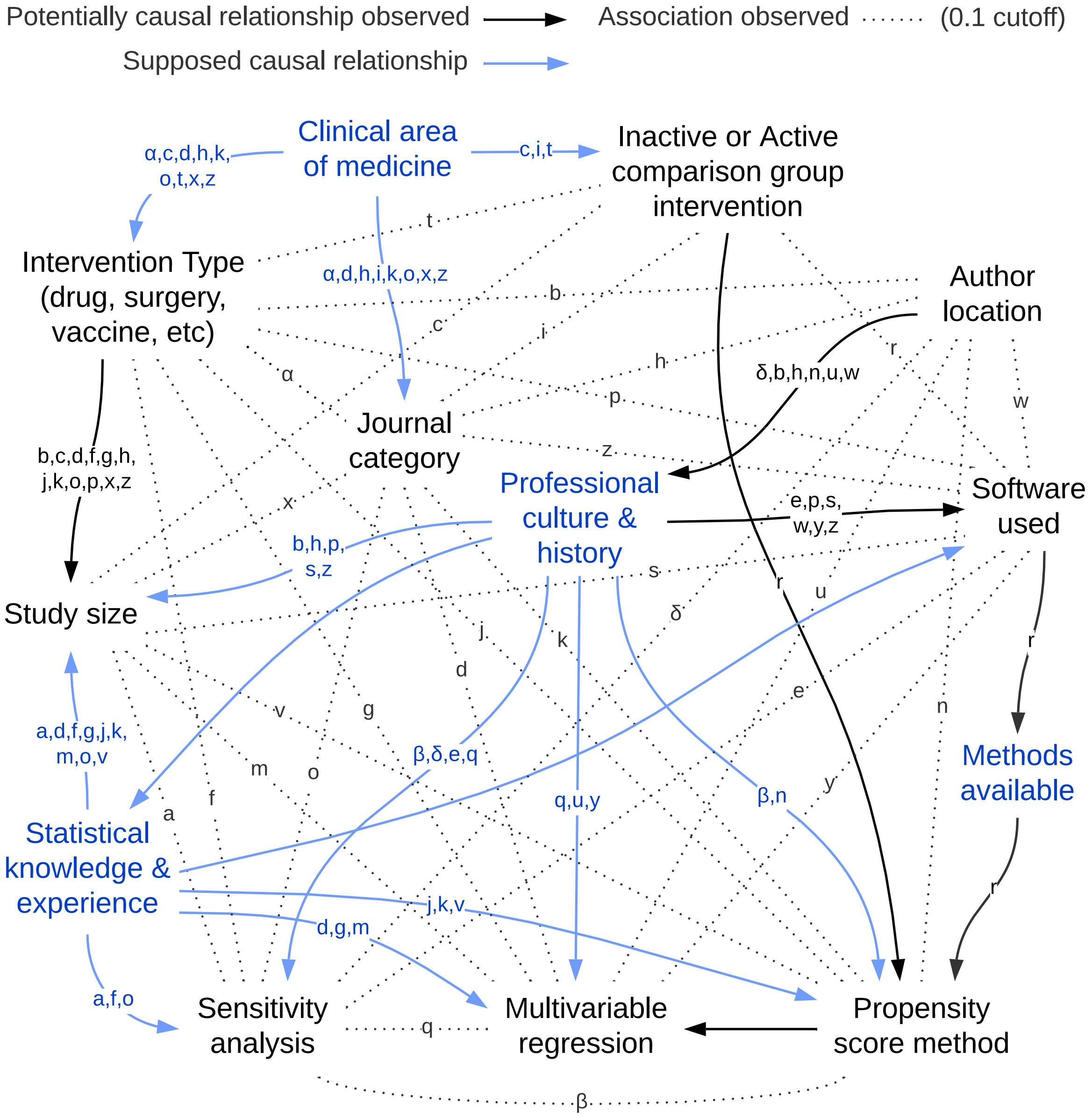
5.3.7 Explanations for observed associations

As was made clear in Chapter 4, people often try to explain the associations that they observe. And in reality, if we seek to better understand the factors underlying a phenomenon such as the choice of methods when analysing data, the only reason to test for associations is to suggest possible causes. Thus, taking an exploratory data analysis approach,[604] we created a causal diagram (Figure 5.3) to see if it could assist in understanding the multiple interconnected relationships suggested by the associations.

In the diagram, the blue variables are theorised common causes or mediators of the observed associations, and the black variables are those we analysed. Using a 0.1 p-value cutoff, each observed association is drawn as a dotted line and labelled with a letter, except for those associations considered causal which were drawn as black arrows. As discussed in section 3.3, an association between two variables is assumed to exist if a path can be traced along some of the arrows in the diagram, regardless of the direction of the arrows. As a way of checking that all of the associations have a possible explanation, and also, of illustrating this principle, the causal path for each association is labelled on each connecting arrow. Most are explained by common causes, but a few rely on a collider structure, such as ‘h’ and ‘r’.

Drawing a causal diagram that starts with the associations allows for a more detailed and considered causal structure to be suggested. It is important to note, however, that the diagram contains only one set of plausible relationships, both causal and associational, where the latter could have arisen only by chance. Hence, it is recommended that more than one such diagram be created to help prevent such overconfidence in the initial causal explanations that come to mind. For example, a new variable could be introduced such as ‘Location of influential propensity score method developers’, which may help to explain the increased use of propensity score methods found in articles by North American authors. Alternatively, a causal diagram could be drawn that includes an arrow from ‘Professional culture & history’ to ‘Inactive or Active comparison group intervention’. This is plausible, though it would increase the apparent problem that no association was found between ‘Author location’ and ‘Inactive or Active comparison group intervention’. Such an association is already a problem, however, with a path connecting the two variables able to be traced in the diagram of Figure 5.3.

Figure 5.3 Causal diagram containing one possible explanation for observed associations



5.3.8 Temporal trends

Full-text search results of articles from 2014-2017 (Table 5.13) suggest a possible small increase in propensity score methods, including words often associated with their use (though not exclusively), such as 'balance' and 'standardised difference'. A small increasing trend is also suggested for multiple imputation and sensitivity analysis. Other methods show no obvious trend or were found in small numbers only.

In terms of statistical software packages, the use of SAS remained steady and the most used, whereas Stata and R showed an increasing trend and SPSS may be slowly decreasing.

Table 5.13 Full-text search of words or terms in articles identified with PubMed query

Percentages relate to column N and possible trends are highlighted in magenta

Words or combinations searched for in article full text	2014* (N=2617)	2015* (N=2563)	2016† (N=2342)	2017† (N=2084)	Trend‡
propensity score	180 (6.9%)	210 (8.2%)	207 (8.8%)	211 (10.1%)	0.000
propensity score and					
matching	127 (4.9%)	153 (6.0%)	147 (6.3%)	167 (8.0%)	0.000
greedy matching	31 (1.2%)	30 (1.2%)	26 (1.1%)	25 (1.2%)	0.973
nearest neighbour matching	27 (1.0%)	42 (1.6%)	36 (1.5%)	39 (1.9%)	0.029
high dimensional	8	7	5	2	
inverse probability weighting	20 (0.8%)	27 (1.1%)	35 (1.5%)	36 (1.7%)	0.001
balance	69 (2.6%)	83 (3.2%)	91 (3.9%)	87 (4.2%)	0.002
standardised difference	37 (1.4%)	53 (2.1%)	57 (2.4%)	56 (2.7%)	0.001
marginal structural model	15 (0.6%)	11 (0.4%)	21 (0.9%)	13 (0.6%)	0.368
g-formula	0	1	3	4	
g estimation	1	0	1	0	
instrumental variable	12 (0.5%)	12 (0.5%)	14 (0.6%)	18 (0.9%)	0.063
latent class	3	5	15	10	
structural equation	8	10	5	4	
difference in difference	9	7	10	6	
multiple imputation	119 (4.5%)	111 (4.3%)	123 (5.3%)	122 (5.9%)	0.018
sensitivity analysis	557 (21%)	546 (21%)	593 (25%)	609 (29%)	0.000
directed acyclic graph	17 (0.6%)	13 (0.5%)	23 (1.0%)	13 (0.6%)	0.557
machine learning	3	10	9	10	
Bayesian	43 (1.6%)	45 (1.8%)	56 (2.4%)	44 (2.1%)	0.098
stepwise	227 (8.7%)	245 (9.6%)	201 (8.6%)	168 (8.1%)	0.314
SAS	781 (30%)	751 (29%)	718 (31%)	634 (30%)	0.466
Stata	343 (13%)	352 (14%)	378 (16%)	327 (16%)	0.002
SPSS	655 (25%)	622 (24%)	524 (22%)	461 (22%)	0.006
R	150 (5.7%)	186 (7.3%)	186 (7.9%)	239 (11.5%)	0.000
bias analysis	2	1	4	7	
alternative explanation	27 (1.0%)	29 (1.1%)	33 (1.4%)	20 (1.0%)	0.874
significant(ly)	1214 (46%)	1187 (46%)	1100 (47%)	974 (47%)	0.705
References from PubMed	2747	2739	2882	2664	
Articles for full-text search§	2617 (95%)	2563 (94%)	2342 (81%)	2084 (78%)	

* Articles from PubMed query submitted in May 2017; † PubMed query submitted in Oct 2018; ‡ Chi-squared test for linear trend on rows with a combined count of 40 or more; § Number of articles available through open access or the University of Sydney's journal subscriptions

5.4 Discussion

In this review of 288 health intervention cohort studies, across every area of medicine, we surveyed all statistical methods and related tools to provide a snapshot of current practice in causal inference related data analysis. Aside from methods that used propensity scores, employed by a third of the articles in our sample, most statistical methods and tools associated explicitly with causal inference, such as marginal structural models, instrumental variables, and causal diagrams, remain seldom used. And while a small increasing trend was observed for propensity score methods over 2014-2017, no such trend was apparent for the seldom used methods.

The way studies handled missing data was unfortunately not explained clearly in the majority of the articles, and of those that did report it, the generally preferred method of multiple imputation[605] was performed in a similar number of studies (21/288) as those stating that they excluded subjects with missing data from the analysis (22/288), a technique also called ‘complete-case analysis’ and known to produce biased estimates where the reason for data being missing relates both to the chance of receiving the intervention and also to the outcome the subject did or would have recorded.[606] Given that complete-case analysis has been observed to be the most commonly used technique in health research,[607] it seems likely that it is under-reported in this review. However, multiple imputation showed a slight increasing trend over 2014-2017.

Another technique that showed an increasing trend was the performing of a sensitivity analysis, considered by many to be an essential tool for causal inference,[86,580] and found in 44% of the 288 studies. However, we only recorded whether a study called one of their analyses a ‘sensitivity analysis’, so the quality and relevance of these analyses was not assessed. For example, in some cases, we thought it may have been more appropriate to call an analysis a ‘subgroup analysis’, rather than a ‘sensitivity analysis’, but assessing this in detail for all 128 articles would have taken more time than we had available. There are currently few methodological reviews that focus on the quality of sensitivity analyses in health research.

Of the other methods found, survival analysis dominated with use in 52% of the studies. And despite the well-known risk of bias in observational studies,[608,609] 11% did not adjust for any confounders, relying instead on simple statistical methods such as t-tests and chi-squared tests.

Of the propensity score methods, matching is an intuitive method and was one of those suggested when propensity scores were first introduced in 1983.[610] Similar to previous reviews,[583,587,611–614] matching was the most popular propensity score method, found in 54/92 (59%) propensity score studies, with the propensity score as a covariate the second most popular, in 25/92 (27%). This is also similar to earlier studies[587,611–614] despite warnings for over 10 years that using the propensity score as a covariate can lead to biased results.[611,615] Inverse probability of treatment weighting (IPTW) was found in 14/92 (15%) studies but was a relatively unknown method in health research before 2008.[581]

In an exploratory exercise, plausible associations were examined between the methods used and study features that might in some way influence the decisions that precede a causal inference, such as factors that can make particular analysis choices more likely. One methodological grouping of interest was the use or non-use of a propensity score method, because of their explicit association in the literature with causal inference and rapid increase in uptake. The other two groupings of interest were the use or non-use of multivariable regression, and whether or not a sensitivity analysis was performed, because the use of either can reduce the chance of a biased result.

Firstly, propensity score methods were more likely to be encountered if the authors resided in North America compared to Europe (39% vs. 24%), the two locations where most authors were based. This association does not appear to be related to the types of interventions investigated, such as pharmaceutical studies, because the proportion of studies investigating drugs was about the same for both continents. Further associations between various study features led to the creation of a causal diagram to help form at least one plausible explanation (Figure 5.3). It seemed to help noticeably and thus may be a useful tool for some types of exploratory data analyses.

5.4 Discussion

One strength of this review is that at 288 articles, our sample was reasonably large; and because we searched highly ranked journals that were not restricted to specific disciplines or open access only, it is likely to be more generalisable across all of health research. One limitation might be that a single person extracted all of the data from the sample. However, this was offset by the use of full-text automatic search software, with key terms flagged when extracting from each article. Another limitation is that only cohort studies were included in the sample; however, to date, most statistical methods specifically targeting causal inference have been aimed at cohort designs.[616] And, we should note that the exploratory nature of this review has produced many comparisons, so some low p-values are likely to have occurred by chance alone.

Finally, where feasible and given the study design, the use of statistical methodology that is most suited to answering the research question—by reducing confounding and selection bias most effectively—can make accurate research conclusions more likely; though not if they are used inappropriately. Factors likely to increase uptake of advances in methodology or to promote improvements in how methods are used, such as more consistent checking of assumptions, include the appropriate use of methods by recognised opinion leaders and easily accessible and understandable statistical code.[617–619]

Chapter 6
Causal interpretation in health intervention cohort studies

6.1 Introduction

While there are many biases that can influence the generation of research evidence, many of these and more can alter the interpretation and communication of that evidence, with examples from Chapter 4 including myside bias, overconfidence bias, the Dunning-Kruger effect, and a bias for causal explanations. This chapter reviews again the sample of health intervention cohort studies from Chapter 5, but this time looking at the causal interpretation of that evidence and the communication of those inferences, with a focus on the strength of causal language in study conclusions. Our goals are first, to better understand how causal inferences are expressed in writing, and the words or grammatical features that convey their strength or certainty; and second, to explore whether the statistical methods or other study features might influence the strength of causal inference communicated by study authors.

6.1.1 Causal interpretations

A causal interpretation of the results is a causal inference, and both terms are often thought to mean, at least casually, that a black-and-white decision has been made favouring the existence of a causal effect.[620] However, research can only provide supportive evidence, with uncertainty never dispelled entirely. In the case of a comparative health intervention study, the aim is to provide evidence to help answer one, or both, of the following questions:

a) whether the intervention caused the health of subjects to **improve**, or

b) whether the intervention caused the health of subjects to **worsen**

6.1 Introduction

Using the definition of a causal inference proposed in Chapter 1: "that the evidence available supports either the existence, or the non-existence, of a causal effect", any conclusion in a health intervention study that addressed one of the two questions above could be considered a form of causal inference. An exception would be any study that concluded that their results do not suggest anything about the intervention and the outcome, as if the study had not been done; but if these studies exist, they are rare.

In addition, a conclusion that suggests a causal association does not exist can have as much of an influence over a person's use of that intervention as a conclusion that the causal association does exist. Hence, the term 'causal inference' need not be restricted to decisions that a causal relationship does exist. Note that this is not to disagree with the fact that a large p-value does not provide direct evidence of 'no effect',[294,520,621] which may have occurred, for example, from a lack of power; we only point out that conclusions of 'no effect', whether justified from a combination of evidence or not, are nevertheless inferences about a causal relationship.

Some also assume that by concluding 'an association' was found between the intervention and outcome, a causal interpretation of the data has not been made.[622] But while this is unlikely to be misinterpreted as a strong causal finding, stating that an association has been found between possible cause and effect events, can only be understood as the causal effect of either the intervention, a confounder, a collider, or a combination of the three. Combined with our preference for causal explanations (see Ch.4), use of the euphemism 'association', without reference to the causal question under study, seems unlikely to be interpreted completely non-causally.

And lastly, while evidence that is considered weak by some people may well be thought of as no evidence by others, in the absence of any other information, even weak evidence will be used for a causal inference when a decision is required.

Thus, all conclusions that address a causal question in research can be usefully thought of as a 'causal inference', even if the conclusion talks only of associations, favours no effect, or the evidence is weak. Such conclusions, however, will convey differing levels of uncertainty, and this is where the strength, or certainty, of the causal language being used, rather than whether causal language is used, may be more useful when judging whether a conclusion is appropriately worded.

The strength of a 'causal inference', the certainty of the belief that a causal effect exists, will vary from mostly uncertain to supremely confident. When these beliefs are expressed in the conclusions of research articles, the strength or confidence conveyed will likewise affect how the findings are interpreted by readers. Support for this notion comes from one study on spin,[623] though not by another.[624] However, studies in psychology also suggest that people have a confidence heuristic,[625] such that increased confidence in an author or speaker tends to be more persuasive, based on the assumption that their confidence is determined by their knowledge and the certainty that this provides. Hence, a confident study conclusion, assuming there are no apparent reasons to doubt its validity, would seem more likely to leave a reader perceiving that strong evidence had been found, hence the confident conclusion. In other words, overconfidence can be catching.

6.1.2 Causal language

There are many influences on the causal inferences researchers make and the words they choose to communicate them. Our decisions, including when we write, aim to fulfil our motives,[626] and one of the fundamental human motives that evolution left us most likely involves a desire, or drive, to increase the respect other people have for us.[627–632] In psychology, this desire is frequently merged with similar concepts into the unattractive sounding *desire for status*, defined in various ways, such as "the respect, admiration, and voluntary deference an individual is afforded by others, based on that individual's perceived instrumental social value";[627] or alternatively, "the prominence, respect, honour, and influence that individuals enjoy in the eyes of others".[633] A more commonly known motive is the concept of *self-actualization*, popularised in 1943 by Abraham Maslow as part of his "hierarchy of needs".[634] It suggests that a fundamental drive in life is to fulfil, or 'actualize', our unique full potential. However, recent evidence suggests that underlying this drive may really be the desire for status or respect.[635] And like all people, statisticians and researchers will be partly motivated by this need, along with others, such as compassion for people in need,[636] when they make and communicate a research study causal inference.

6.1 Introduction

One way the desire to increase our status can be satisfied is to make a research finding sound more important. But, at the same time, a statistician or researcher will be reluctant to risk losing the respect of their peers through a perception that they deliberately mislead readers. When an article favours the first of these competing motives, the word *spin* might sometimes be used to describe the writing. First appearing in public relations and politics in the 1980s, *spin* became "shorthand for a particular kind of political public relations, with the negative connotation of spinning a yarn – lacking truthfulness, not to be trusted, of suspect motivation".[637] As this meaning entered the general lexicon, it was taken up in health research to describe intentional, or unintentional, reporting that could distort the interpretation of study findings and mislead the reader.[21,408,623,624,638–663]

Boutron and Ravaud (2018)[640] recently catalogued a variety of forms that spin could take in biomedical articles. They include the reporting of post-hoc hypotheses as though prespecified; selective reporting of analyses that favour the investigator's hypotheses or those that display significant p-values; and biased interpretation of the results, such as ignoring regression to the mean, confounding, or overstating small study effects. An additional example of spin might involve avoiding a discussion of missing data. All of these would increase the apparent certainty of the result and thus increase its perceived importance as a scientific finding.

The location of such spin is also important. When a person makes a decision about an intervention, the evidence they weigh will often come from the conclusions of other people, be they friends, doctors, journalists, or in the case of evidence from a research article, the study investigators. However, critically reading a research article takes time and effort, and that is assuming the full-text article is available in the first place, either through open-access or by a subscription the person may use. But abstracts are always free, and in health research they also tend to be structured, making it easy to quickly absorb the content of interest. Consequently, as put succinctly by Peter Gøtzsche (2009),[664] a Cochrane collaboration co-founder: "Most users of the scientific literature read vastly more conclusions than they read abstracts, and vastly more abstracts than full papers". This is a commonly held view[665–668] and there is also some empirical support.[669,670] It may be why the abstract is where spin is most likely to be found,[639] precisely because the wording used in the abstract, and especially the study conclusion, is where authors are most likely to influence readers.

However, it seems reasonable to assume that spin is not employed consciously by many authors, although it would be hard to determine when it is. Other than from a desire to impress peers, which may not be a conscious motive when wording a conclusion, spin may also result from an inadequate understanding of methodological principles, such as failing to fully understand how missing data can produce bias, or from following a reporting practice that is commonly observed in their field.[640]

When conclusions are written with exaggerated confidence the consequences for findings that are false may be multiplied. Some even suggest that a majority of research money and researcher effort has been wasted because of false or exaggerated findings.[3,392,671,672] Thus, greater caution in the interpretation of results and conclusions has been recommended, with words implying uncertainty considered essential.[673–676] To put it another way, the uncertainty needs to be conveyed in a way that leaves the reader appropriately uncertain.

Overlapping with issues of spin are legitimate concerns about the use of "causal language", especially in non-randomised research. Overconfidence in the accuracy of results can be seen in many health research articles,[109,313,423] and is not a recent problem.[677] One response to the overstatements seen in articles, and encouraged by a causality-shy statistics profession, has been the development in research publishing and teaching of a convention where causal language is generally discouraged.[678] But an increasing focus on causal inference in recent decades has seen this practice criticised. On one side of the argument are those maintaining a preference for associational language only, to avoid overinterpretations and leave more of the inference making to the reader.[650,679–683] On the other side, some believe that the use of causal language to describe research with a causal aim, rather than leading to increased overinterpretations, will increase the chance that the statistical model used will be appropriate to the causal goal, instead of one better suited to prediction modelling, and that inferences will ultimately be less ambiguous to the reader.[90,622,678,684–693] At the heart of this issue is the meaning of the term 'causal language', and which approach will provide the least misunderstanding and the better science.

Language is often ambiguous or vague in its meaning,[132,254] with interpretations of study conclusions likely to vary, at least a little. Hence, before we judged the 'strength of causal inference' in study conclusions, we conducted a brief review of the words and grammatical features that might help determine the strength of a causal inference.

Recommendations for greater caution when interpreting research have often come from statisticians, and training in statistics has long emphasized caution regarding correlation and causation.[233,694] Anecdotally, there also appears to be a common assumption that statisticians are more cautious when inferring causality. In general, with their greater methodological knowledge, statisticians would seem more likely to use statistical methods that, at least in theory, control for more confounding and selection bias, and this can strengthen causal inference. Naturally, this is only the case if they also have a good understanding of the study design and implementation, as well as sufficient knowledge of the subject matter to enable judgements about confounding. From another perspective, it might be said that statisticians are more likely to understand when the assumptions of a methodology are not met and to be able to take advantage of their greater knowledge of alternative methods that may reduce the potential for bias, thus potentially improving the strength of the results. This suggests the possibility that statisticians might be **more** likely to infer strong causality than non-statisticians. As this does not appear to be the case, it suggests another possibility: that the use of statistical methods more capable of controlling for confounding and selection bias will, in turn, result in changes to how causal inferences are formed, such that more cautious causal language is then used. Hence, our second goal is to examine whether the statistical methods used, or other study features, might affect the 'strength of causal inference' communicated by study authors.

6.2 Methods

Chapter 5 contains details on the selection and screening of studies for this review, and the general method of extracting data from the final 288 studies.

6.2.1 Additional data extracted

For each conclusion in the abstract of every article:

- providing that it addressed the question of whether the evidence found in the study did, or did not, support the inference that the intervention had an effect on some participants' health
- the following additional information was extracted:
    - ▹ the text of each *conclusion*
    - ▹ the *outcome* that the intervention might have affected, either:
        - ↳ a health **benefit:** some aspect of improved health, or
        - ↳ a **harm** to health: an adverse effect
    - ▹ the *result* after comparing group outcomes, as determined by the study investigators, either:
        - ↳ **similar** (a null result, a result of no difference**),** suggesting the evidence found did not support the inference that the intervention caused, or caused a change in, the outcome
        - ↳ **different**, suggesting evidence was found that supported the inference that the intervention may have caused, or caused a change in, the outcome

6.2.2 Review of causal language

Before assessing the 'strength of causal inference' in study conclusions, a brief review was conducted to better understand the words, grammatical features, and word combinations, that might help convey the strength of a causal inference to the reader. For this task, we reviewed relevant literature from linguistics,[152,695–705] machine learning-based natural language processing,[706–714] health research[532,624,641,648,650,652,654,679,682,693,715] and psychology;[716,717] as well as from Wikipedia,[718,719] online dictionaries,[720–724] and other grammar related websites.[725,726] The information gathered is summarised at the beginning of the Results section.

6.2.3 'Strength of causal inference' ratings

Over the last decade, a number of health research reviews have given ratings to articles that relate in some way to strength of causal language. Of six reviews identified,[§] only one by Li et al.[715] specifically rated the 'strength of causal inference', while the others rated 'spin',[650,654] "misleading abstract conclusions",[652] 'biased research reporting',[532] or the 'use of causal language'.[679] All rated causal language according to specific criteria using two or three reviewers, with initial disagreements discussed until a consensus was reached. Ochodo et al.[654] used five additional reviewers for some articles each, and Li et al. used a panel of 34 researchers to rate the 'strength of causal inference' using a Likert scale ranging from 1 to 7, with scores then analysed in a model.

Instead of relying only on the ratings of one or more people, an initial goal was to investigate whether a more objective and repeatable method could be developed. A full-text search algorithm was designed, using SQL code and the database, that it was hoped could replace or assist with the human rating process. To reduce the number of words processed by the rating algorithm, and to avoid some possible bias when three reviewers gave subjective ratings, each conclusion was modified as follows:

1. words that did not relate to the 'strength of causal inference' were removed
2. words that described the intervention(s) became 'intX', 'intY', or 'intZ', and
3. words that described the outcome became 'outcome'

Using the understanding of causal language gained from experience and the review (section 6.2.2), we extracted the words or word combinations (abbreviation: words/combinations) in the modified conclusions that might **imply a <u>specific</u>** 'strength of causal inference' when contained in a conclusion. At the same time, we assigned to the word/combination the specific 'strength of causal inference'; initially 'Weak', 'Moderate', or 'Strong', but later it was changed to 'Not strong' or 'Strong'. When combinations of words were thought necessary to convey the correct causal strength, they could be any number of characters, up to and

[§] Our search was not systematic or exhaustive, however, so some may have been missed including the entire modified conclusion, if required. The words/combinations were then labelled with the grammatical (parts-of-speech) categories they belonged to, focusing on those most relevant to causal inference, such as copula verb, epistemic modal verb, intensifier, comparative, and others. The database tables that stored this word/combination information, the algorithm ratings, and the human ratings, are displayed in Appendix B.1.

The automatic rating algorithm we developed was:

- For <u>each</u> modified conclusion:
    - For <u>each</u> word/combination in the table, and starting with the word/combination that has the most characters:
        - ↳ Search the modified conclusion for word/combination, and if found:
            - Assign the causal strength of that word/combination to the modified conclusion, but only if
                - a. no strength is recorded for that conclusion, or
                - b. the previous highest strength recorded is 'Not strong' and the new highest strength is 'Strong'[**]
            - Delete the word/combination in the modified conclusion
        - ↳ Repeat using the word/combination in the table with the next most number of characters, until no word/combination still exists in the modified conclusion
    - Repeat using the next modified conclusion

An iterative process was used to improve the word/combination table that the algorithm relied on. This involved comparing the agreement between the rating given by the algorithm and my own rating of the conclusion, and if they differed, followed by either (a) modifying one or more of the word/combination strength ratings, (b) adding a word combination to the table with an associated causal strength that more accurately reflected the strength

[**] Slightly more complicated when the ratings were ‘Weak’, ‘Moderate’, or ‘Strong’

implied by the modified conclusion, or (c) modifying my own rating if it seemed appropriate. The process was repeated until agreement for all modified conclusions was reached.

At this point, two additional reviewers (statisticians Laurent Billot and Jannah Baker) were asked to rate the modified conclusions. Initially, only an intuitive judgement of 'Weak', 'Moderate', or 'Strong' was given. We avoided using criteria in order to gain a glimpse of how variable an article's interpretation might be when people only read the conclusion.

Disagreements between the three reviewers (T.W., L.B. and J.B.) and the automatic algorithm were then discussed a number of times, and the word/combination table continued to be improved until a consensus was reached between the ratings of the three reviewers and the automatic algorithm. During this process, the scale was changed to a two-level rating system of 'Strong' (confident) or 'Not strong' (cautious) because agreement with the three-level scale could not be reached; and for a few modified conclusions, a two against one majority was used when unanimous agreement still could not be achieved. Interrater agreement between individual ratings was determined using the intraclass correlation coefficient (ICC) and Cohen's kappa, with agreement also compared between the first rating each reviewer gave, the second after discussing the different interpretations, and the rating that resulted when each was converted to the binary scale.

Finally, each article was given a consensus rating of 'Strong' causal inference if any of the article's conclusions (or the article's single conclusion) had a rating of 'Strong', otherwise a rating of 'Not strong' causal inference was given. Associations between the 'strength of causal inference' and various study features were conducted using chi-squared tests.

We used Stata 15.1 to construct graphs and calculate statistics.

6.3 Results

6.3.1 Review of causal language

The first thing to note is that, as in all academic disciplines, opinions in linguistics can vary about the best way to categorise words and other features of grammar.[703] However, the information summarised here is not particularly controversial, though specific terminology may not be common to all. For example, the same category of word might be called a "modal verb"[698] by some and a "modal auxiliary" by another.[131]

Many grammatical categories were examined for their relevance to causality, and those found to be important are described in Table 6.1, Table 6.2 and Table 6.3.

Table 6.1 Grammatical categories of causal words and word combinations

Category	Description	Examples from articles
Verb	Loosely defined as 'doing' or 'action' words.[697]	"IntX **produced**"; "InX **conferred**"; "was **observed**"
Adverb	Adds more information about a verb, and sometimes an adjective, another adverb or a sentence.[720]	"IntY **negatively** affected the outcome"
Noun	Physical things, abstract ideas, events.[698]	"suspicion"
Adjective	An attribute of a noun.[697]	"IntX is a **viable** option"
Copula verb	Links a subject to a specified state, quality, nature, role, etc.[727] Main forms:[728] "be", "am", "is", "are", "being", "was", "were", "been". Related forms: "seem" and "appear".[152]	"IntX **was** associated with an improvement in the outcome"
Evaluative verb	Expresses the writer's attitude towards a statement that the writer accepts as true; often followed by "that".[703,729]	"showed"; "indicating"
Intensifier	Modifier of an adjective or adverb that expresses the degree to which the quality expressed by that adjective or adverb is present.[703]	"highly"; "marked"; "substantial"

Epistemic modality (or mood) refers to when words express the degree of reality of a statement; or how possible, believable, or actual it is, in the opinion of the writer.[703] Such words (Table 6.2) are often involved in expressions of causal strength.

Table 6.2 Words expressing modality

Category	Examples from review articles
Modal verb	“can”; “could”; “may”; “might”; “must”; “should”;
Modal adverb	“positively”; “possibly”
Modal adjective	“causal”

During this review and the development of the rating algorithm, it became clear that causal strength would tend to depend on word combinations instead of individual words, so categorisation of phrases and sentences were also examined. Various theories of syntax and grammar use the term *predicate* in different ways.[152,730] *Traditional grammar* refers to *predicate-subject* combinations, while *predicate calculus* (also called *predicate logic*) makes use of *predicate-argument* structures. We used predicate-argument combinations to categorise the word combinations taken from the modified conclusions, because this system appeared to be more common in articles on causal relations in English (for example[712,731,732]). Below is an example that is used to explain three terms for different word combinations:

For the sentence: **IntX is associated with the outcome**

- the predicate:
  - ○ is associated with
- the verb phrase:
  - ○ is associated with the outcome
  - ○ with the verb in blue and its dependent in gold
- the predicate-argument:
  - ○ IntX is associated with the outcome.
  - ○ with the green words representing the arguments of the predicate in red

Table 6.3 Grammatical categories of word combinations

Category	Examples from review articles
Noun phrase	“association between”
Verb phrase	“findings do not support”; “have different effects”
Predicate	“findings suggest”; “has a detrimental effect”
Predicate and argument	“a significant difference was observed”; “intX increased”

The typical aim in linguistics articles is to simply describe causal language; and in natural language processing the common goal was to find causal relations in a large body of text. Our target, instead, was to judge the strength of a causal relationship already assumed to exist in a piece of text.

The six health research reviews discussed in 6.2.3 each gave ratings to articles at least partly related to causal language. For example, Lazarus et al.[650] classified spin in abstract conclusions and they included an assessment of whether “causal language” was used, defined as the use of “modal auxiliary verbs”; “causal relationship” words like “effective”, “improve”, or “enhance”; or a tone suggesting a “strong result (e.g., “The results demonstrate” or “This study shows that”)”. They considered that causal language was not used when only a statistical association was reported.

Cofield et al.[679] likewise judged whether causal language was used, defining language implying causation as: “effect, effect modifier, modify, increase, decrease, improve, influence, impact”; and non-causal language as “associated, related, correlated, predicts”.

Li et al.[715] used a “Likert scale that ranged from 1 (the investigators inferred no causal relationship) to 7 (the investigators inferred a strong causal relationship)” in order to rate the ‘strength of causal inference’ in abstracts. However, unlike in our review, their consideration of the strength of the inference was combined with whether an effect was detected and the size of the effect. Hence, no effect was rated 1.0 regardless of whether a strong belief in this inference was expressed. In terms of language used, they provided the following examples for each rating range:

- 1.0–2.7: 'no significant change', 'unable to'; 'unsuccessful'; 'no effect'; 'no impact'
- 2.8–4.6: tentative tone with weak modal auxiliary verbs, e.g., 'may'
- 4.7–5.1: mostly tentative; terms like 'suggests', 'seems to', 'appears to be', 'is possible', 'has the potential'; strong modal auxiliary verbs like 'can'
- 5.2–5.8: infers moderate to strong causality, e.g., 'resulted in', 'demonstrates', 'was found to be', 'feel confident', 'believe the results show', 'support', 'strongly support', 'have shown', 'indicate that', 'provide strong evidence', 'constitute objective evidence'; strong modal auxiliary verbs like 'can' or no modal auxiliary verbs
- 5.9–7.0: definitive tone with terms like 'is effective', 'more efficient', 'clear evidence', 'had an impact', 'robust', 'significant', 'substantial effect'

6.3.2 'Strength of causal inference'

From 288 article abstracts, 338 distinct conclusions were identified, with 115 (40%) rated as 'Strong', as opposed to 'Not strong'. Most articles contained only one main conclusion in the abstract (Table 6.4), but those articles containing multiple primary interventions or multiple main outcomes (e.g., benefits and harms) also had more than one study conclusion.

Table 6.4 Number of conclusions in the abstract

Conclusions	Articles
1	244
2	39
3	4
4	1
Total	*288*

6.3 Results

Table 6.5 contains some examples of the modified conclusions and associated consensus strength ratings. Included are the key word combinations that the algorithm used to provide an automatic 'strength of causal inference' rating. Despite many iterations involving modifications to the key word table, automatic strength ratings and personal ratings from the three reviewers, it became apparent that the word/combination data relating to the causal strength of different words, taken from linguistics and natural language processing and enhanced in our study, were not always able to distinguish the different strengths of conclusions accurately. In fact, sometimes a whole modified conclusion (18/338) needed to be available to the algorithm to accurately give the appropriate 'strength of causal inference'.

The most frequently occurring words and word combinations in the modified conclusions can be seen in Table 6.6, along with the strength rating used by the automatic algorithm, and a few of the relevant grammatical categories they belong to. The words 'association' and 'associated' are common and, in fact, with 'associate', these words appear in 129/338 (38%) of the modified conclusions and 124 of these — almost all — were given the rating 'Not strong', which is 57% of the 217 conclusions given this rating.

Table 6.5 Examples of modified conclusions and consensus 'strength of causal inference'

Study conclusion in abstract	Modified conclusion	Algorithm key words	Strength rating
Improved sports function and patient-reported outcome measures are obtained when an autograft is used.	Outcomes are obtained when intX is used.	'are obtained when'	Strong
In conclusion, high-dose thromboprophylaxis nearly halves the rate of VTE in morbidly obese inpatients.	IntX nearly halves the rate of the outcome.	'intX nearly halves the rate of the outcome'	Strong
We found no significant overall difference in adjusted mortality between patients transported by the police department compared with EMS	We found no significant difference in the outcome between intX compared to intY.	'found no significant difference'	Not strong
Early initiation was associated with greater all-cause mortality and greater all-cause hospitalizations.	IntX was associated with greater outcomes.	'was associated with'	Not Strong
Dronedarone has not exposed patients to increased risks of death or liver disease.	IntX has not exposed patients to increased risks of the outcome.	'IntX has not exposed patients to increased risks'	Strong
Influenza vaccination was effective against hospitalization and mortality among the frail elderly.	IntX was effective against the outcome.	'was effective'	Strong
Live birth rates were significantly higher for IVF patients compared with IUI conversion when two follicles were present	Outcome rates were significantly higher for intX patients compared with intY.	'outcome rates were significantly higher'	Not strong
Rates of fetal and neonatal outcomes were similar in vaccinated and non-vaccinated women.	Rates of the outcome were similar in intX and intY patients.	'rates of the outcome were similar'	Not strong
Postoperative weight-bearing restrictions did not affect mid-term cartilage repair outcomes	IntX did not affect the outcome.	'intX did not'	Strong
For late fetal death, regular supplement use after conception may decrease risk, but numbers were small.	For the outcome, intX may decrease risk, but numbers were small.	'intX may decrease risk'	Not strong
5-year disease-free survival rates were not significantly different for patients undergoing transplantation using 3 types of donors	Outcome rates were not significantly different for patients undergoing intX and intY.	'were not significantly different'	Not strong
Past exposure to thiopurines increases the risk of myeloid disorders 7-fold among patients with IBD.	Past exposure to intX increases the risk of the outcome.	'intX increases'	Strong

Table 6.6 Most frequent words and word combinations found in the modified conclusions

Word or word combination	Strength	N	Copula Verb Be	Epistemic Modal Verb	Evaluative Verb	Predicate	Predicate Argument
was associated with	Not strong	50				✓	
is associated with	Not strong	24				✓	
suggest that	Not strong	10				✓	
potential	Not strong	8					
is not associated with	Not strong	7				✓	
intX did not	Strong	7					✓
intX does not	Strong	6					
are associated with	Not strong	6				✓	
association between	Not strong	6				✓	
observed	Not strong	6					
resulted in	Strong	6				✓	
showed	Not strong	6			✓		
was not associated with	Not strong	6				✓	
was independently associated with	Not strong	5				✓	
outcomes are	Strong	5	✓				
intX may be associated with	Not strong	5					✓
intX reduced	Strong	4					✓
can	Strong	4		✓			
appears	Not strong	4		✓			
conferred	Strong	3					
could	Not strong	3		✓			
findings suggest	Not strong	3				✓	
improve	Strong	3			✓		
reported	Not strong	3					
is associated with an increased risk of	Not strong	3				✓	
suggests that	Not strong	3				✓	
results in	Strong	3				✓	
the outcome was similar	Not strong	3					✓
therefore	Strong	3					

Table 6.7 Examples of the varying strength of words/phrases depending on context

Word or Phrase	Rating given	Modified conclusions	Rating given
improved	Strong	IntX **improved** the outcome.	Strong
		IntX was associated with **improved** outcomes.	Not strong
causal	Strong	We report a strong relationship between the outcome and intX. Patients treated with intX worldwide should be advised about this association and further studies should evaluate the potentially **causal** role of intX in these findings.	Strong
		The estimated **causal** effects of intX and intY were.	Not strong
potential	Not strong	IntX was associated with increased outcomes. These data underscore the **potential** for intX to promote the outcome.	Strong
		This study provides support for the **potential** for intX to exert outcome effects.	Not strong
observed	Not strong	In this prospective study of intX, we **observed** a significant increase in the rate of the outcome, a risk that must be weighed against the benefits of preventing adverse outcomes.	Strong
		A significant difference was **observed** in the outcome between intX and intY.	Not strong
showed	Not strong	IntX **showed** beneficial effects comparable with intY.	Strong
		IntX and intY **showed** similar safety with no differences in the outcome.	Not strong
there was no difference	Not strong	**There was no difference** in the long-term effectiveness of IntX and IntY in this population.	Strong
		There was no difference in the outcome after treatment involving IntX or IntY.	Not strong

To highlight the dependence that individual words have on context for their meaning, including the ‘strength of causal inference’ they might impart to a conclusion when added, Table 6.7 lists a number of examples where specific words initially given a strength rating, were found in conclusions that ended up with opposite ratings.

6.3.3 Interrater agreement

The first rating by the three reviewers, using a 3-item scale of ‘Weak’, ‘Moderate’, or ‘Strong’, produced relatively poor agreement with Kappa = 0.19 and the ICC = 0.32. After discussion resolved some initial differences in word interpretation, and in some cases corrected errors, agreement improved with Kappa = 0.44 and ICC = 0.66. To achieve further agreement, the ratings were converted to a binary scale with ‘Weak’ or ‘Moderate’ converted to ‘Not strong’, with ‘Strong’ remaining as it was. This resulted in Kappa = 0.75 and ICC = 0.76. From there, the remaining differences were either successfully resolved, or for 59 (17%) modified conclusions, a two against one majority was used to provide a consensus strength rating.

Some of the modified conclusions over which the three reviewers initially disagreed are shown in Table 6.8.

Table 6.8 Examples of modified conclusions where reviewers initially gave different ratings

Modified conclusions	Reviewer 1	Reviewer 2	Reviewer 3	Final consensus
The estimated causal effects of intX and intY were.	Strong	Moderate	Weak	Not strong
Comparison of intX and intY revealed no differences in the outcome. These findings should provide helpful information for clinicians.	Strong	Moderate	Weak	Not strong
Patients receiving intX showed significantly lower outcome risk compared with intY.	Weak	Moderate	Strong	Not strong
The rate of the outcome was significantly lower using intX.	Weak	Moderate	Strong	Not strong
The outcome was superior in patients receiving intX.	Weak	Moderate	Strong	Not strong
Patients treated with intX reported deterioration of outcomes in comparison with intY.	Weak	Moderate	Strong	Not strong
Patients born after intX had a higher risk of the outcome compared with intY patients, but favourable outcomes compared to intZ.	Weak	Moderate	Strong	Not strong
An almost 4-fold increase in the outcome was observed after intX compared with intY.	Weak	Moderate	Strong	Not strong
Patients initiating intX were more likely to develop the outcome.	Weak	Moderate	Strong	Not strong
Individuals who received intX had a greater risk of the outcome.	Weak	Moderate	Strong	Not strong
A significant difference was observed in the outcome between intX and intY.	Weak	Moderate	Strong	Not strong
The improvement of outcomes was superior after intX than after intY.	Weak	Strong	Strong	Strong
We found that intX performed better than intY.	Weak	Strong	Strong	Strong
IntX showed a statistically significant higher performance than intY.	Weak	Strong	Strong	Strong
Our findings support intX.	Moderate	Moderate	Strong	Strong
IntX predicted outcomes at follow-up.	Moderate	Strong	Strong	Strong
IntX was less harmful than intY.	Moderate	Strong	Strong	Strong
IntX was associated with the outcome. These findings have significant implications.	Strong	Moderate	Moderate	Not strong

6.3.4 Statistical methods used and reported

Figure 6.1 suggests that articles expressing strong causal inference were less likely to have used statistical methods designed to improve control of adjustable biases. Of the three categories: multivariable regression, propensity score methods (compared to other multivariable regression methods), and sensitivity analysis, each control for these biases in different though related ways.

When a study used an inactive control as the comparative group intervention—'no intervention' or 'usual care'—then the proportion expressing strong causal inference was found to be around half that of 'Not strong' (Figure 6.2), compared to studies with two or more active interventions compared, where 'Not strong' and 'Strong' causal inference were approximately equal.

Also in Figure 6.2, when studies focused on unintended harms or adverse effects of an intervention, such as drug side-effects or long-term health risks, they were less likely to use strong causal language in their conclusions than if they focused on the positive health benefits of an intervention, such as improved symptoms or survival.

The final graphs in Figure 6.3 suggest no link between the 'strength of causal inference' and authors who reported their method of missing data handling. But a lower chance of using strong causal language was found for articles that had adequately described their methodology, where we thought a clear picture of the methods they used could be obtained from their reporting.

An alternative way to compare these proportions is to calculate an odds ratio using univariate logistic regression, and Table 6.9 presents odds ratios with corresponding confidence intervals for each of the comparisons in Figure 6.1, Figure 6.2 and Figure 6.3. We also explored a number of possible multivariable models, however, with considerable uncertainty over the causal structure of the relationships between the variables, it was decided that too many possibilities existed and this would make interpreting such models difficult. Some relationships are briefly explored in Table 6.10 and Table 6.12. The exercise included an attempt to create a causal diagram, and it was the difficulty of doing this that led us to two realisations. One was to increase our doubt that some of the variables are really causes of the outcome (strong causal language), with the type of software used and whether the methodology was adequately described considered the most unlikely to be causes. The second was the level of uncertainty over which variables might be causes of other variables, such that they might act as confounders.

There was only very weak evidence suggesting that the result of comparing group outcomes had an effect on the ‘strength of causal inference’ in study conclusions (Table 6.10). When stratified by the type of outcome there was some difference; however, this largely just reflected the difference seen in the second graph of Figure 6.2, where ‘strong’ causal language was much more likely if the outcome was a health benefit than if a harm to health was the outcome.

There appears to be no obvious association between strong causal inference and study size (Table 6.11), while for intervention type (Table 6.12), a difference can be seen between some study types, notably drugs, and a number of the other intervention types such as surgery. Also displayed was the relationship between intervention type and whether an inactive control was used. In most cases, studies with intervention types associated with strong causal language were also more likely to not use an inactive control.

No clear difference in ‘strength of causal inference’ is apparent between different author locations, in terms of the continent where they all reside (Table 6.14). But journals in the categories of Infectious Diseases (60%), Gastroenterology & Hepatology (59%), and Surgery (57%) had the highest proportion of studies with causal inferences rated strong, while Critical Care Medicine (13%), Urology & Nephrology (13%), and Cardiac & Cardiovascular Systems (24%) journals had the lowest proportion.

Finally, studies that used SAS, Stata or R, appeared to use weaker causal language, on average, compared to studies using SPSS (Table 6.15).

Figure 6.1 Methods used and 'strength of causal inference'

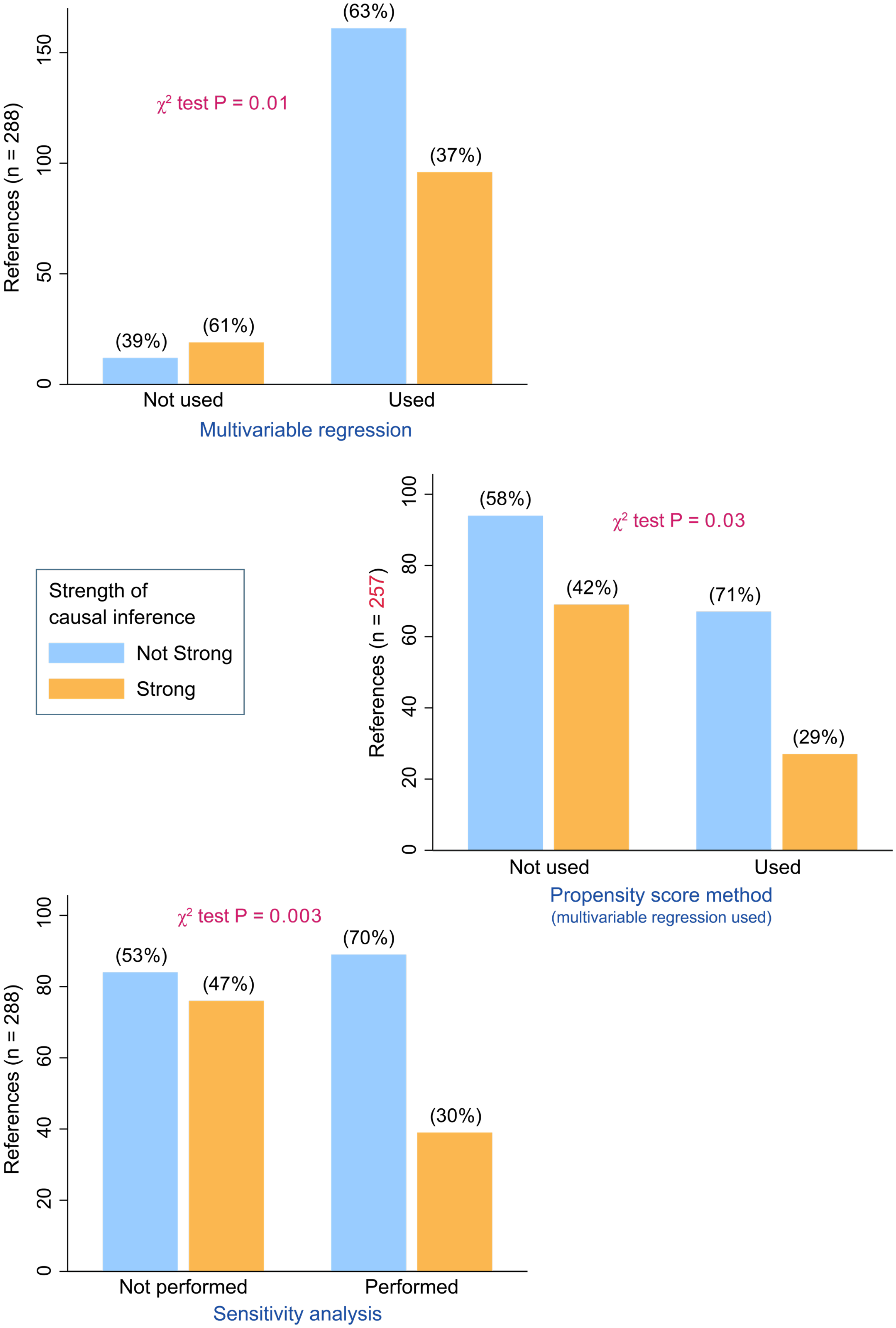


Figure 6.2 Study design features and ‘strength of causal inference’

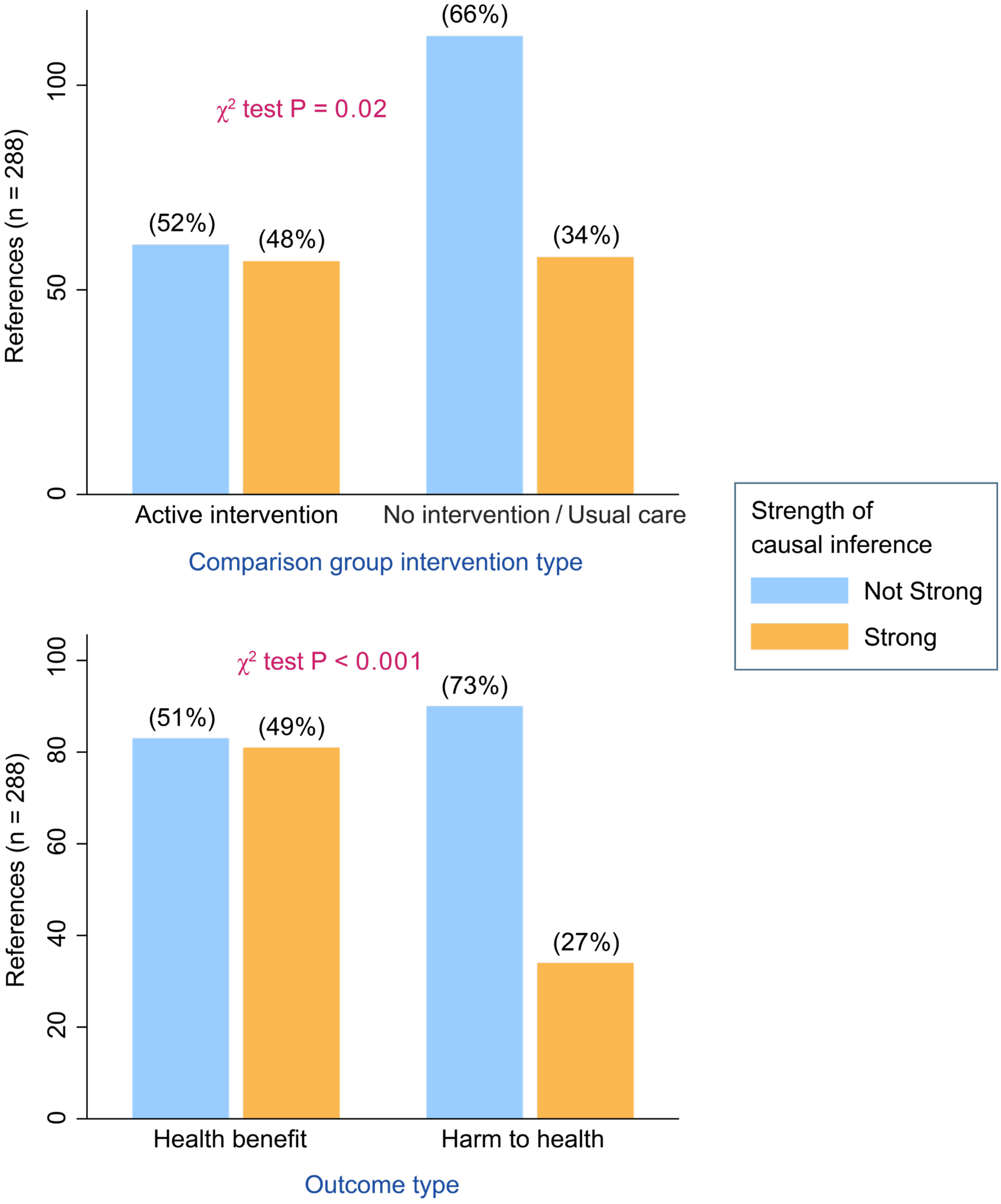


Figure 6.3 'Strength of causal inference' and reporting of methodology

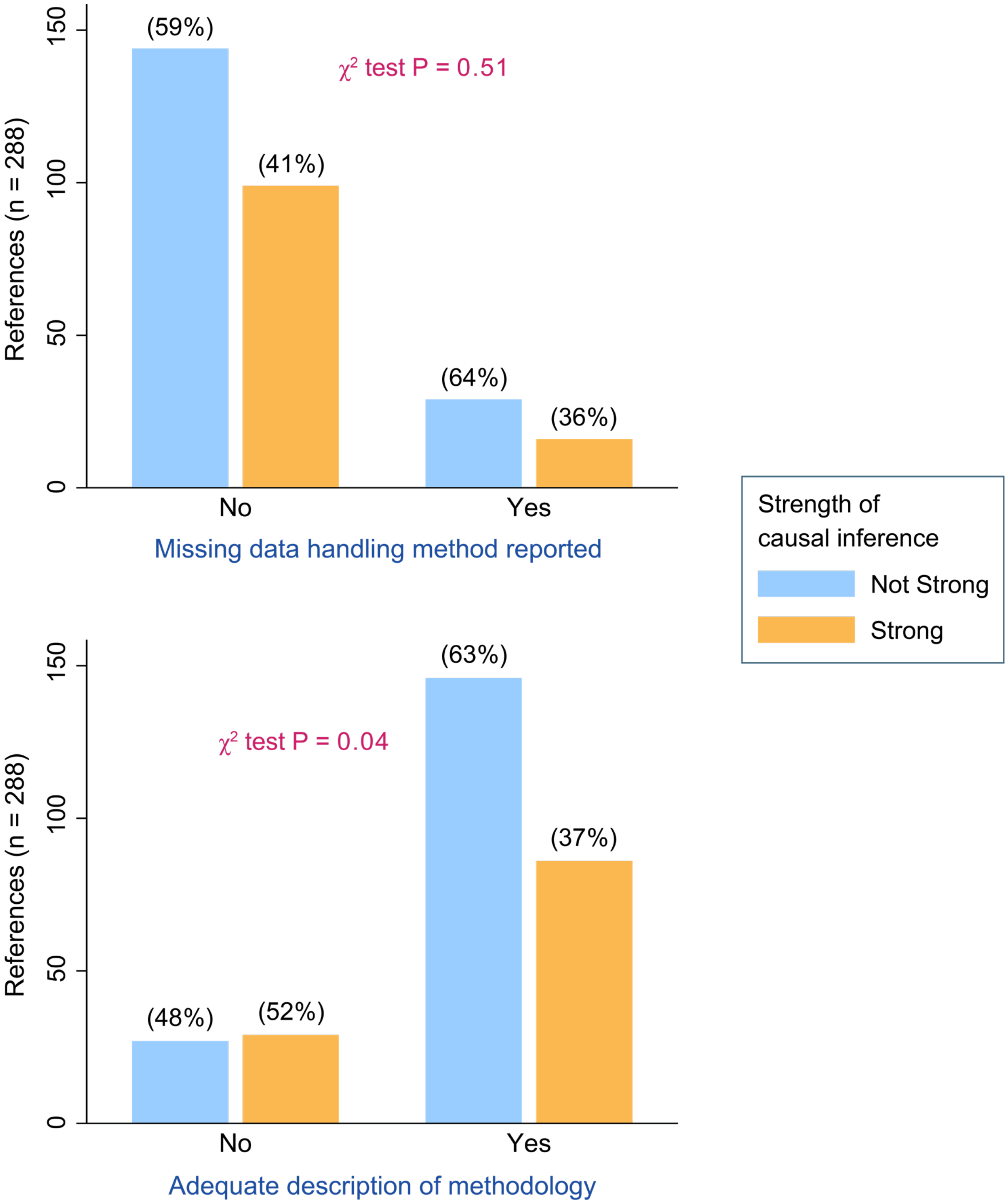


Table 6.9 Results from logistic regression with outcome: 'Strong' causal language

Univariate logistic model results for each variable

	Odds ratio	**95% CI**	**P**
No multivariable method used (compared to use of a multivariable method)	2.7	(1.2–5.7)	0.012
Multivariable but no propensity score method used (compared to use of a propensity score method)	1.8	(1.1–3.1)	0.031
No sensitivity analysis performed (compared to performing one)	2.1	(1.3–3.4)	0.004
Methodology not adequately described (compared to providing adequate description)	1.8	(1.0–3.3)	0.045
Comparison group used active control intervention (compared to inactive intervention or usual care)	1.8	(1.1–2.9)	0.016
Outcome is improvement in health or health benefit (compared to a harm to health)	2.6	(1.6–4.3)	0.000
Group results similar or no difference reported (compared to a difference found between groups)	1.3	(0.8–2.2)	0.27

Table 6.10 Group outcome comparison result and the 'strength of causal inference'

Percentages relate to row N

	N	'Strong' causal language
Overall		
Similar (null result)	92	41 (45%)
Different	196	74 (38%)
Total	288	*P = 0.27*†
If outcome is **harm** to health		
Similar	44	13 (30%)
Different	80	21 (26%)
Total	124	*P = 0.69*
If outcome is health **benefit**		
Similar	48	28 (58%)
Different	116	53 (46%)
Total	164	*P = 0.14*

† Chi-squared test

Table 6.11 Study size and 'strength of causal inference'

Percentages relate to row N

Study Total Subjects	N (288)	'Strong' causal language
200 - 799	75	35 (47%)
800 – 4,999	75	27 (36%)
5,000 - 29,999	63	35 (56%)
30,000 - 10,912,834	75	18 (24%)

Table 6.12 Intervention type and 'strength of causal inference' plus type of control

Percentages relate to row N; highlighted values: high = magenta, low = blue

Intervention Type	N (288)	Strong causal language	Inactive control
Assisted reproductive tech.	19	10 (53%)	8 (42%)
Drug	120	41 (34%)	76 (63%)
Mix	15	5 (33%)	10 (67%)
Other*	56	17 (30%)	40 (71%)
Radiation therapy	6	3 (50%)	2 (33%)
Surgery	60	31 (52%)	22 (37%)
Vaccine	12	8 (67%)	12 (100%)

* For example, hospital procedures that do not fall under the other intervention types; interventions relating to quality or timing; other health services

Table 6.13 Journal Category and the 'strength of causal inference'

Percentages relate to row N; highlighted values: high = magenta, low = blue

	N (320†)	Strong causal language
Cardiac & Cardiovascular Systems	17	4 (24%)
Critical Care Medicine	15	2 (13%)
Gastroenterology & Hepatology	17	10 (59%)
Infectious Diseases	15	9 (60%)
Medicine, General & Internal	28	7 (25%)
Obstetrics & Gynecology	35	15 (43%)
Other categories	116	48 (41%)
Peripheral Vascular Disease	19	9 (47%)
Surgery	42	24 (57%)
Urology & Nephrology	16	2 (13%)

† Some journals were in multiple categories

Table 6.14 Other study features and the 'strength of causal inference'

Percentages relate to row N

	N	Strong causal language
All articles	288	115 (40%)
Author Continent		
North America	132	46 (35%)
Europe	82	34 (41%)
Asia	26	11 (42%)
Other or Multiple*	48	24 (50%)
Total	288	*P = 0.31*†

* 'Multiple' if any of the authors were from different continents; † Chi-squared test;

Table 6.15 Software use and 'strength of causal inference'

Percentages relate to row N; highlighted values: high = magenta, low = blue

	N (333†)	Strong causal language
Not Specified	42	20 (48%)
Other	20	10 (50%)
R	35	13 (37%)
SAS	110	39 (35%)
SPSS	70	37 (53%)
Stata	56	15 (27%)

† Some articles used more than one software package

6.3.5 Results if 'Strong' not used for 'no effect found'

To see what the results would look like if a causal inference was defined as Li et al.[715] had defined it—only for conclusions after evidence of 'an effect' was found, with the weakest rating given to conclusions of 'no effect'—all 'Strong' causal strength ratings were changed to 'Not strong' if a difference in the average outcome between the intervention groups was not observed.

The shift of 41 studies that did not find evidence of an effect, from a rating of 'Strong' to 'Not strong' causal inference, makes an obvious difference to the relative numbers in Table 6.16. This shift also had an impact on some of the comparisons made in this chapter.

Table 6.16 Alternate definitions of a causal inference and group outcomes comparisons
Values referred to in the text are highlighted in magenta

		Causal inference definition can encompass		
Group outcomes comparisons	**N**	'Effect' and 'No effect' (Ch.6)		'Effect' only (Li et al.)
		'Not strong' : 'Strong' ratio		
Similar (null result)	92	51 : 41	→	92 : 0
Different (evidence of causal effect)	196	122 : 74	→	122 : 74

For example, a substantial difference occurred with the relatively low number of articles that did not use a multivariable regression method (Table 6.17). By seeming chance, the ratio has been reversed. Combined with the p-value moving to the other side of 0.05, the inference would change to either one of 'no effect', or one where the inference is not clear; as opposed to the weak evidence we found, using our definition of a causal inference, of a much greater proportion of those who didn't use multivariable regression also favouring strong causal language in the conclusion.

Table 6.17 Alternate definitions of a causal inference and multivariable regression
Values referred to in the text are highlighted in magenta

		Definition of a causal inference can encompass			
Multivariable regression	**N**	'Effect' and 'No effect' (Ch.6)		'Effect' only (Li et al.)	Change in inference
		'Not strong' : 'Strong' ratio			
Not used	31	12 : 19	→	19 : 12	reversed
Used	257	161 : 96	→	195 : 62	no change
*P-value for difference**		*0.01*		*0.08*	

* Chi-squared test

However, with most of the comparisons in this chapter exhibiting larger numbers in each category, the shift of some articles from the 'Strong' to the 'Not strong' column was more evenly balanced, with a similar proportion of articles shifting in each category. Nevertheless, three other comparisons that did change with the different definition of a causal inference are worth noting. The first involved whether the methodology was considered adequately reported (Table 6.18), with the inference changing from an association with the 'strength of causal inference' to no clear association, when the definition changes.

Table 6.18 Alternate definitions of a causal inference and reporting of methodology

Values referred to in the text are highlighted in magenta

		Definition of a causal inference can encompass			
Description of methodology	**N**	'Effect' and 'No effect' (Ch.6)		'Effect' only (Li et al.)	Change in inference
		'Not strong' : 'Strong' ratio			
Inadequate	56	27 : 29	→	39 : 17	equal to unequal
Adequate	232	146 : 86	→	175 : 57	no change
*P-value for difference**		*0.04*		*0.37*	

* Chi-squared test

Table 6.19 Comparison group type with alternate definitions of a causal inference

Values referred to in the text are highlighted in magenta

		Definition of a causal inference can encompass			
Comparison group intervention type	**N**	'Effect' and 'No effect' (Ch.6)		'Effect' only (Li et al.)	Change in inference
		'Not strong' : 'Strong' ratio			
Active control	118	61 : 57	→	84 : 34	equal to unequal
Inactive control	170	112 : 58	→	130 : 40	no change
*P-value for difference**		*0.02*		*0.31*	

* Chi-squared test

In Table 6.19, changing the definition of a causal inference likewise changed the inference for having an inactive control group defined in a study, from a possible effect to not clear.

Lastly, SPSS appears to have had the greatest proportion of articles change from ‘Strong’ to ‘Not strong’, and with it the inference that SPSS users used stronger causal language in conclusions, on average, than the users of SAS, Stata or R.

Table 6.20 Software type with alternate definitions of a causal inference

Values referred to in the text are highlighted in magenta

	Definition of a causal inference can encompass				
	N	‘Effect’ and ‘No effect’ (Ch.6)		‘Effect’ only (Li et al.)	Change in inference
		‘Not strong’ : ‘Strong’ ratio			
Not Specified	42	22 : 20	→	29 : 13	equal to unequal
Other	20	10 : 10	→	13 : 7	equal to unequal
R	35	22 : 13	→	26 : 9	no change
SAS	110	71 : 39	→	82 : 28	no change
SPSS	70	33 : 37	→	47 : 23	equal to unequal
Stata	56	41 : 15	→	47 : 9	no change
*P-value for difference**		*0.04*		*0.33*	

* Chi-squared test

6.4 Discussion

The idea motivating this review is that scientific progress depends not only on researchers avoiding bias, but that they also convey the uncertainty that remains when a study is reported. In summary, after a brief review of causal language in the literature, our first objective was to rate the ‘strength of causal inference’ implied in the final study conclusions. The second objective involved assessing whether the ‘strength of causal inference’ might be affected by the use of more advanced statistical techniques, as well as with other study features associated with the design, interpretation and reporting. Using a broader definition of a causal inference than many researchers might tend to use, this review suggests that 40% of 288 health intervention cohort studies implied relatively ‘Strong’ causal inference in study conclusions, as opposed to ‘Not strong’. We found that articles using either multivariable regression, propensity score methods (compared to other multivariable regression methods), or sensitivity analysis, were more likely to express ‘Not strong’ causal inference in study conclusions. Some associations were also noted with other study features, such as whether an inactive control intervention was used, and whether the outcome was a benefit to health or a harm. Given the evidence of bias summarised in Chapter 4, some of these cohort study conclusions are probably wrong, but confidence that exceeds the uncertainty will only compound any effect of evidence that is false.

6.4.1 Review of causal language and strength rating

Research sometimes involves an exploration to see what might be possible, and one outcome of this review is that an automatic algorithm that will rate the ‘strength of causal inference’ no longer seems an achievable goal, or at least, not as the sole judge. Partly, this is because the exact meaning of single words depends heavily on context,[132] and the number of possible contexts in a conclusion would seem to be very high. The other apparent reason is that the causal strength implied by a study conclusion was often not clear-cut, with different reviewers interpreting words, and consequently the strength, a little differently. All communication involves some ‘reading between the lines’, with a balance maintained between the risk of losing the reader’s interest with tedious details, and the risk of misinterpretation from insufficient detail.[254] Hence, communication involves the reader (or listener) making inferences about the meaning intended by the writer, and this will often leave multiple interpretations as a possibility.[733] For example, the study conclusion “Intervention X had a lower risk of the outcome” might mean to some readers merely that Intervention X had a lesser association with the outcome than its comparator intervention, or it might imply that an association was observed in the study **because** Intervention X had a lower risk of **causing** the outcome. Said a different way, the inherent ambiguity and vagueness of language[132,254,734,735] means that the wording of conclusions will often not have a single precise and reasonable interpretation. The interpreted meaning will also depend on the reader's knowledge, experience and beliefs, and thus cannot be guaranteed to be the meaning intended by the author.[736] For these reasons, an automatic algorithm to interpret the causal strength of some text will never be able to provide a truly objective interpretation. Hence, perhaps it is better to involve people when rating causal language, to avoid the misperception that ratings provided by an algorithm must be objective and therefore 'correct'. An algorithm can provide decision guidance, however, and it appeared to help in this review.

From the first edition of *Modern Epidemiology* (1986), and quoted in at least two articles since then,[678,692] Rothman lamented that:

> Some scientists are reluctant to speak so blatantly about cause and effect, but in statements of hypothesis and in describing study objectives such boldness serves to keep the real goal firmly in focus and is therefore highly preferable to insipid statements about 'association' instead of 'causation'.

It is clear from Table 6.6, however, that the word 'association' remains common as a means of describing the results to avoid an explicit statement of a causal inference. It does not avoid implicit inferences, however, though they are likely to be interpreted as weak. Still, if words are included that make clear the causal aim of the research, then using the word 'association' will often help to convey an appropriate sense of uncertainty.

6.4.2 Associations and interpretations

In this exploratory review, we wished not only to examine the strength of causal language that a range of studies implied in their conclusions, but to explore the factors that might have an influence on this strength. Understanding the potential causes of a problem, such as overconfidence expressed in some study conclusions, may lead to methods that can reduce the problem.

Following the review on statistical methods, the subject of Chapter 5, such methods were again the main focus in this review. Using propensity score methods compared to other regression methods, using multivariable regression over simple methods, or performing a sensitivity analysis compared to not doing so, were each associated with increased caution in judgements about causal effects. This finding offers support for the general assumption (as perceived in statistical circles, at least) that people with more advanced methodological knowledge tend to be more cautious with causal inference, even though more advanced methodology will sometimes provide better evidence for causality. It may be that, on average, the use of such methods will prompt researchers to consider more potential confounders than if those methods were not used, and this may lead them to consider more alternative explanations of the results, so developing a greater awareness of the uncertainty in their findings.

This may also help explain why less 'strong' causal language was used in articles that had adequately described their methodology, assuming that greater methodological knowledge would increase the quality of their reporting of the methodology; though no relationship was found for whether missing data handling was reported.

Likewise, SPSS developed a menu driven user interface earlier than the other major statistics packages like SAS and Stata[737] and this is perhaps why it often seems to be used in beginner statistics courses, at least in the health sciences. Hence, if less experienced researchers are more likely to use SPSS, this may explain why SPSS was associated with stronger causal language than SAS, Stata or R.

However, while inadequate reporting and software package used might help to predict the use of strong causal language, it does not seem likely that these are causes. More plausible is that they share common causes with the strength of causal language, such as the investigator's level of statistical knowledge.

Regarding features of study design, an association was not detected between the number of study subjects and the strength of causal language. On the other hand, when the comparative group was defined as receiving an inactive control intervention ('no intervention' or 'usual care'), then 'Not strong' causal inference was twice as likely in the conclusion as 'Strong'. Whereas when active controls were used, 'Not strong' and 'Strong' causal inference were approximately equal. Perhaps by comparing the primary intervention to what might happen if nothing is done, at least approximately, helps to provide a better causal contrast for imagining alternative explanations?

An even stronger difference was found between harm outcomes, where 'Not strong' was much more likely than 'Strong' causal inference, compared to those conclusions addressing a health benefit outcome, where the probability of 'Strong' was approximately equal to the probability of 'Not strong' causal inference. A plausible explanation might be that researchers were influenced by confirmation bias, given that all interventions had a history of being used and those involved may have already believed that it was an effective intervention.

Another relationship found in the data was that studies investigating drugs were noticeably less likely to use 'strong' causal language (34%) than other study areas, such as assisted reproductive technology (53%), surgery (52%), and vaccine studies (67%). Possible underlying causes may relate to differences in professional culture, for example, surgery compared to the more heavily regulated pharmaceutical industry. Alternatively, studies with intervention types associated with strong causal language were also more likely to use an active control, similarly associated with strong causal language. It is not clear, however, how the intervention types listed might relate to the journal categories, other than with the most obvious ones like surgery.

Finally, the definition of a causal inference that we used to judge the strength of causal inferences in study conclusions was uncommonly broad. Hence, to gain an idea of how the results might have differed with, for example, the definition of a causal inference used by Li et al.[715]—only for conclusions after evidence of 'an effect' was found, with the weakest rating given to conclusions of 'no effect'—we changed all 'Strong' causal strength ratings to 'Not strong' if no difference in group outcomes was the reported finding. Not surprisingly, a number of inferences did change, all from 'evidence of an effect' to one of 'no effect', such as the link between the use of an active intervention for the comparison group and 'strength of causal inference.

Many factors have been suggested to help explain overconfidence expressed in study conclusions, including the pressure to publish,[429] for which mixed evidence exists.[738,739] Other suggested factors include financial as well as social conflicts of interest.[469,740,741]

Potential mechanisms for a link between considering alternative explanations and caution when judging causal effects, may be related to cognitive biases,[109] a topic covered in detail in Chapter 4. One potential debiasing technique relevant to causal inference might simply be trying to think of alternative explanations for the results, which the use of methods that control for more confounders might encourage. However, while overstatements of evidence appear to be common, they highlight the influence that many cognitive biases can have on causal inference; and while cognitive biases might lead to unjustified causal beliefs, they appear to affect everyone.

While the use of more complex methods aimed at confounder control is one way to encourage this, an additional method that does not require expert statistical training is to simply, and deliberately, think about alternative explanations for the results. This might be any combination of: creating a list of plausible alternative explanations after searching the literature; a causal diagram such as a directed acyclic graph (DAG);[250] or a quantitative bias analysis.[578–580] Each of these may prompt the researcher to think of potential confounders not previously considered, leading to a greater appreciation of the true uncertainty attached to most research results.

For researchers to develop greater experience in adjusting for many confounders, various things must happen, with one clearly being that the process of learning and using new methods needs to be sufficiently simple, easy and quick. Otherwise, the researcher's other professional responsibilities will soon capture and probably hold onto their attention.

A journal might add a requirement to include a named heading in Discussions such as we have used here: "Limitations and alternative explanations". Having such a heading would encourage authors to think more about the factors that increase the uncertainty of their results. This would hopefully increase the caution of authors who might otherwise have formed overconfident conclusions.

6.4.3 Limitations and alternative explanations

The strengths and limitations mentioned in Chapter 5 will apply here as well, and the most important may be that this is intended as an exploratory review rather than a test of hypotheses. However, analyses that are called "exploratory" can still provide evidence relating to a causal question. Nevertheless, using the term "exploratory" warns the reader that the evidence should be considered suggestive, or fairly weak.

Regarding the link between propensity score methods and strength of causal language, one alternative explanation is that papers using propensity score methods may be sent to reviewers more likely to ask authors to add more caution to their causal claims. However, this involves reasoning that is somewhat circular as it assumes that reviewers with expertise in propensity score methods will prefer more caution, thus it assumes the association already exists that this alternative mechanism attempts to explain.

Another plausible alternative explanation is that studies expressing weaker causal inference were more likely to include a professional statistician as one of the authors. But rather than following the circular reasoning of the previous paragraph, that is, that statisticians will generally have more experience with advanced methods, statisticians also differ from researchers who are clinicians in that statisticians do not have to make regular clinical decisions. Such decisions often involve a need for certainty where often there is none. As a result, statisticians may feel more comfortable incorporating uncertainty into their decisions. It is unknown how many studies in our sample used a statistician, however, because professional occupations are usually not included. Additionally, involvement of a statistician is not always acknowledged with authorship.[742,743]

A further difference between the work of statisticians and health researchers is that the number of projects a statistician might contribute to, and potentially be an author on, is often going to be larger because of the nature of their work. A researcher, on the other hand, is more likely to work on a single project and hence, might feel more pressure to publish an important finding to safeguard their career opportunities.

Lastly, researchers less familiar with the relatively more advanced methods for confounder control might also be less familiar with articles recommending caution when making inferences from research.

Reviews assessing the quality of health research are an important means to both monitoring the current standard, as well as viewing whether changes occur over time. Reviews can act to highlight areas that can most be improved, and those that can most **easily** be improved, as well as provide recommendations on how this can be done. In general, improvements in skills only come from deliberate efforts to improve,[744] and these require not only incentives, but also the means by which the factors underlying the occurrence of bias can be countered. This will be the topic of the final chapter of this thesis.

Chapter 7
Case study: Understanding the potential biases in a study

List of acronyms and synonyms

RCT	Randomised controlled trial
Telemonitoring group	Intervention group
TM	Telemonitoring
BP	Blood pressure
SBP	Systolic blood pressure
DBP	Diastolic blood pressure
GP	General practitioner
HbA1c	Glycated haemoglobin A1c
BMI	Body mass index
DAG	Directed acyclic graph
CI	Confidence interval
P	P-value from a statistical test
N	Number of participants
SD	Standard deviation
IQR	Interquartile range
MCAR	Missing completely at random
MAR	Missing at random
MNAR	Missing not at random
MI	Multiple imputation
IPW	Inverse probability weighting
MICE	Multivariate imputation by chained equations

7.1 Introduction

7.1.1 Overview

In Chapters 7 and 8, we present a case study where we apply some of the principles discussed in this thesis. It centres on the analysis of a randomised controlled trial (RCT) of a telemonitoring service and our aim was to provide conclusions that were more accurate and relevant than we might otherwise have delivered. We also wanted to better understand and communicate to stakeholders and clinical researchers the true level of uncertainty that remained following the analysis. This communication goal became more important when the study revealed much more missing data than was expected, and as a result, needed to be analysed as if it were an observational study.[745] Although the level of missing data would reduce the certainty of our conclusions, the human predilection for causal thinking, discussed in Chapter 4, may have left some of the staff involved believing that a causal relationship existed based primarily on their anecdotal observations during the trial. In reality, whether true or not, those causal inferences are likely to have been influenced by confounding and selection bias,[745] as well as by the cognitive biases that can influence causal judgements, including confirmation bias[746] and overconfidence bias.[305]

Our view was that, of greatest value, might be an analysis that properly assessed the potential sources of confounding, selection bias, measurement error, and cognitive biases and, where possible, controlled for as much confounding and selection bias as could be determined, while ensuring that the level of uncertainty remaining was well understood and communicated.

To facilitate an extended discussion of bias relating to the case study, the presentation is divided into two chapters. In Chapter 7, we focus on describing the study and the data, including the measures taken in response to missing data. The overall aim is to promote an understanding of the potential biases this study is exposed to. In Chapter 8, our focus shifts to the analysis of the data and presentation of the results; using models to reduce the potential for bias and sensitivity analyses to better understand and communicate the uncertainty. We also explore the concept of time-dependent confounding in a separate analysis that uses the parametric g-formula.

More specifically, in this chapter we will:

1. briefly present the background and design of the case study
2. fully describe the data collected and the data that is missing
3. explain how we assessed and inferred aspects of the missing data mechanism
4. examine the possible effects of the missing data in terms of biased results and increased uncertainty, and the use of multiple imputation to try to reduce such effects
5. display the causal diagrams we constructed to more easily identify and communicate potential sources of bias

7.1.2 Pragmatic trials

Telemonitoring trials, the type of trial assessed in this case study, have returned mixed results over the last two decades. There have been at least 20 randomised controlled trials (RCTs)[747–766] and 4 observational studies[767–770] assessing either home blood glucose or blood pressure measurement and all combined with some form of remote assessment and support. The HCF Telemonitoring RCT was a pragmatic trial that offered remote home blood glucose or blood pressure self-measurement, with associated telemonitoring by nurses. Originally introduced by Schwartz and Lellouch,[771] the term 'pragmatic trial' refers to a randomised controlled trial where the intervention: (a) resembles those that are already in routine use and may be combined with other interventions, as would occur in normal clinical practice; (b) where the main aim is to inform routine clinical decision making, as opposed to testing whether the intervention really can cause improvements in some people; and (c) is trialled with a broad patient group that is sufficiently representative of those encountered in normal clinical practice.[772]

In many cases, the analysis of a pragmatic trial relies on routinely collected data. Using such data often has substantial advantages, such as less interference with usual care and fewer expenses from a reduced need for onsite staff training regarding data collection and management.[773] Relying on this type of data comes with a range of limitations, however, because the primary focus when the data is collected is on clinical care rather than answering a research question. For example, data for some confounders may not be adequately collected, such as particular diagnoses, medications or lifestyle factors, unless prompted by a voiced health concern from the participant.[773,774] This may mean that important baseline data is not available for some, or even all, study participants. Participant outcome data is also more likely to be missing if it does not represent a major life event such as death, and this is often the case in pragmatic trials.[775]

7.1.3 Missing data

Missing data is one of the main concerns when using routinely collected data,[773,774] but the mechanisms can be difficult to understand[280,776,777] and are often not handled adequately.[776,778–783] The loss of information from missing baseline, intervention or outcome data leads not only to a reduction of precision and power, but more importantly, it can also result in biased estimates.[784] Whether such bias occurs depends primarily on why participant values are missing, often called the *missing data mechanism* or *missingness mechanism*.[785] From a system developed by Rubin in 1976,[786] these reasons are commonly classified into three types using the slightly confusing[781,785] terminology of Little and Rubin (1987, 2002).[787] They are *missing completely at random* (MCAR), *missing at random* (MAR), and *missing not at random* (MNAR). As well as having ambiguous labels, the three types are also frequently described in noticeably different ways. Hence, to assist with clarity, each missingness mechanism type will be described in a variety of ways below. We will also avoid formal mathematical definitions, partly because this chapter is focused more on the practical application of methods and concepts, and partly because in recent years, differences have been highlighted in the way these terms have been formally defined by various authors,[788–791] but these details are beyond the scope of this case study.

Missing participant values are considered to be MCAR when they are, in effect, a random sample of the complete data.[785] In this case, the missingness mechanism does not depend on the values of any observed or unobserved variables in the causal network under study, including the missing values.[792] This also implies that there are no systematic differences between the missing and the observed values.[781] Describing missing participant data as MCAR is usually not a plausible assumption in health research, however.[785,792,793]

7.1 Introduction

When missing participant data can be explained by the observed participant data, the missingness mechanism is labelled MAR.[781] In this case, systematic differences do exist between the missing and observed values, however, conditioning on the measured values of the other variables removes the association between a value being missing and what that value would have been.[785] A few statistical techniques used to handle missing data, including multiple imputation, can provide unbiased estimates if the missingness mechanism is MAR and other assumptions are met. But if it is MNAR, such techniques may or may not provide unbiased estimates, depending on the nature of the missing data.[794]

If the missing participant data cannot be explained by what has been observed, then we say it is MNAR.[792] This means that the probability that a participant's value is missing is related to the value itself,[785] and that value cannot be predicted from the observed data, making statistical adjustment not possible.[552] But while this makes it more likely that particular values are missing compared to other values, and that may lead to biased estimates, such bias is not always inevitable as it depends on the specific causal structure and the parameter being estimated.[280,792]

Another term commonly encountered in the literature is *ignorability*, which is often used to mean that the missing data values are MAR or MCAR.[792] But the formal mathematical definition is a little different and means that inferences made from a parametric model of the observed data do not differ from inferences made from a joint model describing the observed data and missingness mechanism.[788] Missing data that is MNAR is sometimes referred to as *informative missingness*, meaning that the fact that the values are missing contains information about what that value is.[142]

As an example, in the Glucose arm of the Telemonitoring trial, if occasional blood glucose measurements for some participants were the only missing data, and the reason was that their glucometer happened to have a defect that led to underestimated measurements, the missing data would likely have been MCAR. Alternatively, if blood glucose measurements were missing only for people who held a full-time job and it was inconvenient to take measurements sometimes, then if employment status was fully recorded the data might be described as MAR. But if blood glucose measurements were sometimes missing because participants had eaten foods that they knew would result in a high reading and thought they would feel embarrassed providing such a reading, then assuming there was no record of diet the data would be MNAR.

Missing data from loss to follow-up or dropout is the main mechanism by which a randomised controlled trial can become as susceptible to selection bias as an observational study,[745] so a careful assessment of missing data is essential. The vital question is whether the results would have changed if the missing data had, instead, been obtained. In general, however, it is not possible to tell from the observed data whether the values that were missing were MAR or MNAR.[792] Instead, as put by Sterne et al. (2009):[781]

> The onus rests on the data analyst to consider all the possible reasons for missing data and assess the likelihood of missing not at random being a serious concern.

One tool that can assist in this assessment is causal diagrams, and a range of articles are now available that focus on causal diagrams for missing data.[244,279,280,795,796]

Once the nature of the missing data has been ascertained, attempts can be made to reduce its influence. Over the last four decades, numerous authors have divided missing data methods into two groups. Methods often labelled *ad hoc* include the older, simpler methods, like complete-case analysis; all developed before the advent of modern computers.[787] The more sophisticated and more recently developed methods, like multiple imputation, comprise the second group.[142,776,779–781,787,793,797–806] Additionally, an increasing number of authors are now referring to this second, model-based, group as the *principled* missing data methods.[776,780,793,798,800,801,803–806]

7.1.4 ‘Ad-hoc’ missing data methods

The easiest method to employ when faced with missing data is complete-case analysis, where participants with missing values for any variable of interest are simply excluded from the analysis. While complete-case analysis can produce unbiased estimates in some situations,[792] including when missingness depends on some of the covariates but not on the outcome,[794] in general, its validity relies upon the assumption that the missingness mechanism is MCAR, an unrealistic assumption in most health research studies,[785] though bias will not always occur.[807] Complete-case analysis also discards potentially useful information that is partially available on excluded individuals and as a result, will usually produce less precise estimates compared to methods such as multiple imputation.[792] Despite these problems, numerous recent reviews have found that complete-case analysis is by far the most common method used to handle missing data.[776,778–780,782,783,808,809] Underlying this are a number of factors related to mental effort and what has been called the *law of least mental effort,*[487] from among the many names mentioned in Chapter 4 describing this influence. The decision to use only complete cases means no additional mental work will be required, making it an attractive option, assuming the analyst's colleagues and peers also choose this method, as clearly most do. In addition, and also relating to avoiding mental effort, complete-case analysis is effectively the default method in most common software packages.[552,779] For example, when estimating most regression models, participants will simply be excluded if any of the variables in the model for that person do not have values.[810] And for non-statistician researchers, methods such as multiple imputation may well appear daunting to learn.

A related method is to simply drop variables with missing values from any model being constructed. But this can easily lead to bias if an important confounding variable is one of those removed, so it is far from the best option in most circumstances.[800]

The *missing indicator method* is another 'ad hoc' method.[779,797,800] It involves setting the missing values to a fixed number, such as zero, but the specific value does not matter. An indicator variable is then created for each covariate with participant values missing; with its value set to 1 if the corresponding covariate value is missing, and it is set to 0 otherwise.[811] For missing baseline data in randomised trials it is considered a valid method that will enable all participants to be included in the analysis, because the covariate values are not related to treatment allocation.[799,812] Its use in observational studies, however, is strongly discouraged because there is a considerable risk that it will introduce bias, even if the missing data is MCAR.[813–815]

The remaining 'ad hoc' methods can be grouped under the label 'single imputation', where missing values are filled in by a single value, thus allowing data from all participants to be used in most statistical modelling procedures. A popular option in longitudinal studies, though also strongly discouraged,[816] is called *last observation carried forward* (LOCF), or *last value carried forward* (LVCF), where for each participant, the missing values of any variable that is measured repeatedly are filled in with the last value that was recorded for the same participant.[785] *Mean value imputation*, on the other hand, involves filling in missing values with the mean of that participant's previous non-missing values, though it still tends to provide biased results.[779]

One problem with single imputation methods is that they do not adjust the uncertainty of the estimates, for example, by widening the confidence intervals, to take account of the fact that missing values have been imputed, because many of the imputed values will probably differ from the values that would have been recorded, had they not been missing.[785] This will increase the precision of the estimate, but that increase will not reflect a decrease in the true uncertainty of the estimate. More importantly, however, single imputation methods can also produce estimates that are biased, including when the missing data is MCAR.[779]

The term *ad hoc* is defined by the Cambridge English Dictionary to mean "happening only for a particular purpose or need, not planned before it happens",[817] and thus seems an apt label for methods that are, perhaps in many cases, more of an automatic response to missing data than methods applied with adequate forethought and planning.

7.1.5 'Principled' missing data methods

In recent decades, the term *principled* has been increasingly applied to specific methods for missing data, such as multiple imputation, inverse probability weighting, and likelihood-based methods such as mixed models.[605,776,780,782,792,793,798,800,801,803–806,808,818–824] Often the term 'principled methods' is used in a direct contrast with 'ad hoc methods'.[776,780,793,798,800,801,803–806,819] At first, this seemed a curious word to use in the sense of a label or a name, with most articles not explaining why they used it to group these methods, and the word 'principled' has not been regularly used to group any other method types. The Oxford English Dictionary[17] provides two definitions for *principled*, with the first regarding a person who shows a recognition of right and wrong, and the second regarding a method that is "based on a given set of rules". The second definition could, of course, describe the use of most statistical methods. Somewhat similarly, the Cambridge English Dictionary,[817] suggests

”based on principles, or (of a person) having good personal standards of behaviour” as its definition in American English.

A search of the literature reveals that the word’s connection with missing data methods possibly originates with Little and Rubin in 1983,[825] where they suggest:

> A principled approach to the problem of missing data in large databases requires a plausible model for the missing data mechanism and estimation procedures that remove or minimize biases introduced by the incompleteness of the data.

In 2000, Little and Rubin[798] suggested that the methods are principled because “they are based on explicit assumptions about the data and missing-data mechanism”. And Kenward and Carpenter (2007)[801] give an explanation that contrasts ‘principled’ with ‘ad hoc’:

> Principled methods are based on statistical models for the data … Unprincipled methods are characterized by ad hoc procedures – typically manipulating the data so that the analysis originally intended for fully observed data can be run.

It is quite possibly meant to convey multiple meanings, both an adherence to principles or rules, but also implicitly suggesting that the analyst who uses such methods will be displaying good personal standards of behaviour. Possibly a useful strategy to encourage non-statistician researchers to try to use the more sophisticated methods for missing data.

Perhaps the most common of these methods is multiple imputation,[826] with the ability to use common software packages having become an option in recent years.[785] But before a method such as this is employed, one of the first steps is to define the intervention, outcome and other covariates as precisely as possible, so we know the questions that are really being answered by the analysis.[827,828] For many trials this is straight forward but for some, such as this one, the amount of missing data meant that there were many equally valid definitions. Likewise, unless we defined the intervention and outcome that was of most interest in a per-protocol analysis, or included a number of possibilities, our analysis may not have provided the range of answers that would satisfy stakeholders, given the limits of an intent-to-treat analysis when there is missing data.[829]

7.2 Methods

7.2.1 Trial design

An Australian not-for-profit private health insurer, HCF, offers a chronic disease support program to its members called My Health Guardian.[830] Provided by the health management company Healthways, it features online as well as telephone support, and within this setting, a randomised controlled trial (RCT) of a telemonitoring program was conducted. It contained two intervention arms: one assessing the effects of home blood glucose self-measurement, and the other assessing home blood pressure self-measurement, each with associated telemonitoring. The aim was to assess the effectiveness of a telemonitoring service offered to suitable members from mid-2014 to mid-2016. From the point of view of the trial sponsor, HCF, they wanted to know whether the program produced meaningful improvements in the health of some of its members. This information would help them decide whether to keep the program running, make changes or end it.

Aim

To assess whether exposure to the Telemonitoring intervention lowers the mean blood glucose and/or blood pressure level in people with type 2 diabetes and/or hypertension after a minimum duration of 6 months, compared to people who were not exposed to the intervention.

Eligible participants

Those eligible were HCF members diagnosed with type 2 diabetes mellitus (Glucose arm) or hypertension (BP arm) who were participating in the My Health Guardian program and monitored their blood glucose and/or blood pressure at home. Participants also needed to reside in a Telstra mobile service area to be able to participate.

Intervention

The intervention was a telemonitoring service consisting of a Wi-Fi enabled glucometer and/or blood pressure monitor that was able to transmit blood glucose and/or blood pressure measurement data to the health service provider Healthways, combined with ongoing tailored advice via telephone calls from registered nurses employed by Healthways.

When attempting enrolment of participants via a telephone call, a verbal instruction to participants regarding measurement frequency was:

> There will be no need to increase the frequency of your readings due to your participation in this program and it is anticipated that you will continue monitoring your blood glucose levels and/or blood pressure as your treating doctor has recommended

Design

For both the Glucose and BP arms, members were randomised to be offered the telemonitoring program either in the early enrolment period, joining the Telemonitoring (TM) group if they accepted, or they were offered the program 12-24 months later, in which case they became part of the Control group if they accepted.

As mentioned above, this type of RCT is sometimes called a pragmatic trial, and in this case, the only difference between this trial and the telemonitoring service, as it would otherwise have proceeded, was the addition of randomisation that determined the enrolment period in which members were offered the program. Hence, no blinding, allocation concealment or any other strategies to avoid potential bias were employed, and no attempt was made to increase the chance that either baseline or outcome data would be collected. The data was routinely collected over the phone or by email, with no face-to-face contact between nurses providing the service and the participants receiving it.

Enrolment

For the Glucose arm, the enrolment period for the Telemonitoring group was from 1 July 2014, until all contactable randomised members had been offered the program and from 1 July 2015 for the Control group, with blood glucose outcome data collected from July 2014 to Dec 2015. For the BP arm, enrolment also occurred from 1 July 2014 for the Telemonitoring group and from 23 November 2015 for the Control group, with blood pressure (BP) outcome data collected from July 2014 to Feb 2017.

Figure 7.2 in the Results illustrates the enrolment process.

Outcomes

Before any comparative analyses were performed, the outcomes used were chosen following an assessment of data accuracy and level missing. The definition of each outcome variable can be seen in Table 7.3. The outcomes originally specified before the trial started are shown below.

For the **Glucose arm**, the primary outcome was

a) **HbA1c** (glycated haemoglobin A1c), measured at the end of the trial period by each participant. It is a marker for the average plasma glucose concentration over the previous 3 months. Values range from 4.0–12.0% and most people with diabetes aim for 6.5–7.0%.

Secondary outcomes for the **Glucose arm** were

b) **Blood glucose** measurements taken at home by the participant using the glucometer provided. The target for people with type 2 diabetes is 6-8 mmol/L before meals and 6-10 mmol/L two hours after starting meals.

c) **BMI** (body mass index), calculated from the last weight measurement recorded by each participant at the end of the trial period. This was included as an outcome because the intervention included lifestyle advice, via the telemonitoring component, that might lead to reductions in BMI.

The pre-specified outcome for the **BP arm** was simply an undefined measure of systolic blood pressure for each participant. The BP arm analysis occurred after the Glucose arm analysis revealed considerable missing data, so the BP arm outcome was not specified more precisely until the amount of data collected could be examined. The final outcome definitions are shown in Table 7.4.

For some participants, occasional BP measurements had also been collected routinely over the phone prior to starting the Telemonitoring trial as part of the My Health Guardian chronic disease support program. Because these values were recorded before any exposure to the intervention, their possible use as an alternative outcome for the Control group was also examined.

Baseline data

The telemonitoring service was provided as an add-on service to the existing My Health Guardian disease management service that participants were already using, so the trial was able to take advantage of the existing data collection system for participant information. For both arms, the baseline variables were Age, Sex, Ethnicity, HbA1c, BMI, Diabetes type, Hypertension, Hyperlipidaemia, Cardiovascular Disease, Arthritis, Back Pain, Walking Pain, Eye Problem, Insulin or Analogue, Number of diabetes drugs, Pain relief drug, Employment status, Self-employed, Moderate exercise, Smoking history, and Risk level. The baseline variable 'risk level' was also referred to by Healthways nurses as the "risk summary score". When the telemonitoring nurses conducted a clinical assessment over the phone, they also reviewed the risk level at the end of the call to determine if it should be changed, though it was not clear how this was done. We were otherwise told that the risk level was determined by a proprietary Healthways algorithm. The risk of hospitalisation is perhaps similar to what it implies. From the nurse's point of view, it determined the length of time before the nurse would call the member again, for example, 1 week or 1 month.

7.2.2 Outcome data availability

Glucose arm

Following updates from Healthways, a significant amount of missing data was expected and so we planned to make an assessment of the outcome data that was available, before finalising outcome definitions. The primary aim was to compare the mean of each outcome of the Telemonitoring group, following at least 6 months exposure to the intervention, with the mean of each outcome of the Control group taken before they had been exposed to the intervention. Ideally, the measurements to be compared would have been recorded close to a common date, to avoid possible confounding in case measurements taken far apart in time were influenced by changes to the telemonitoring service that might have occurred over time. The participants may have been encouraged to ask their GP (general practitioner) for a blood test for HbA1c, either at the end of the trial period if they were in the Telemonitoring group, or when they enrolled if they were in the Control group, but it is unclear whether this encouragement occurred if it was not required for clinical reasons at the time. Nevertheless, it was unknown how many participants would have requested the test at that time anyway, because some might not have understood that it was needed, did not want to see their GP at that time, they simply forgot, or perhaps they did not want to for some other reason. In those cases, it was hoped that a fairly recent measurement would still be available, but it was not known how far back in time we would have to go to find a measurement for most participants. This was also the case for weight measurements that were used to calculate BMI. As a result, the amount of outcome data that was available within various date ranges needed to be ascertained before outcomes could be fully defined with specific date range criteria.

Blood glucose measurements vary considerably, with a heavy dependence on the type and timing of food and the time of the measurement.[831] With blood glucose, there was no guarantee that the measurements were taken after fasting, so using a single blood glucose measurement as the outcome may have provided an inaccurate assessment of glucose control. Hence, a number of possible definitions were considered once the amount and nature of the data available could be assessed.

BP arm

For the BP arm, we wished to find out whether the program of home blood pressure self-measurement with telemonitoring, caused at least some participants to make changes to their lifestyle, diet, medication adherence, exercise level or other relevant factors, that resulted in lower mean blood pressure over time. The only outcome measurements available were the self-measurements, produced by part of the intervention (the remote monitoring and advice by nurses was the other part). Unlike management of blood glucose, no generally applicable guideline exists for blood pressure measurement frequency or the best time of day for measuring. The advice instead, is likely to depend on a range of factors related to a person's condition, treatment and personal preference.[832] Hence, we used all measurements recorded by participants, either separately, or averaged with, at most, one measurement per day.

7.2.3 Variable definitions

Given the level of missing data, we wanted to assess whether the original primary outcome definitions would provide the most accurate information about the intervention for all participants, or whether other outcome definitions would be preferable. Ideally, the outcomes would accurately reflect the participant's true mean blood glucose or blood pressure, and whether it changed in response to the intervention.

We also needed to define other variables more precisely following an assessment of the data available. For example, in a per-protocol analysis, the intervention in both arms could be defined as home self-measurement at least once a week or once a month. It could also be defined as just being given the measuring device with instructions. In each case, the number of participants eligible to be included would be different. Note that this is not important for an intention-to-treat analysis, often the preferred method because it maintains the advantages of randomisation.[833]

We also wanted the intervention definition to reflect how it would be viewed in the community. In this case, however, it was not clear what this would be, so multiple definitions were used in the analysis to give a fuller picture of the intervention's effect. And although HCF expressed interest in an intention-to-treat effect that evaluated the telemonitoring program as a whole, such estimates are most relevant when the program is to be continued unchanged and with the same level of dropout expected.[829]

7.2.4 Causal diagrams

A causal diagram can be very useful in the design stage to help identify additional potential sources of confounding or selection bias that might be measured, but while this did not happen here, with the trial commencing before this PhD project got underway, causal diagrams are nevertheless useful at every stage of a research project and can fulfil various purposes. For this analysis, we constructed diagrams for each arm of the trial to guide model construction, interpret the results and to help communicate the uncertainty that remained following the analysis.

All but one of the diagrams was a directed acyclic graph (DAG), the most common form of causal diagram to such a degree that many consider the terms synonymous (see for example[14,85,253,267]). For the Glucose arm we created the DAG in Figure 7.9, but for the BP arm, we decided to try a different approach and created an alternative causal diagram (Figure 7.10), though similar to a DAG, that might be used to help guide a statistician or researcher when they first consider the causes relating to their research question. This causal diagram was designed to make the initial collection of potential sources of confounding and selection bias easier by grouping potential sources of bias to help trigger thoughts and memories. Though not a DAG, such a diagram could act as an easier starting point that could then be used to create a conventional DAG for the analysis. By lowering the cognitive effort required to perform each step, the benefits of which were discussed in Chapter 4, more statisticians and researchers might give causal diagrams a try.

Two other DAGs can be seen in Chapter 8.

7.2.5 Missing data patterns

A missing data pattern, as displayed in the data matrix of Figure 7.1 and in Table 7.8, is called *monotone* if the variables and patterns can be reordered so that it exhibits the pattern on the left in Figure 7.1, where if the variable X_j has been observed for a participant, then all variables X_k for $k < j$ have also been observed for that participant.[800] The advantage of a monotone missing data pattern is that methods for handling such missing data can be easier to apply than methods for *non-monotone* patterns,[787] which include all patterns that are not monotone. In most health research settings, however, monotone missing data is uncommon.[800]

Figure 7.1 Monotone and non-monotone missing data patterns

Adapted from Figure 1 in Horton and Kleinman (2007); Val = observed value, ‘ - ‘ = missing

	Monotone				Non-monotone			
Pattern	Y	X_1	X_2	X_3	Y	X_1	X_2	X_3
1	Val	Val	Val	Val	Val	Val	Val	Val
2	Val	Val	Val	-	Val	Val	Val	-
3	Val	Val	-	-	Val	Val	-	Val
4	Val	-	-	-	Val	-	-	Val

Missing data patterns can also be used to explain how missing data increases uncertainty, particularly to researchers or other people with an interest in the results. To attempt this for the HCF staff coordinating the telemonitoring project, we modified a table produced by the SAS procedure PROC MI that shows missing data patterns and explained the problem by referring to specific rows as examples.

7.2.6 Multiple imputation

Multiple imputation and inverse probability weighting

Of the recommended 'principled' methods for missing data that are available, multiple imputation (MI) and inverse probability weighting (IPW) are two of the more commonly utilised.[785] IPW involves weighting complete cases by the inverse of the probability that they are a complete case.[834] In general, IPW is simpler and easier to implement than MI[785,826,834] and has advantages in specific situations, such as when the only missing data is from dropouts/attrition/loss-to-follow-up and missingness is MAR. In this case, when only the outcome is missing, a recent simulation study has suggested that MI was unable to correct for attrition bias and, in fact, performed no better than complete-case analysis.[795]

On the other hand, the validity of IPW relies on there being a sufficient number and variety of complete cases to enable the positivity assumption to be satisfied. That is, for all possible combinations of the full data, there is a non-zero probability that a complete case with those values has been observed.[792] This is because, ideally, we'd like each distinct subtype of individual with missing data to have a representation in the complete cases. We did not expect this to be the case with the HCF dataset, however.

In addition, MI is generally more efficient than IPW, producing estimates with greater precision.[792] This occurs partly because MI works by using information from all participants, whereas IPW uses only the information from complete cases. It is also partly because the MI model makes an assumption about the distribution of the missing data given the observed data, which IPW does not, and this leads to increased efficiency, though dependent on this assumption being true.[834]

MI was first proposed by Don Rubin in 1978[835] and through the 1980s he led its development into a powerful statistical tool.[818,836] Rubin originally developed the method to handle nonresponse in surveys and, wanting to avoid the problems associated with single imputation, he developed this method that instead replaced missing values with a distribution of two or more likely values.[837]

The MI procedure is based on two separate models: the *substantive model*, which is used, in effect, in the complete-case analysis of the filled-in data following imputation of the missing values, and upon the results of which, inferences will be made; and the *imputation model*, from which the distribution of the missing data given the observed data is derived.[801,834] One advantage of MI is that the imputation model can potentially contain variables that are predictive of missingness but not causal (if the substantive model is causal),[801] often called *auxiliary variables*.[822]

There are three steps involved when using MI. The first step uses a Bayesian approach to create multiple copies of the dataset. During this procedure, missing values are replaced with values randomly sampled from the posterior distribution produced using the imputation model.[781,801] The process is then repeated until the desired number of datasets have been created. The second step then involves fitting the substantive model to each of the completed datasets using standard methods of analysis, with the resulting estimates combined in the third step using simple rules (Rubin's rules)[822] to produce a final estimate. This includes a standard error that acknowledges the uncertainty implied by some participant values not being known.[605,781,801] A curiosity of MI is that it involves the combination of a Bayesian step with a frequentist step.[801]

Multiple imputation by chained equations (MICE)

The first step is often the more difficult one and depends on an appropriate imputation strategy and model being chosen.[605,822] A common strategy or method to use when the missingness pattern is non-monotone, and when a combination of categorical and continuous variables have missing values, is called *multiple imputation by chained equations* (MICE),[605,822,838] or sometimes *multivariate imputation by chained equations* (MICE).[820] It is also called *fully conditional specification*.[839] The MICE procedure involves the fitting of a series of regression models where each variable with missing values is modelled separately, conditional upon the other variables in the imputation model. This allows continuous variables to be modelled using linear regression while binary variables can be modelled using logistic regression, and so forth.[820]

Model checking

The implementation of MI in modern software packages still has many limitations, including the absence of procedure features that can be used to check imputation models.[822] Nevertheless, tables and graphs have been recommended for a number of years as one way to check that the imputation model has produced 'reasonable' values.[820,822] But rather than act as formal statistical tests, they can instead act as flags that can indicate when there may be problems with the imputation model that needs checking.[840]

7.3 Results and Discussion

7.3.1 Enrolment and device measurements

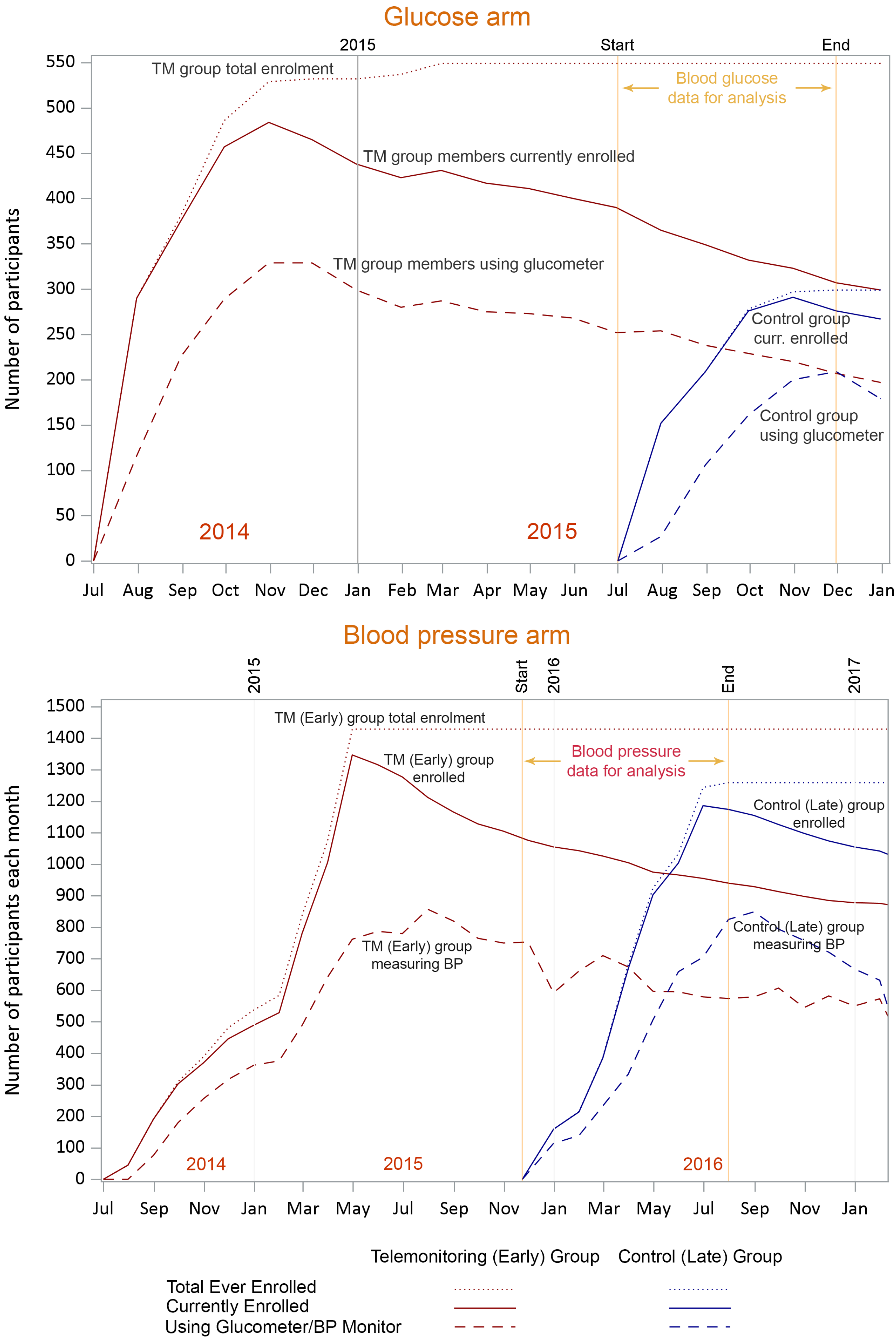
For both arms, collection of Telemonitoring device blood glucose or blood pressure measurements occurred following enrolment and at the discretion of the participant. The flow of participants can be seen in the Appendix. Figure 7.2 displays the number of participants in each group that were currently enrolled each month and the number who recorded at least one measurement during that month. The cumulative total enrolment is also shown. A table in the Appendix provides more detailed participant numbers and percentages for the Glucose arm. Figure 7.2 also shows that around a third of currently enrolled participants did not use the Telemonitoring device in any one month, even with non-adhering participants steadily dropping out.

To check the range of values recorded, the distribution of all Telemonitoring device BP measurements was examined (see Appendix), while the distributions of both observed and imputed blood glucose, HbA1c and BMI outcome values, following multiple imputation, can be seen later in this chapter.

Health (hospitalisation) risk level and order of enrolment

The participants were randomised both to when they would be invited (Early or Late period) as well as the order in which they were to be invited. The My Health Guardian service provided by Healthways operates so that participants are contacted more frequently when their health is worse, as judged by the custom measure of health status labelled ‘risk level’ in the analysis dataset. Because adhering to a randomised ordering of people to invite might have conflicted with normal service delivery, we checked for evidence of a non-random ordering of invitations in the analysis of the BP arm. This was done by plotting the participant’s ‘risk level’ at the time of enrolment against the date of enrolment, for each group, and calculating a Lowess line of best fit. Suspicion that the enrolment order might not have been followed did not occur until after the Glucose arm analysis was complete.

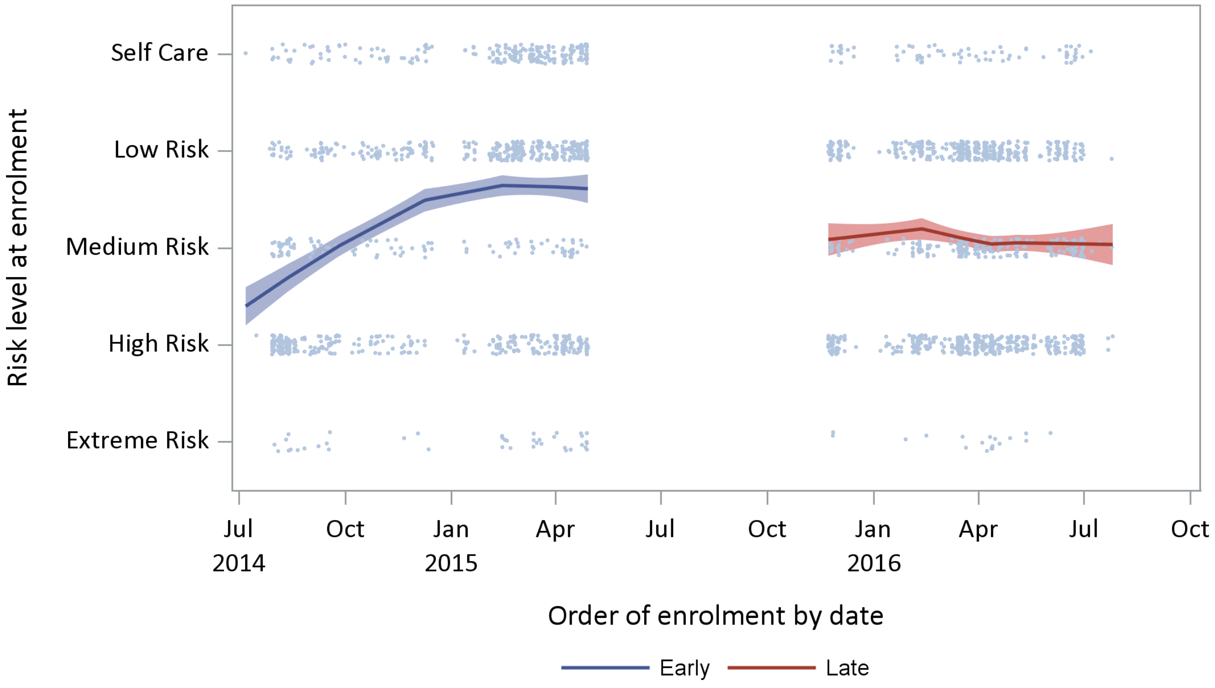
Figure 7.2 Trial enrolment and number of participants measuring by month



In Figure 7.3, Telemonitoring group participants in the BP arm who were enrolled early appear to have been less healthy, on average, than those who enrolled later. This suggests that a randomised ordering of enrolment did not occur for this group. Such a pattern was not apparent with the Control group, however. Initial enrolment of the Telemonitoring group may have been balanced by the better than average health of participants who were enrolled toward the end, suggesting that the specific group the members were randomised into was nevertheless adhered to.

With people of worse overall health starting earlier, on average, it may have increased the chance that participants with higher blood pressure would have dropped out by the time of the analysis period. It is unknown whether this occurred or whether average BP outcomes were affected.

Figure 7.3 BP: Risk level at the time of enrolment Lowess lines of best fit



Note: Early = Telemonitoring group; Late = Control group

To assess this possibility, Table 7.1 lists the correlations between risk level, dropout and BP outcome. The correlation between the risk level at 1 July 2014 and the risk level at the time of enrolment suggests that the risk level varies quite a bit over time. Neither risk level appears to relate to the likelihood that the participant would withdraw from the trial or stop using the Telemonitoring device before at least 7 months (approx. 208 days). Likewise, neither risk level was strongly correlated with the outcomes measured in the analysis period of 23 Nov 2015 to 31 Jul 2016, though less than half of the participants were able to be included in this calculation. This suggests that the non-random ordering of enrolment is unlikely to have had a significant impact on the outcomes measured.

Table 7.1 BP: Correlations between risk level, dropout and BP outcome

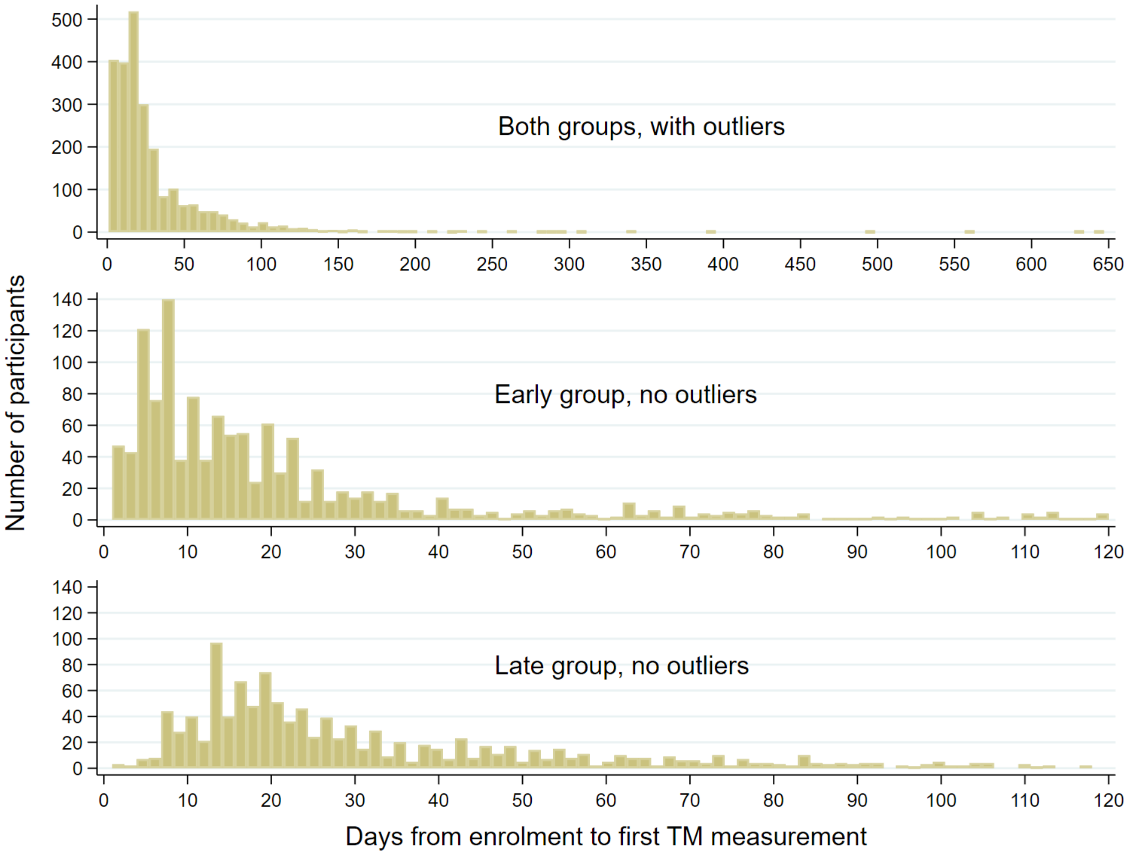
		Risk level at enrolment	No analysis outcome (dropped out)	Mean SBP outcome	Mean DBP outcome
TM (Early) group					
Risk level at 1-Jul-14	*r**	*0.41*	*-0.002*	*-0.11*	*-0.02*
	P	<.0001	0.96	0.01	0.55
	N	1039	1039	624	624
Risk level at enrolment	*r*		*-0.04*	*-0.09*	*0.04*
	P		0.25	0.03	0.31
	N		1039	624	624
Controls (Late) group					
Risk level at 1-Jul-14	*r*	*0.20*	*-0.04*	*-0.05*	*-0.03*
	P	<.0001	0.13	0.09	0.39
	N	1158	1158	968	968
Risk level at enrolment	*r*		*-0.01*	*-0.07*	*-0.01*
	P		0.70	0.02	0.81
	N		1158	968	968

* r = Pearson correlation coefficient; P = P-value; N = number of participants

Time from enrolment to first measurement

Another possible concern with enrolment is the gap in time between the enrolment date and the date on which participants first used the glucose or BP measuring device. In the Glucose arm there was a mean of 34 days and median of 20 days between enrolment and the first device measurement. For the BP arm, the problem of non-adherence is illustrated in Figure 7.4, which shows the large variation in time between enrolment and many participant's first measurement.

Figure 7.4 BP arm days between enrolment and first telemonitoring (TM) measurement



Note: Early group = Telemonitoring group; Late group = Control group

Time of day that measurements were taken

Glucose arm

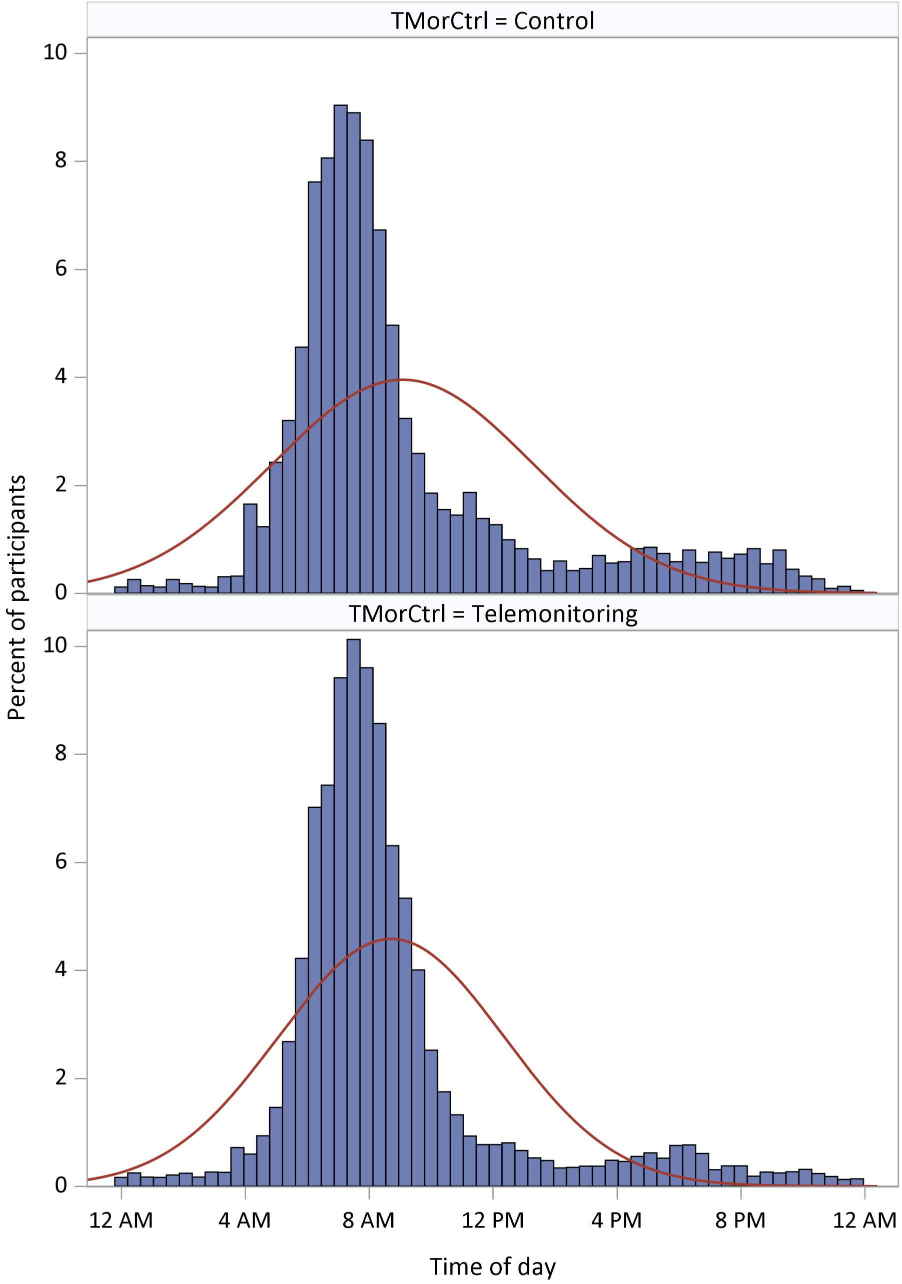
While it was not known if blood glucose measurements were taken after fasting in this trial, the first measurement of any day seems the one most likely to be taken after fasting, and it is a common recommendation made to people with diabetes. Therefore, with some participants expected to record multiple blood glucose measurements on some days, only the first measurement from such days was used in the analysis. We do not know how often multiple measurements were carried out, however, so the number of measurements taken at each time of the day for all first-in-day blood glucose measurements was examined to enable visual assessment.

For the Glucose arm, Figure 7.5 shows the distribution of the time of day that first-in-day blood glucose measurements were taken, displayed for each group. These are from the

435/549 (79%) of the Telemonitoring group and 256/299 (86%) of the Control group with at least 2 measurements recorded. No clear difference is apparent.

One potential concern in comparing the Telemonitoring and Control groups is that the range of blood glucose measurements that make up each individual's mean blood glucose, might not be equal in terms of the time-of-day they were taken (later in the day measurements are less likely to be after fasting), or date of the year (in case the treatment changes over time in subtle but influential ways).

Figure 7.5 Glucose arm time of day of all first-in-day blood glucose measurements



BP arm

With the BP arm, to see whether a pattern was apparent in Telemonitoring device measurements by hour of day, and whether any differences could be seen between the Early and Control groups, some penalized B-spline lines of best fit are shown in Figure 7.6. Blood pressure is known to drop at night for most people, though this can be blunted in people with hypertension, however, only a small number of measurements were recorded at this time so to avoid distorting the lines of best fit, they were constructed using only measurements from 6 AM to 11 PM.

A small drop is suggested around 2 PM and 9 PM and may relate to a postprandial BP drop after lunch and dinner. The most obvious difference between the groups is a higher morning blood pressure in the Control group. However, all available Telemonitoring device measurements were used to construct Figure 7.6, so the Telemonitoring group contains a greater proportion of measurements from those people who continued to measure for many months. The Telemonitoring group measurements also include some 12 months or more after enrolment, the maximum length of time most Control group participants could have contributed measurements. This means the Telemonitoring group BP device data contains more measurements collected after a longer period of exposure to the intervention, as well as more measurements from the type of people who, for many reasons, decided to continue measuring.

To look at more comparable groups, we restricted the Telemonitoring group measurements to those in the first 12 months only (until 31 July 2015) and constructed Figure 7.7. This shows a more similar pattern between the groups.

Figure 7.6 BP arm time of day of BP measurements and relationship to blood pressure

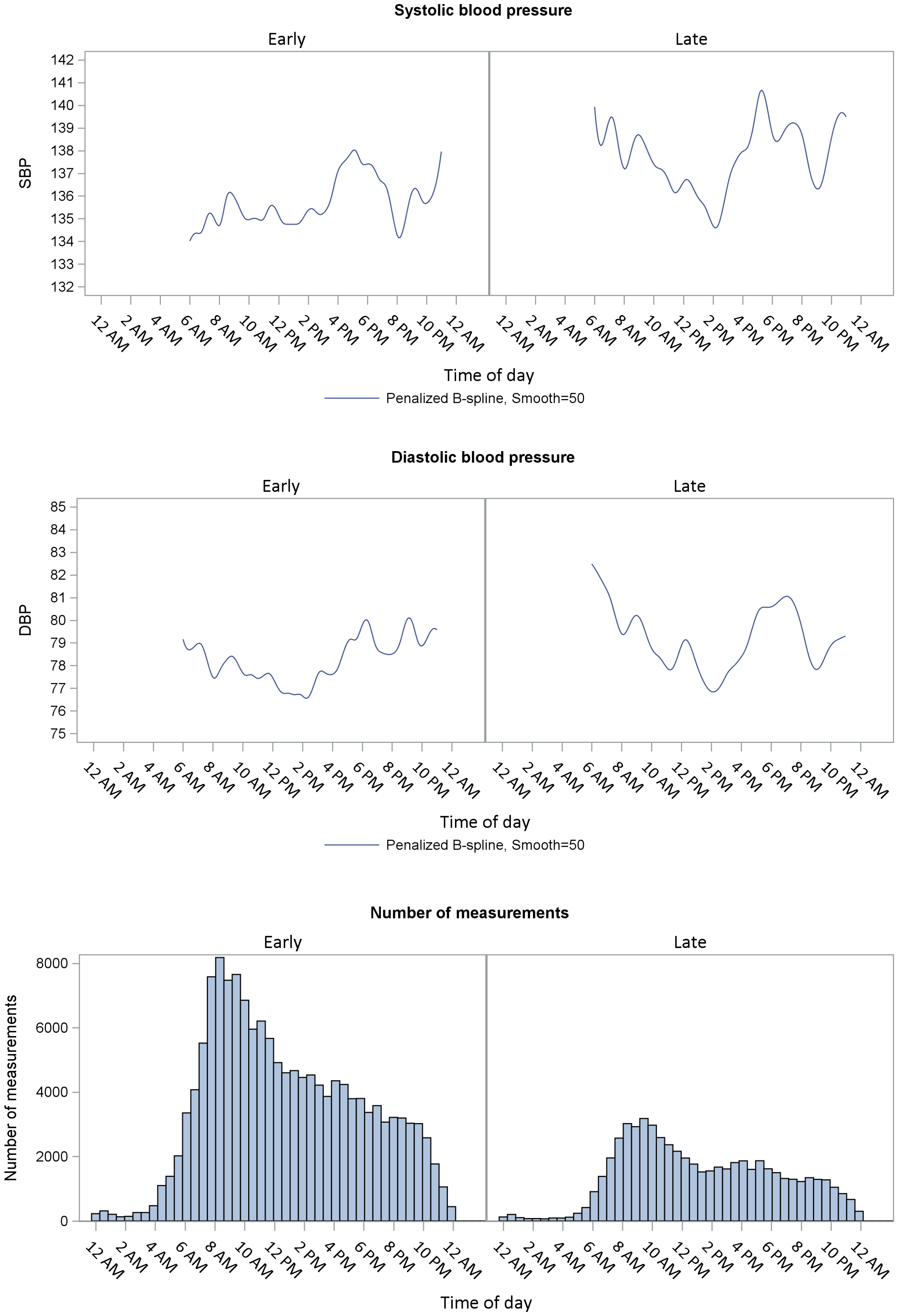
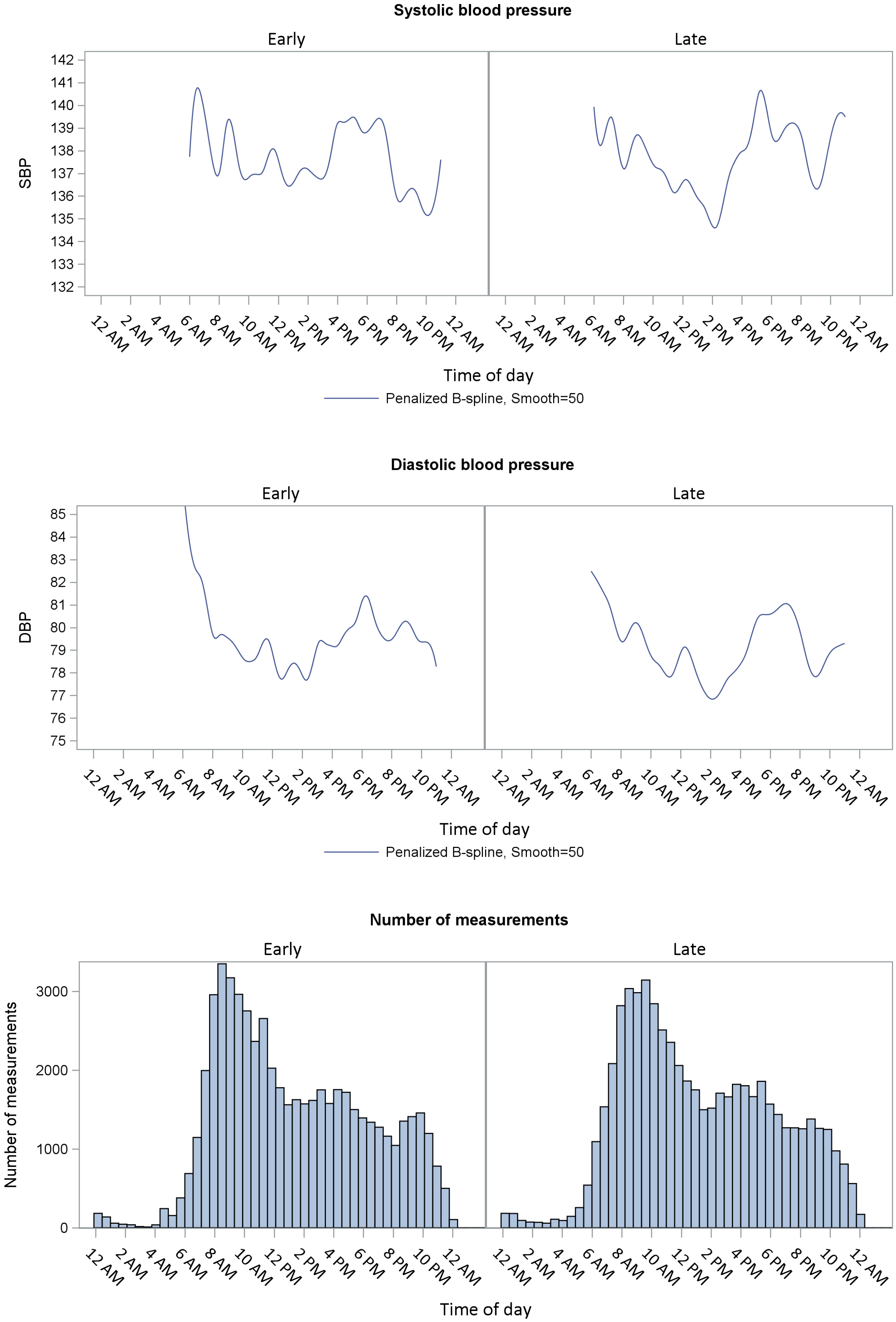


Figure 7.7 BP arm time of day of measurements and BP – initial 8-12 months only



Systolic blood pressure

Diastolic blood pressure

Number of measurements

7.3.2 Outcome data availability and definitions

Glucose arm

The amount of outcome data that was available for analysis in the Glucose arm is shown in Table 7.2. HbA1c, the original primary outcome, was available for just 25% of the participants, though this was only if we used HbA1c results recorded between April and December 2015, a fairly wide time window. A narrower time window would have reduced the number of participants with HbA1c results even further (not shown). And if, to avoid problems with regression to the mean, we had used our preferred criteria of only participants with baseline values available, then HbA1c measurements would have come from only 10% of those enrolled. Thus, the large majority of participants would have been missing an outcome, with some of the reasons for not having a value possibly related to the HbA1c result they would have provided had they arranged for the blood test. For example, age, employment status, and psychological and motivational factors might all increase or decrease the chance that someone would get the blood test done, and all might plausibly relate to an HbA1c level that was different to the group average. In this case, an analysis might have provided biased estimates if the available HbA1c results differed from the results missing, and one group had more of their results missing. This scenario was plausible because the HbA1c measurements in the Telemonitoring group did not include those from participants who dropped out early, unlike the Control group's measurements which were recorded before any dropouts occurred.

While body mass index (BMI) had the most complete data of the originally specified outcomes, from 38% of participants if baseline values were required, the plausible outcome with the most data available was blood glucose recorded using the telemonitoring enabled glucometer. But as these were from using the intervention, this was far from ideal because it would mean that the outcome was only available from participants who used the intervention — and many did not — introducing possible bias into an intention-to-treat analysis that would, by definition, include all participants. It also presented a problem with the Control group's outcome values because they would be recorded following exposure to the intervention, albeit for a much shorter time than for the Telemonitoring group.

Table 7.2 Glucose arm outcome data available with possible definitions

Chosen definitions used for the analysis are highlighted in red

		Glucose participants with data		
Outcome variables	Date range of ≥1measurements	TM* N=549	Controls N=299	Total N=848
HbA1c	1 Apr - 31 Dec 2015	113 (21%)	96 (32%)	209 (25%)
HbA1c	1 Apr - 31 Dec 2014 and 1 Apr - 31 Dec 2015	62 (11%)	22 (7%)	84 (10%)
BMI	1 Jul 2013 to 30 Jun 2014 and 1 Apr to 31 Dec 2015	171 (31%)	151 (51%)	322 (38%)
Glucometer blood glucose	1 Jul - 31 Aug 2015	264 (48%)	107 (36%)	371 (44%)
Glucometer blood glucose	1 Jul - 30 Sep 2015	269 (49%)	173 (58%)	442 (52%)
Glucometer blood glucose	1 Jul - 31 Oct 2015	270 (49%)	227 (76%)	497 (59%)
Glucometer blood glucose	1 Jul - 30 Nov 2015	271 (49%)	263 (88%)	534 (63%)
Glucometer blood glucose	1 Jul - 31 Dec 2015	276 (50%)	267 (89%)	543 (64%)

* Telemonitoring group

An additional problem was that the results may not then have been applicable to the participants in the Telemonitoring group who either didn't use the glucometer or didn't use it for long enough to have measurements recorded in the selected time window of 1 July 2015 to 30 November 2015. This would have derived from needing to either restrict the analysis to those participants with outcome values, as well as all other covariates; or use some form of imputation, in which case, the data would still be based on the outcomes recorded, though the use of multiple imputation would make use of the recorded covariate data, as well, if used. Nevertheless, we decided that there would be a lower chance of bias if the primary outcome was changed from HbA1c to one of the definitions of mean blood glucose, though we still retained HbA1c and BMI as additional outcomes.

After many time windows were considered, those we chose for the definitions of each outcome are highlighted in red in Table 7.2. There were many ways we could have defined and calculated a mean blood glucose level for each participant at the end of the trial, with the superior definitions not immediately clear. But after some experimentation, and with group comparisons at times, unavoidable, the two outcomes used in analyses 1 and 2 were chosen and are summarised in Table 7.3, along with equivalent details for the HbA1c and BMI analyses. Although using a prespecified analysis plan was not feasible in this case, due to the level of missing data and, in particular, uncertainty about how best to define and analyse the available data, we feel that our focus on understanding the sources of uncertainty and potential bias and the desire to communicate this to the stakeholders, will have compensated for any motivation to produce a particular result. Nevertheless, we hope to describe the analysis process in a way that would reveal the possible influence of cognitive biases discussed in Chapter 4.

Analysis 1

The first outcome choice, the mean of many measurements, had the advantage of more accurately reflecting the mean blood glucose level for each participant, at least compared to a mean of only a few measurements or just a single measurement. The disadvantage, in this case, was that the mean for the Control group came from measurements taken after the participant had been exposed to the intervention, albeit for a much shorter time period. We attempted to compensate for this in analysis 2.

Analysis 2

Ideally, we would have preferred to compare the outcome for Telemonitoring group participants at the end of the trial with the outcome measured for Control participants just before they start, with one group fully exposed to the program and the other group having no exposure. But because the outcome, in this case, was measured using the health intervention being studied, some exposure of the Control participants was unavoidable. Thus, we needed this exposure to be as small as possible, however, it was not clear just how small the exposure could be that would still provide a reasonable approximation to the participant’s mean blood glucose level.

Initially, a single measurement for each group was used, with the first measurement taken by Control participants compared to either the last or middle measurement taken by Telemonitoring participants. But when measurements close in time were used instead for each participant’s value, the result of the comparison varied noticeably, depending on which particular single measurements were chosen. A number of factors may have accounted for this, such as the type and timing of a meal, exercise, or glucose lowering medication. In some cases, the first-in-the-day measurement that we used would have been after fasting, but while this is recommended by medical practitioners, we do not know how many of the measurements were before breakfast. Instead, as a compromise between avoiding the variability of single measurements while minimising the exposure that Control participants had to the intervention, we decided to use 5 measurements using the criteria in Table 7.3. It is worth noting, however, that individual blood glucose trajectories were quite variable in most cases (see the Appendix for some individual blood glucose trajectory examples), suggesting many factors may play a role in the blood glucose level that was measured.

The distribution of measurement dates, using this criterion, is shown in **Figure 7.8**. The first measurement dates in the Control group are well spread over the time period July to November 2015, owing to the progressive enrolment of participants and variations in time between enrolment and first use of the glucometer. The Telemonitoring group participants, on the other hand, mostly had regular measurements throughout this period, so to make the distribution of measurement dates as similar as possible, we chose to select the 5 measurements closest to the middle of this period.

Figure 7.8. Distribution of measurement dates for Analysis 2

First 5 measurements for Control participants and middle 5 for the Telemonitoring group

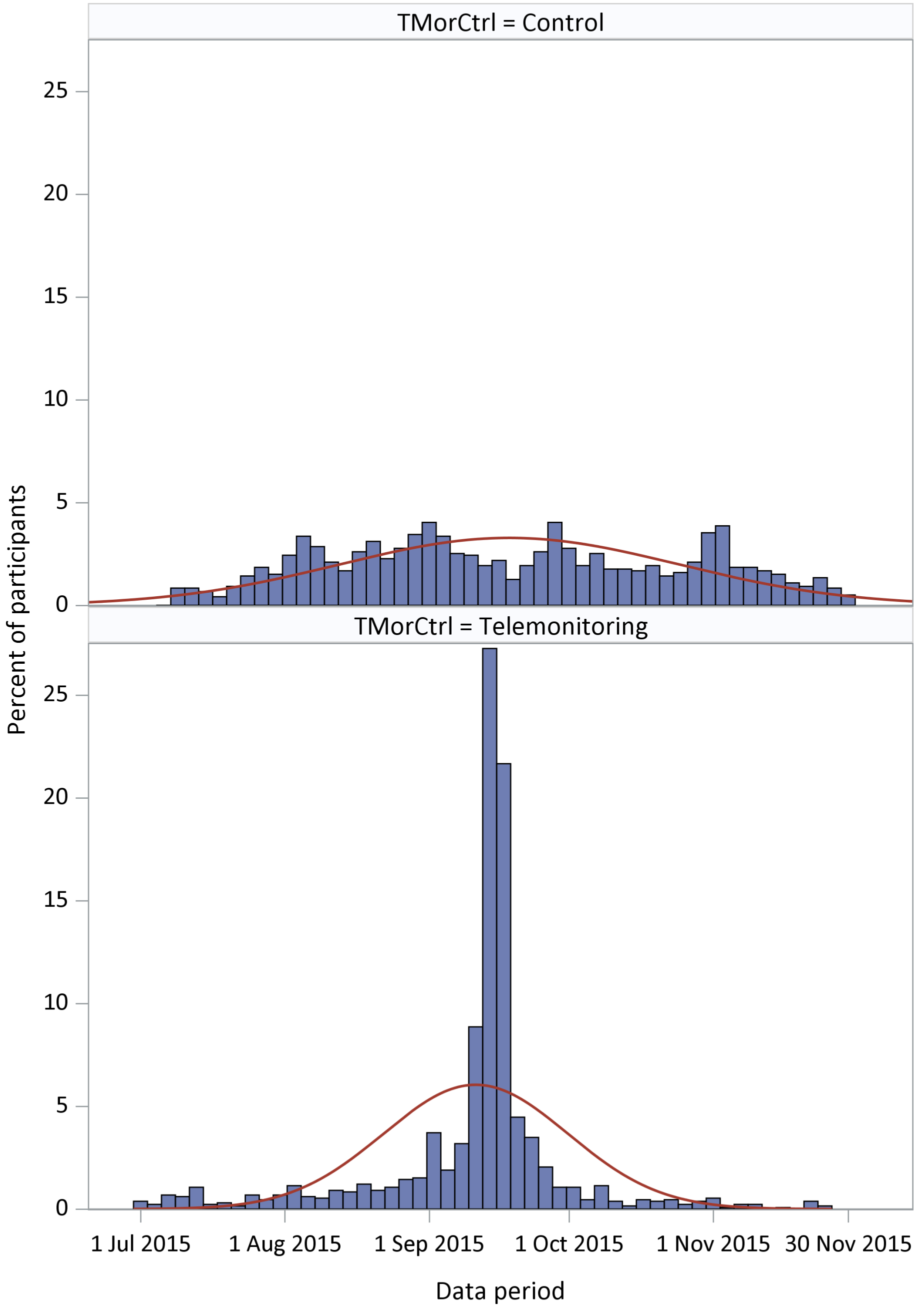


Table 7.3 Glucose arm outcome definitions for group comparisons

	Analysis 1	Analysis 2	Analysis 3	Analysis 4
Participants	All*	All*	All*	All*
Measurements used to create outcome†	All first-in-day blood glucose 1-Jul-15 to 30-Nov-15	First-in-day glucose 1-Jul-15 to 30-Nov-15 satisfying below	Any HbA1c 1-Apr-15 to 31-Nov-15	BMI from height and • Last weight 1-7-13 to 30-7-14 (baseline) • First weight 1-04-15 to 31-12-15
Telemonitoring outcome	Mean blood glucose	Mean of middle 5 measurements → closest to 15-Sep-15‡	HbA1c (mean if more than one)	Change in BMI from baseline
Control outcome	Mean blood glucose	Mean of the first 5 measurements from 1-Jul-15 to 30-Nov-15	HbA1c (mean if more than one)	Change in BMI from baseline

* all randomised participants following multiple imputation; † outcome measurements included before multiple imputation; ‡ mean of 5 measurements closest to 31-Mar-2016 (middle of 23-Nov-2015 to 31-Jul-2016) - approximate middle of analysis period

BP arm

Summarised in Table 7.4 are the three variations of the group comparison analysis for the BP arm, varied to reduce our reliance on one choice of model. The methodology used, and results are reported in Chapter 8.

Analyses 5 and 6

Analysis 5 (numbering is continued) repeated the method we considered the most valid from the Glucose arm analysis. In Analysis 5, it was plausible that all 5 measurements could have been from the one day. The only difference between Analysis 5 and Analysis 6 was that when multiple measurements were recorded on any one day, in Analysis 6 those measurements were averaged so that participants had at most one measurement per day.

All participants were used in analyses 5 and 6 following multiple imputation, and as such, they provided intention-to-treat estimates. However, considerable uncertainty remained because these estimates were based on the strong assumption that the missing data were MAR or MCAR. Before we used multiple imputation, 38% of the Telemonitoring group and 16% of the Control group had no BP measurements recorded in the analysis time window of 23 November 2015 to 31 July 2016. The Control group enrolled during this time window, so only participants who did not use the device were without BP measurements. However, participants in the Telemonitoring group enrolled between 7 and 16 months prior to 23 November 2015, so many had either withdrawn from the study or ceased using the device before the Control group began enrolling.

There are many plausible reasons that might explain why these participants effectively dropped out before they could provide outcome values, and some of these reasons may be related to their blood pressure levels so that, had they stayed in the trial, their blood pressure measurements in the analysis time window might have been different, on average, to those participants who stayed in the trial and ended up providing BP measurements for the analysis. This suggests the possibility that the intent-to-treat results from Analyses 5 and 6 may be biased because of, in effect, differential loss to follow-up.[829] If the results suggest a difference may exist in mean blood pressure between the groups, this possible bias means we should be less sure of whether the difference was caused by the Telemonitoring intervention, or by the dropping out of Telemonitoring group members with higher mean blood pressure, leaving an overall lower mean in the Telemonitoring group outcome values.

Analysis 7

With the actual blood pressure values of many participants who dropped out or stopped measuring unknown, there is no way to know what the mean blood pressure would have been if those participants had remained in the trial. Nevertheless, to try to increase the similarity, or exchangeability, of the Telemonitoring and Control groups, and hence reduce the potential influence of selection bias from dropout, in Analysis 7 we restricted the Telemonitoring group participants to those with outcome measurements recorded in the analysis time window, though we still used all of the imputed baseline values generated for this subgroup through multiple imputation. And we restricted Control group participants to those who similarly went on to record their blood pressure for at least as long as the shortest time a Telemonitoring group member with measurements in the analysis window had measured for. This was calculated as the shortest number of days between the last enrolment of an included Telemonitoring group member and the first Control group enrolment, which in this case was 208 days. In addition, to ensure that the participants' mean values were more likely to represent their mean blood pressure, the analysis was restricted to participants with measurements recorded on at least 5 separate days.

Outcome measurement distributions

For the four outcome measurements: blood glucose, HbA1c, BMI, and blood pressure, distributions are presented in the Appendix. These were created to assist with familiarity of the data and the units used for measurements.

Table 7.4 BP arm outcome definitions for group comparisons

	Analysis 5	Analysis 6	Analysis 7
Participants	All*	All	All who measured on at least 5 days, at least 208 days after enrolment
Blood pressure measurements†	Any from 23-Nov-15 to 31-Jul-16	Any from 23-Nov-15 to 31-Jul-16, but unlike analysis 5, daily averages were used instead of measurements	
Telemonitoring outcome	Middle 5 BP measurements‡	BP measurements from middle 5 days§	
Control outcome	First 5 BP measurements¶	BP measurements from first 5 days**	

* following multiple imputation; † included for multiple imputation; ‡ mean of the 5 measurements closest to 31 March 2016 (approximate middle of analysis period); § mean of the 5 days with measurements closest to 31 March 2016; ¶ mean of the first 5 measurements in the analysis period; ** mean of the first 5 days with measurements in the analysis period

7.3.3 Baseline characteristics

Some baseline data for both arms was collected prior to 1 July 2014 through the My Health Guardian program run by Healthways, and definitions of these covariates are included in Table 7.5 and Table 7.6 for the Glucose arm, and in Table 7.7 for the BP arm.

Tables in the Appendix show how diagnoses and medications were classified and derived from Healthways data. It should be noted that all participants had some diagnoses data so, for the purposes of the analysis, we assumed this information was complete; that is, we assumed that the absence of a certain diagnosis was not due to missing data but instead, due to that person not having the corresponding condition. It was also assumed that those participants without medication data were, in fact, not taking any. Given the number of participants, however, it is highly likely that errors and omissions exist in the Diagnoses and Medications data.

The level of baseline data missing for Glucose arm participants is shown separately in Table 7.5 to enable an initial assessment. This also indicates the quantity of data we needed to impute using multiple imputation.

Table 7.5 Glucose arm number of participants missing baseline data

Baseline data variable	Glucose participants missing data Telemonitoring N = 549	 Controls N = 299	 Total N = 848
Age (at 1 July 2014)	0	0	0
Sex	0	0	0
Ethnicity	115 (21%)	88 (29%)	203 (24%)
BMI (last weight recorded July 2013 - June 2014)	183 (33%)	98 (33%)	281 (33%)
HbA1c (last recorded Jul 2013 - Jun 2014)	396 (72%)	225 (75%)	621 (73%)
Diagnoses (onset before Jul 2014) (Diabetes type, Hypertension, Hyperlipidaemia, CVD, Arthritis, Back pain, Walking pain, Eye problem)	7 (1%)*	5 (2%)*	12 (1%)*
Medications (began taking before Jul 2014) (Insulin/Analogue; Diabetes drugs; Pain relief drug)	36 (7%)†	28 (9%)†	64 (8%)†
Employment status (before July 2014)	255 (46%)	155 (52%)	410 (48%)
Moderate exercise (before July 2014)	278 (51%)	144 (48%)	422 (50%)
Smoking status (before July 2014)	270 (49%)	141 (47%)	411 (48%)
Risk level (last recorded July 2013 - June 2014)	0	0	0

* No diagnoses recorded with onset before July 2014; † No medications recorded with start date before July 2014

Table 7.6 presents the demographic and observed baseline data of participants with at least one home blood glucose measurement from July to November 2015. We restricted participants in this table to those with outcome data in the analysis time window because it is the outcomes of these participants that formed the basis for missing data imputation when multiple imputation was used. Around half of the participants were missing important baseline information, such as whether they engaged in moderate exercise or were current or past smokers, while around three quarters did not have baseline HbA1c results.

This table also highlights a potential problem with an intention-to-treat analysis for the Glucose arm with outcome data available for 88% of the Control group but only 49% of the Telemonitoring group. As with the BP arm, the difference derives from the fact that Control group participants enrolled during the analysis time window, so those who measured at least once but later dropped out, did not have missing outcome data. Telemonitoring group participants, on the other hand, enrolled between 4 and 12 months before the analysis time window started, so many had either dropped out or stopped measuring by 1 July 2015. And some of the participants might have dropped out for reasons predictive of their blood glucose level, such as poor motivation to measure because they had not been taking their medication and did not want to see anticipated unfavourable glucose readings. If this was the case, then with more participants dropping out from the Telemonitoring group, the participants from each group with available data would no longer have been exchangeable due to selection bias from dropout, and the intent-to-treat estimates would be biased. And the missing data mechanism would likely have been MNAR because variable such as motivation were not measured.

Some differences are suggested by the range of p-values in Table 7.6, though with the number of tests conducted, some of the low p-values may be due to chance. Nevertheless, overall the differences suggest that the Telemonitoring group participants providing data were, on average, slightly older and not as healthy. This can be seen with the variables Age, Diabetes type, previous diagnosis of Hypertension, Hyperlipidemia, Cardiovascular disease or Arthritis, the prescription of a Pain relief drug, and the 'hospitalisation' risk level. To reduce potential confounding from the imbalances in the available data, which may carry over into the imputed data, these variables were incorporated into the regression models constructed in Chapter 8.

For the BP arm, comparison of the baseline characteristics between the intervention and control groups, detailed in Table 7.7, suggests some differences also existed. For Analysis 7, we listed the baseline characteristics separately given the restricted participant inclusion. But although we used this restriction in an attempt to make the comparison groups more exchangeable, in the end, if we judge by comparing the range of p-values between analyses 5 and 6 and analysis 7, this goal does not appear to have been achieved.

Table 7.6 Glucose arm baseline characteristics before multiple imputation

For participants with ≥1 home blood glucose measurement from 1 Jul to 30 Nov 2015; Some variable categories are not shown, with the full details in the Appendix.

Baseline characteristics	Telemonitoring N = 271 (49% of 549)	Controls N = 263 (88% of 299)	P-value
Sex, Male	169 (62% of 271)†	171 (65% of 299)†	0.530
Age, mean (SD)	68.8 (9.2)	65.7 (11.1)	0.001
Ethnicity, Caucasian (Missing: 22%)*	202 (87%)	165 (88%)	0.267
HbA1c, mean (SD) (DCCT %) (Missing: 73%)	6.7 (1.2)	6.8 (1.2)	0.944
BMI, mean (SD) (Missing: 31%)	30.5 (5.6)	30.4 (5.4)	0.838
Diabetes Type 2	248 (92%)	237 (90.5%)	0.074
Hypertension	157 (58%)	57 (22%)	< .0001
Hyperlipidemia	80 (30%)	56 (22%)	0.037
Cardiovascular disease	145 (54%)	107 (41%)	0.003
Arthritis (any type)	131 (48%)	100 (38%)	0.018
Back pain‡	55 (20%)	58 (22%)	0.672
Walking pain‡	48 (18%)	36 (14%)	0.235
Eye problem‡	34 (13%)	27 (10%)	0.418
Insulin or Analogue	45 (17%)	41 (16%)	0.814
Pain relief drug	155 (57%)	122 (46%)	0.015
Number of Type 2 diabetes drugs			
0 drugs prescribed	71 (26%)	92 (35%)	0.263
1 drugs prescribed	127 (47%)	113 (43%)	
2 – 4 drugs prescribed	73 (27%)	58 (22%)	
Employment status (Missing: 81%)			
Full-time, Part-time or Self-employed	8 (17%)	12 (23%)	0.734
No employment	15 (31%)	13 (25%)	
Retired	25 (52%)	28 (53%)	
Moderate exercise (Missing: 88%)	9 (22%)	4 (16%)	0.752
Smoking status (Missing: 45%)			
Never smoker	88 (58%)	89 (61%)	0.860
Past smoker	56 (37%)	50 (34%)	
Current smoker	7 (5%)	6 (4%)	
Risk level			
Extreme Risk	11 (4%)	11 (4%)	0.011
High Risk	63 (23%)	49 (19%)	
Medium Risk	17 (6%)	10 (4%)	
Low Risk	100 (37%)	77 (29%)	
Self-Care	80 (30%)	116 (44%)	

* missing from included participants (total=534); † % of non-missing; ‡ related diagnosis

Table 7.7 BP arm baseline characteristics before multiple imputation

	Analyses 5 & 6			Analysis 7		
Baseline characteristics	**TM** **N = 1,429**	**Controls** **N = 1,259**	**P**	**TM** **N = 773**	**Controls** **N = 617**	**P**
Sex, Male	727 (51%)	661 (52%)	0.40	426 (55%)	370 (60%)	0.07
Age, mean (SD)	70.6 (9.9)	69.1 (9.5)	<.0001	70.6 (9.1)	69.4 (9.0)	0.01
Ethnicity, Caucasian (m%)*	1,036 (73%) (21%)	809 (64%) (29%)	0.58	577 (75%) (19%)	424 (69%) (24%)	0.23
BMI, mean (SD) (m%)	29.4 (6.3) (38%)	29.3 (5.3) (38%)	0.74	29.2 (5.9) (37%)	28.8 (4.5) (35%)	0.38
Diabetes type 2	139 (10%)	145 (12%)	0.009	70 (9%)	46 (7%)	0.04
Systolic BP, mean (SD) (m%)	132.6 (13.7) (46%)	132.2 (13.2) (48%)	0.57	132.3 (13.4) (42%)	132.4 (13.2) (42%)	0.88
Diastolic BP, mean (SD) (m%)	75.1 (9.4) (48%)	76.0 (8.7) (49%)	0.08	75.0 (8.9) (43%)	76.2 (8.8) (44%)	0.06
Cholesterol, mean (SD) (m%)	4.5 (1.6) (92%)	4.5 (1.3) (93%)	0.80	4.4 (1.4) (90%)	4.4 (1.2) (91%)	0.92
Hyperlipidemia	504 (35%)	373 (30%)	0.002	283 (37%)	199 (32%)	0.09
Cardiovascular disease	616 (43%)	543 (43%)	0.99	359 (46%)	279 (45%)	0.65
Arthritis (any type)	712 (50%)	562 (45%)	0.007	393 (51%)	295 (48%)	0.26
Back pain	342 (24%)	257 (20%)	0.03	196 (25%)	132 (21%)	0.08
Walking pain	166 (12%)	147 (12%)	0.96	91 (12%)	88 (14%)	0.17
Eye problem	159 (11%)	107 (9%)	0.02	89 (12%)	55 (9%)	0.11
Insulin or Analogue	229 (16%)	164 (13%)	0.03	113 (15%)	85 (14%)	0.66
Pain relief drug	801 (56%)	580 (46%)	<.0001	447 (58%)	318 (52%)	0.02
Employment status (m%)	(46%)	(53%)		(45%)	(45%)	
Full-time	69 (5%)	57 (5%)	0.34	33 (4%)	36 (6%)	0.03
Part-time	47 (3%)	50 (4%)		25 (3%)	32 (5%)	
Self-employed	43 (3%)	26 (2%)		25 (3%)	16 (3%)	
No employment	409 (29%)	295 (23%)		241 (31%)	156 (25%)	
Retired	211 (15%)	168 (13%)		104 (13%)	99 (16%)	
Moderate exercise (m%)	388 (27%) (51%)	345 (27%) (57%)	0.004	223 (29%) (50%)	211 (34%) (50%)	0.004
Smoking status (m%)	(56%)	(62%)		(55%)	(56%)	
Never smoker	380 (27%)	299 (24%)	0.82	218 (28%)	163 (26%)	0.33
Past smoker	231 (16%)	178 (14%)		122 (16%)	108 (18%)	
Current smoker	12 (0.8%)	7 (0.6%)		6 (0.8%)	2 (0.3%)	
Risk level (m%)	(6%)	(7%)		(6%)	(6%)	
Extreme Risk	68 (5%)	35 (3%)	<.0001	35 (5%)	11 (2%)	0.009
High Risk	284 (20%)	178 (14%)		140 (18%)	87 (14%)	
Medium Risk	102 (7%)	107 (9%)		59 (8%)	54 (9%)	
Low Risk	496 (35%)	467 (37%)		281 (37%)	243 (39%)	
Self-Care	393 (28%)	378 (30%)		210 (27%)	184 (30%)	

* missing %

7.3.4 Causal diagrams

Following an assessment of missing data and determination of definitions to be used for the interventions, outcomes and covariates, the causal diagrams in Figure 7.9 and Figure 7.10 were constructed.

Figure 7.9 Causal diagram for the Glucose arm blood glucose outcome group comparisons

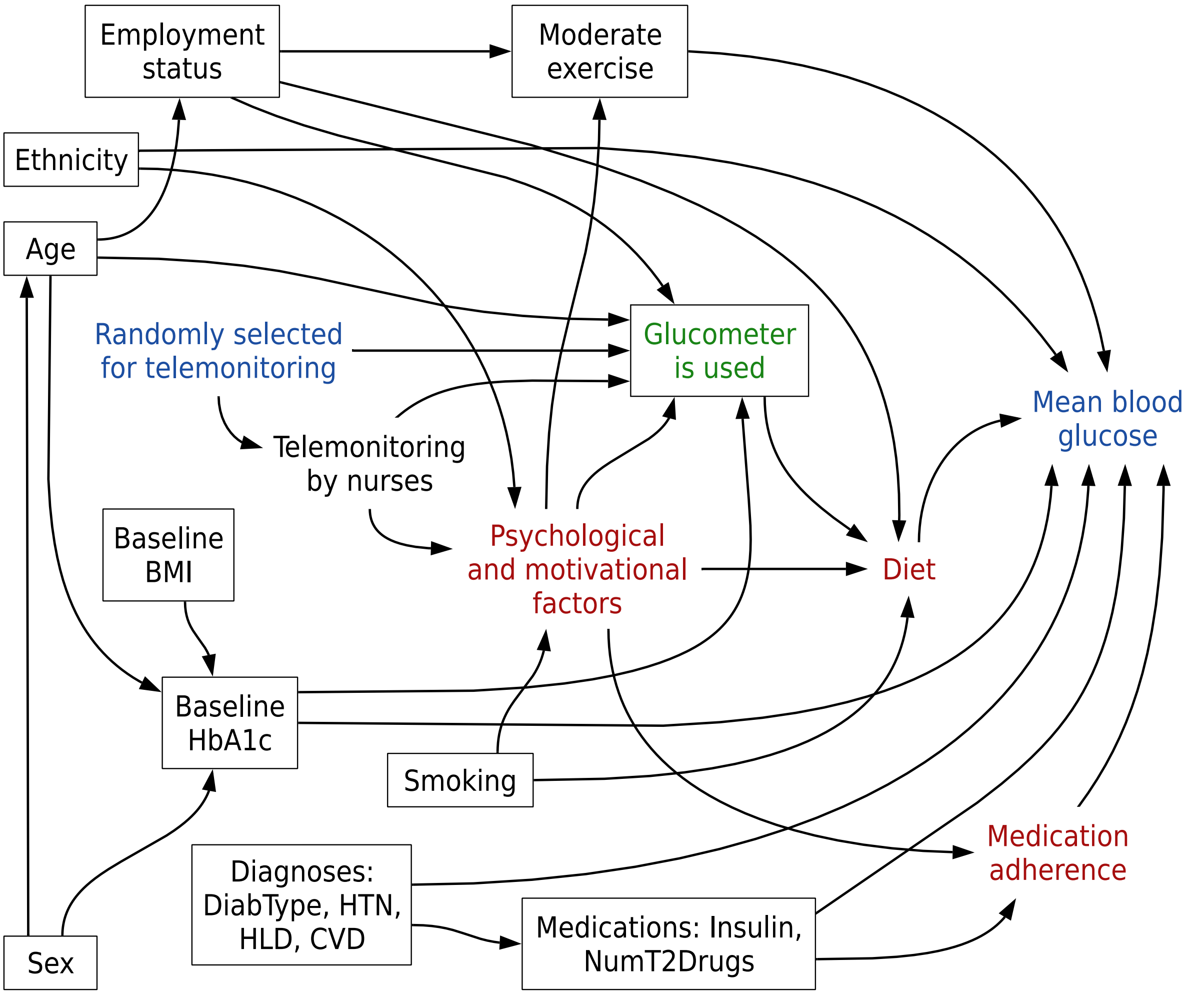


Figure 7.9 shows a directed acyclic graph (DAG) that includes both measured and unmeasured variables. Though it is initially complex to look at, this feature appeared to be an advantage when trying to convey the level of complexity to stakeholders. Nevertheless, a

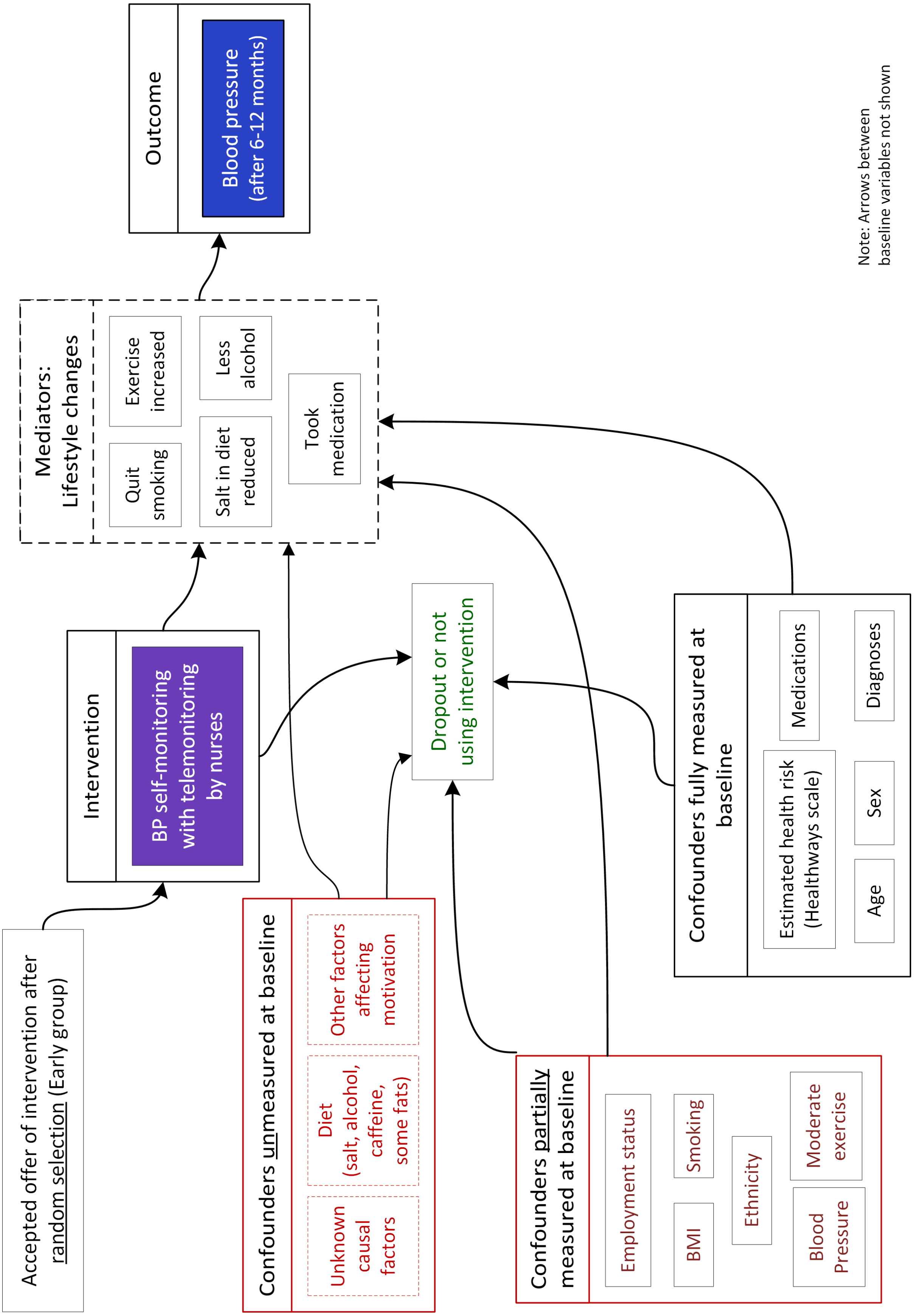
simpler version was also constructed, shown in the Appendix, and used as an example of a causal diagram in a Three Minute Thesis presentation.

Features to note are:

- variables that are conditioned on are surrounded by a box
- the intervention and outcome are coloured blue
- unmeasured variables are coloured red
- the green coloured "Glucometer is used" variable is conditioned on because there is missing outcome data, and this produces selection (collider) bias

In Figure 7.10, a simplified first step type of causal diagram is shown that was thought might have been an easier starting point for researchers not experienced in creating DAGs. Such a strategy might also appeal to some who do have such experience but find using the intermediate step easier.

Figure 7.10 A simplified first step causal diagram for the BP arm group comparisons


Outcome
Blood pressure (after 6-12 months)
Mediators: Lifestyle changes
Exercise increased
Quit smoking
Less alcohol
Salt in diet reduced
Took medication
Intervention
BP self-monitoring with telemonitoring by nurses
Dropout or not using intervention
Accepted offer of intervention after random selection (Early group)
Confounders unmeasured at baseline
Other factors affecting motivation
Diet (salt, alcohol, caffeine, some fats)
Unknown causal factors
Confounders partially measured at baseline
Employment status
Smoking
BMI
Ethnicity
Moderate exercise
Blood Pressure
Confounders fully measured at baseline
Medications
Diagnoses
Estimated health risk (Healthways scale)
Sex
Age

Note: Arrows between baseline variables not shown

7.3.5 Missing data patterns

To help understand the nature of the missing data mechanism, Table 7.8 was constructed using the SAS procedure PROC MI. As expected, the missing data pattern is clearly non-monotone.

Table 7.8 Glucose: Blood glucose missing data patterns

Each variable within each row is a mean or proportion; Overall blood glucose mean was 8.4

Pattern #	Partici-pants N = 534	Prop. of participants in TM group	mean Age	prop. Male	mean baseline HbA1c	mean baseline BMI	prop. with Hyper-tension	mean # Type2 Drugs	mean Employ-ment category	mean Mod-erate Exercise	mean Blood Glucose outcome
1	5	0.4	63	0.6	6.4	32.6	0.4	1.2	2.0	0.8	8.4
2	22	0.5	70	0.7	6.5	30.6	0.4	1.5	3.2	.	8.3
3	14	0.9	70	0.6	7.2	33.1	0.7	1.0	.	0.8	8.5
4	52	0.5	65	0.7	6.8	30.1	0.4	0.8	.	.	8.4
5	1	0.0	50	1.0	9.6	.	1.0	0.0	0.0	1.0	8.9
6	2	1.0	70	0.5	5.3	.	1.0	1.0	0.5	.	7.9
7	2	1.0	72	0.5	7.1	.	1.0	0.5	.	1.0	8.5
8	22	0.6	68	0.8	6.6	.	0.5	0.9	.	.	8.1
9	12	0.5	69	0.6	.	27.0	0.6	1.3	2.9	0.8	7.8
10	31	0.5	69	0.5	.	31.8	0.3	1.2	3.4	.	8.6
11	19	0.6	68	0.6	.	31.7	0.2	0.7	.	0.9	8.0
12	136	0.5	69	0.7	.	30.5	0.4	0.9	.	.	8.5
13	1	0.0	73	1.0	.	33.1	0.0	0.0	4.0	.	7.4
14	1	1.0	84	1.0	.	24.7	1.0	0.0	.	1.0	5.4
15	1	1.0	80	0.0	.	28.0	1.0	0.0	.	.	6.9
16	1	1.0	84	0.0	.	.	0.0	2.0	4.0	1.0	6.9
17	8	0.5	74	0.8	.	.	0.4	1.0	3.4	.	8.7
18	4	0.8	63	0.3	.	.	0.5	1.8	.	0.5	9.3
19	84	0.5	68	0.5	.	.	0.4	0.8	.	.	8.5
20	3	0.3	71	0.7	6.6	30.3	0.0	1.0	4.0	.	8.3
21	2	0.5	63	0.0	6.0	31.9	0.5	0.5	.	1.0	7.1
22	17	0.5	62	0.5	7.2	27.9	0.4	0.5	.	.	8.4
23	1	0.0	53	0.0	5.6	.	0.0	1.0	.	.	6.2
24	3	0.3	67	0.7	.	30.4	0.7	1.7	2.7	0.0	6.1
25	9	0.4	72	0.8	.	30.2	0.3	0.9	3.0	.	8.2
26	2	0.0	55	0.5	.	24.1	0.0	1.0	.	1.0	8.2
27	40	0.4	66	0.8	.	30.2	0.4	0.6	.	.	8.3
28	3	0.3	64	1.0	.	.	0.3	1.3	3.0	.	8.3
29	36	0.3	61	0.5	.	.	0.3	0.8	.	.	9.0

Table 7.8 was also used to help convey to HCF researchers and other interested staff how missing data increases uncertainty. Highlighting two rows as an example, we can see that the mean blood glucose outcome is 7.8 for one and 9.0 for the other. However, while the mean number of Type 2 diabetes drugs might help to explain the difference, 1.3 versus 0.8, mean BMI and the proportion engaging in moderate exercise is not known for one of the groups. Hence, we cannot know whether the difference in the outcome is because of the difference in the proportion using the intervention, 0.5 versus 0.3, or whether differences in the unknown variable values might have some influence. Thus, it increases the uncertainty.

7.3.6 Multiple imputation

The assessment of missing data for the Glucose arm revealed that only 5 participants out of 848 had complete data, so conducting a complete-case analysis with all covariates included in the model would not have been possible, even had we wanted to do so. This number can be seen as the top missing data pattern in Table 7.8. With some brief experimentation, we found that to use 120 (out of 848) participant's results, we would have needed to leave the following covariates out of the model: Employment status, Moderate exercise and baseline BMI. And to use 415 participants would have required the further dropping of baseline HbA1c. All of these are potential confounders and so leaving them out might have led to biased results.

All of the variables for which data was collected were included in the imputation model, including the outcome,[841] with the chained equations approach employed to create the imputations. For the Glucose arm, systolic BP, diastolic BP and total cholesterol were also included in the imputation model as auxiliary variables. With the aid of the causal diagram in Figure 7.9, these variables were not considered sufficiently plausible causes of blood glucose levels and so were not included in the substantive (analysis) model. Nevertheless, they were considered possible predictors of blood glucose through non-causal associations, and hence may have been able to contribute to the imputation model.

Initial experimenting with multiple imputation using SAS produced some variation in the estimates depending on the number of imputations specified and the seed number used (Table 7.9). The estimates and standard errors for each of the three seed numbers appeared to be quite similar when the number was increased to 100 imputations. However, with p-values of 0.05, 0.06 or 0.07, although essentially providing the same information about the estimate and the model, they may have been viewed differently by those who would see 0.05 as ‘significant’, and 0.06 and 0.07 as ‘not significant’. It also left open the question of exactly which estimate to report.

Another problem that we encountered when using PROC MI in SAS was that quite a few specific seed numbers led to errors when generated automatically for 200 imputations, with the 3 estimates listed in Table 7.9 some of the few where no error was encountered. Consequently, a SAS macro was created that combined the multiple imputed datasets produced from system generated seeds and 50 imputations, the number at which most PROC MI runs finished successfully.

With the goal to obtain stable estimates, the number of imputations was increased until the estimates remained relatively unchanged regardless of the seed number used. We speculate that the quantity of missing data may have been a factor that led to PROC MI having convergence problems when it tried to impute 100 or more sets of data. To avoid this problem, imputed data sets, each using a different seed, were combined into a dataset containing 1000 imputations. To check that this produced stable estimates, the procedure was run twice, and the estimates came out almost identical. But to account for slight differences, the final estimates were calculated as an average of the estimates from the two imputed data sets.

The BP dataset was analysed after the glucose analysis was complete and so, with the datasets very similar, we were able to take advantage of the experimentation already carried out. Perhaps with less missing data that needed to be imputed, with the BP dataset we found that 100 imputations were sufficient to generate stable estimates.

Table 7.9 Estimates, imputation number and seeds in SAS PROC MI for the Glucose arm

Estimates are for the effect of being in the Telemonitoring group on mean blood glucose

Imputations	**5**			**10**			**20**		
Seed	*264*	*5545*	*64728*	*264*	*5545*	*64728*	*264*	*5545*	*64728*
P-value	.003	.050	.037	.031	.048	.106	.035	.068	.061
Estimate	-0.48	-0.45	-0.37	-0.39	-0.43	-0.34	-0.37	-0.39	-0.36
Std. Error	0.16	0.21	0.17	0.18	0.21	0.20	0.17	0.21	0.19
Imputations	**50**			**100**			**200**		
Seed	*264*	*5545*	*64728*	*264*	*5545*	*64728*	*264*	*5545*	*64728*
P-value	.028	.070	.065	.056	.065	0.051	.057	.057	.062
Estimate	-0.42	-0.38	-0.38	-0.39	-0.38	-0.40	-0.39	-0.39	-0.38
Std. Error	0.19	0.21	0.20	0.20	0.20	0.21	0.21	0.20	0.20

7.3.7 Checking the imputation model

To check that the imputation model produced values that seemed reasonable, given both subject matter knowledge and the observed data, we compared the observed and imputed data for the Glucose arm using summary statistics and graphs. The distributions of observed and imputed categorical variables for Analysis 1 (Figure 7.11) appear mostly similar. Diabetes type was the only one that warranted checking, but a look at the corresponding numbers to this graph in Table 7.10 suggest it is only because very little data was missing for this variable. The other group comparison analyses for the Glucose arm (analyses 2, 3 and 4) produced very similar categorical variable comparisons.

Figure 7.11 Distribution of observed and imputed categorical variables for Analysis 1

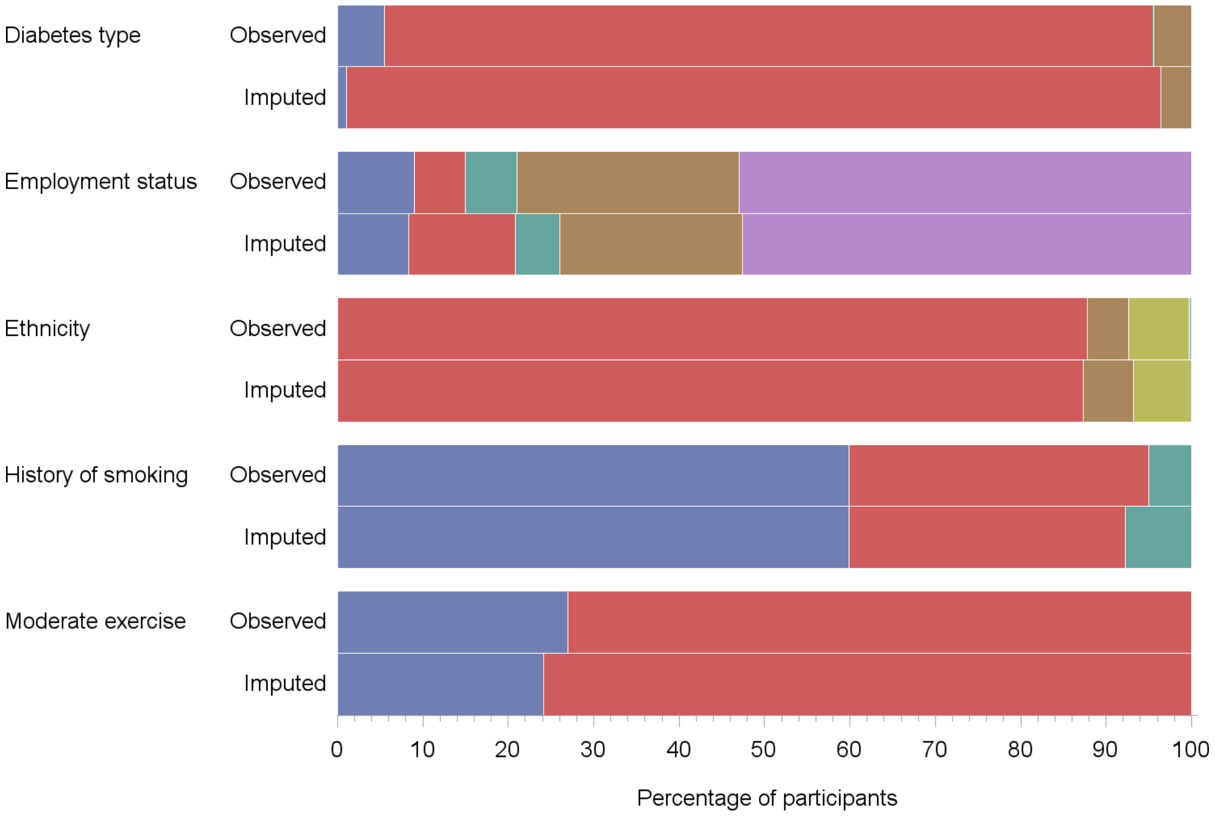


Table 7.10 Categorical variable proportions of observed and imputed values (Glucose arm)

Categorical Variable	Levels	Observed	Imputed
Diabetes type	Type 1	51 (6%)	0.04
	Type 2	760 (90%)	4
	Type 1 & 2	2 (0.2%)	0
	Other/unspecified	31 (4%)	0.1
Employment status	Full-time	13 (8%)	63 (9%)
	Part-time	13 (8%)	96 (14%)
	Self-employed	9 (6%)	35 (5%)
	No employment	42 (26%)	152 (22%)
	Retired	82 (52%)	344 (50%)
Ethnicity	Caucasian	563 (87%)	177 (87%)
	Asian	36 (6%)	13 (7%)
	Other	42 (7%)	12 (6%)
Smoking status	Never smoker	261 (60%)	245 (60%)
	Past smoker	156 (36%)	135 (33%)
	Current smoker	20 (5%)	31 (7%)
Moderate exercise	Yes	68 (76%)	597 (79%)
	No	21 (24%)	162 (21%)

Figure 7.12 Distributions of observed and imputed continuous variables – Glucose arm

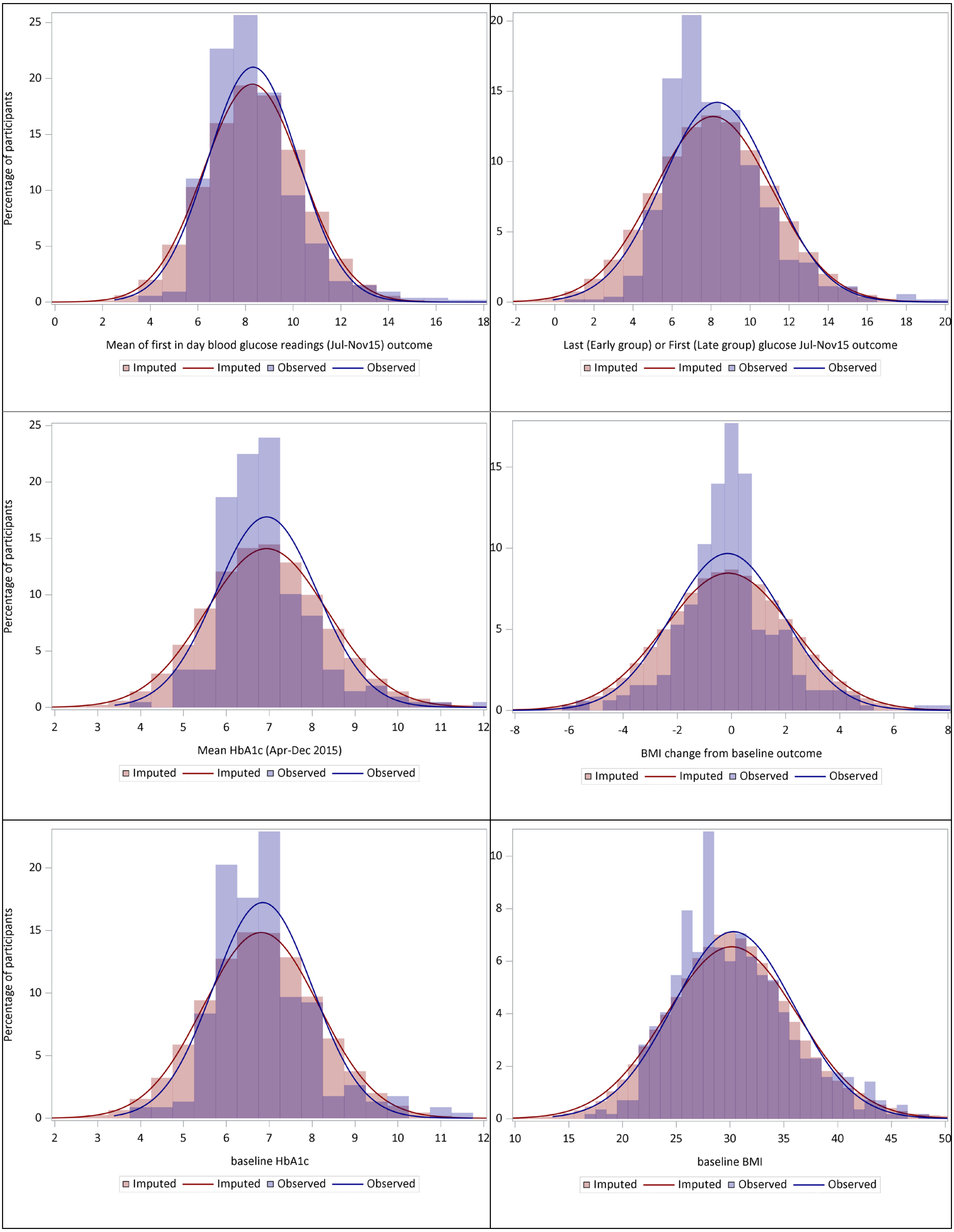


Table 7.11 Continuous variable means for observed and imputed values (Glucose arm)

Variable	Data Source	Mean	Std. Dev	Min	Max	N in dataset	N per imputation
Baseline Variables							
HbA1c	Observed	6.85	1.16	4.0	11.4	227000	227
	Imputed	6.81	1.34	-1.1	14.5	621000	621
BMI	Observed	30.32	5.60	17.3	56.9	567000	567
	Imputed	30.16	6.09	-6.4	59.3	281000	281
Systolic BP	Observed	130.99	14.85	75.0	192.0	365000	365
	Imputed	129.82	16.53	32.0	257.0	483000	483
Diastolic BP	Observed	74.67	9.67	44.0	129.0	365000	365
	Imputed	75.42	11.07	15.0	164.0	483000	483
Outcome variables							
Mean blood glucose	Observed	8.32	1.90	3.9	18.8	534000	534
	Imputed	8.29	2.05	-1.7	20.4	314000	314
Last/first glucose	Observed	8.32	2.81	1.1	23.0	534000	534
	Imputed	8.11	3.02	-6.4	28.1	314000	314
HbA1c	Observed	6.94	1.18	4.2	13.6	209000	209
	Imputed	6.94	1.42	-1.4	16.1	639000	639
BMI	Observed	30.47	6.08	17.2	61.2	465000	465
	Imputed	29.74	6.03	-13.3	58.2	383000	383
BMI change*	Observed	-0.14	2.06	-9.6	8.0	322000	322
	Imputed	-0.12	2.36	-23.9	13.7	526000	526

* BMI change marked as imputed if either the baseline BMI or the outcome BMI was imputed

The continuous variables are compared next for the Glucose arm (Figure 7.12), and this time the observed and imputed values for all four group comparison analyses are displayed. None of them, however, were sufficiently different to suggest something was wrong with the model. This was likewise the case for the corresponding numbers in Table 7.11.

One thing worth noting, however, are the minimum imputed values of some of the variables in Table 7.11, which are negative. Such values for the observed data would not be possible for any of these continuous variables, and an attempt was made, using the options of PROC MI, to restrict the imputed continuous values to realistic ranges. But with each of many attempts leading to a “floating point error” within SAS, the goal was eventually abandoned. It now appears, however, that restricting the range of imputed values may do more harm than good, with a simulation study by Rodwell et al. (2014)[842] finding that restriction techniques can result in bias with highly skewed data.

7.3.8 Exchangeability between groups and between those missing or not missing the outcome

Although designed as a pragmatic randomised controlled trial, factors that have led to considerable missing data have likely made the participants with available data less exchangeable, between those in the Telemonitoring group and those in the Control group, compared to the original randomised sample. This can be seen in the baseline characteristics before MI was used (Table 7.6 and Table 7.7), and it could also be seen following MI (Table 7.12), though not always with the same variables.

In terms of the missing data mechanism, in Table 7.13, we have the baseline characteristics for participants in the Glucose arm, separated into participants with the outcome observed (not imputed), and those who self-selected (in effect) to not have blood glucose outcome data in the analysis period (hence imputed). A number of variables display differences between the groups that suggest some such selection bias may indeed have affected the data.

The level of bias in the results of our analyses cannot, of course, be known, but will depend on missing baseline and outcome data, where those with missing data are, on average, different to those with data, and this difference varies between the groups. To get a sense of this, Table 7.14 presents a comparison of the baseline characteristics between those with measured outcome values in analysis 7, in other words, those who measured their blood pressure for more than 208 days (just under 7 months) following enrolment, and those who did not have outcome values available in the analysis period of 23-Nov-2015 to 31-Jul-2016.

Not surprisingly, those in the excluded column were more likely to have missing baseline data as well. Yet, most importantly, of the available baseline characteristics, around half displayed clear differences between the included and excluded groups. This suggests that the measured outcome values were likely to have been different, on average, to those that would have been recorded, but to an extent that cannot be known because they are missing.

Table 7.12 Glucose arm baseline characteristics – full dataset after multiple imputation

Including a comparison with the p-value calculated before multiple imputation

Baseline characteristics	Telemonitoring N = 549	Controls N = 299	P-value	Pre-MI p-value from Table 7.6
Sex, Male	322 (59%)	193 (65%)	0.093	0.530
Age, mean (SD)	67.9 (10.8)	65.9 (11.4)	0.010	0.001
Ethnicity§, Caucasian (Missing: 22%)*	481 (88%)	259 (87%)	0.753	0.267
HbA1c, mean (SD) (DCCT %) (Missing: 73%)	6.8 (1.3)	6.9 (1.3)	0.723	0.944
BMI, mean (SD) (Missing: 31%)	30.4 (5.8)	30.1 (5.8)	0.546	0.838
Diabetes Type 2	494 (90%)	270 (90%)	0.129	0.074
Hypertension	321 (58%)	64 (21%)	< .0001	< .0001
Hyperlipidemia	162 (30%)	60 (20%)	0.003	0.037
Cardiovascular disease	258 (47%)	115 (38%)	0.017	0.003
Arthritis (any type)	249 (45%)	113 (38%)	0.033	0.018
Back pain‡	108 (20%)	64 (21%)	0.549	0.672
Walking pain‡	77 (14%)	44 (15%)	0.784	0.235
Eye problem‡	59 (11%)	33 (11%)	0.897	0.418
Insulin or Analogue	119 (22%)	50 (17%)	0.085	0.814
Pain relief drug	292 (53%)	139 (46%)	0.062	0.015
Number of Type 2 diabetes drugs				
0 drugs prescribed	162 (30%)	106 (35%)	0.209	0.263
1 drugs prescribed	241 (44%)	126 (42%)		
2 – 4 drugs prescribed	146 (27%)	67 (22%)		
Employment status (Missing: 81%)				
Full-time, Part-time or Self-employed	137 (25%)	91 (30%)	0.221	0.734
No employment	122 (22%)	72 (24%)		
Retired	289 (53%)	137 (46%)		
Moderate exercise (Missing: 88%)	135 (24%)	49 (16%)	0.183	0.752
Smoking status (Missing: 45%)				
Never smoker	329 (60%)	177 (59%)	0.972	0.860
Past smoker	185 (34%)	106 (35%)		
Current smoker	35 (6%)	16 (5%)		
Risk level				
Extreme Risk	20 (4%)	12 (4%)	0.008	0.011
High Risk	127 (23%)	57 (19%)		
Medium Risk	41 (7%)	14 (5%)		
Low Risk	191 (35%)	87 (29%)		
Self-Care	170 (31%)	129 (43%)		

* missing from included participants (total=534); ‡ related diagnosis; § some variable categories with low numbers not shown, see Appendix for full table

Table 7.13 Baseline characteristics for participants with and without observed glucose

Baseline characteristics	With observed blood glucose N = 534 (63% of 848)	Imputed blood glucose only N = 314 (37% of 848)	P-value
Sex, Male	340 (64% of 534)†	175 (56% of 314)†	0.024
Age, mean (SD)	67.3 (10.3)	67.1 (12.2)	0.804
Ethnicity, Caucasian (Missing: 24%)*	367 (88%)	196 (86%)	0.615
HbA1c, mean (SD) (DCCT %) (Missing: 73%)*	6.8 (1.2)	7.0 (1.1)	0.086
BMI, mean (SD) (Missing: 33%)*	30.4 (5.5)	30.1 (5.8)	0.549
Diabetes Type 2	485 (91%)	275 (88%)	0.157
Hypertension	214 (40%)	171 (54%)	< .0001
Hyperlipidemia	136 (25%)	86 (27%)	0.571
Cardiovascular disease	252 (47%)	121 (39%)	0.015
Arthritis (any type)	231 (43%)	131 (42%)	0.667
Back pain‡	113 (21%)	59 (19%)	0.427
Walking pain‡	84 (16%)	37 (12%)	0.127
Eye problem‡	61 (11%)	31 (10%)	0.568
Insulin or Analogue	86 (16%)	83 (26%)	0.0004
Pain relief drug	277 (52%)	154 (49%)	0.435
Number of Type 2 diabetes drugs			
0 drugs prescribed	163 (31%)	105 (33%)	0.023
1 drugs prescribed	240 (45%)	127 (40%)	
2 – 4 drugs prescribed	131 (25%)	82 (26%)	
Employment status (Missing: 81%)*			
Full-time, Part-time or Self-employed	20 (20%)	15 (26%)	0.783
No employment	28 (28%)	14 (24%)	
Retired	53 (52%)	29 (50%)	
Moderate exercise (Missing: 90%)*	13 (20%)	8 (35%)	0.161
Smoking status (Missing: 48%)*			
Never smoker	177 (60%)	84 (60%)	0.964
Past smoker	106 (36%)	50 (35%)	
Current smoker	13 (4%)	7 (5%)	
Risk level			
Extreme Risk	22 (4%)	10 (3%)	0.180
High Risk	112 (21%)	72 (23%)	
Medium Risk	27 (5%)	28 (9%)	
Low Risk	177 (33%)	101 (32%)	
Self-Care	196 (37%)	103 (33%)	

* missing from included participants (total=534); † % of non-missing; ‡ related diagnosis

Table 7.14 BP baseline characteristics of participants included and excluded from Anal. 7

Baseline characteristics	Included (measured BP for more than 208 days after enrolment) N = 1,390	Excluded (dropped out before 208 days had passed after enrolment) N = 1,259	P-value
Sex			
Male	796 (57%)	592 (46%)	<.0001
Female	594 (43%)	706 (54%)	
Age (years)			
Mean (SD)	70.1 (9.1)	69.6 (10.4)	0.23
Ethnicity			
Missing (%)	*291 (21%)*	*372 (29%)*	
Caucasian	1,001 (72%)	844 (65%)	0.99
Asian	33 (2%)	28 (2%)	
Other	65 (5%)	54 (4%)	
BMI (last weight from Jul13-Jun14)			
Missing (%)	*496 (36%)*	*536 (41%)*	
Mean (SD)	29.0 (5.3)	29.7 (6.4)	0.01
Diabetes type			
Type 1	4 (0.3%)	5 (0.4%)	0.002
Type 2	116 (8%)	168 (13%)	
Other/unspecified	12 (0.9%)	10 (0.8%)	
No diabetes	1,258 (91%)	1,115 (86%)	
Systolic BP (last from Jul13-Jun14)			
Missing (%)	*581 (42%)*	*680 (52%)*	
Mean (SD)	132.4 (13.3)	132.5 (13.8)	0.81
Diastolic BP (last from Jul13-Jun14)			
Missing (%)	*606 (44%)*	*699 (54%)*	
Mean (SD)	75.6 (8.9)	75.4 (9.4)	0.81
Cholesterol (last from Jul13-Jun14)			
Missing (%)	*1,258 (91%)*	*1,215 (94%)*	
Mean (SD)	4.4 (1.3)	4.7 (1.7)	0.16
Hyperlipidemia	482 (35%)	395 (30%)	0.02
Cardiovascular disease	638 (46%)	521 (40%)	0.003
Arthritis (any type)	688 (50%)	586 (45%)	0.02
Back pain (related diagnosis)	328 (24%)	271 (21%)	0.09
Walking pain (related diagnosis)	179 (13%)	134 (10%)	0.04
Eye problem (related diagnosis)	144 (10%)	122 (9%)	0.40
Insulin or Analogue	198 (14%)	195 (15%)	0.57
Pain relief drug	765 (55%)	616 (47%)	<.0001

Table 7.14 cont. BP: Baseline characteristics of participants included/excluded from Anal. 7

Baseline characteristics	Included (measured BP for more than 208 days after enrolment) N = 1,390	Excluded (dropped out before 208 days had passed after enrolment) N = 1,259	P
Employment status			
Missing (%)	*623 (45%)*	*690 (53%)*	
Full-time	69 (5%)	57 (4%)	0.80
Part-time	57 (4%)	40 (3%)	
Self-employed	41 (3%)	28 (2%)	
No employment	397 (29%)	307 (24%)	
Retired	203 (15%)	176 (14%)	
Moderate exercise			
Missing (%)	*691 (50%)*	*757 (58%)*	
Yes (before Jul 2014)	434 (31%)	299 (23%)	0.02
Smoking status			
Missing (%)	*771 (55%)*	*810 (62%)*	
Never smoker	381 (27%)	298 (23%)	0.47
Past smoker	230 (17%)	179 (14%)	
Current smoker	8 (0.6%)	11 (0.9%)	
Risk level (last from Jul13-Jun14)			
Missing (%)	*86 (6%)*	*94 (7%)*	
Extreme Risk	46 (3%)	57 (4%)	0.16
High Risk	227 (16%)	235 (18%)	
Medium Risk	113 (8%)	96 (7%)	
Low Risk	524 (38%)	439 (34%)	
Self Care	394 (28%)	377 (29%)	

7.3.9 Conclusions about the trial

This chapter provides an introduction to the HCF Telemonitoring randomised controlled trial, highlighting features of its design and other issues that led to significant problems with missing data. The overall aim of the Glucose arm was to determine if the program was effective in reducing the mean blood sugar of participants. HbA1c is the biomarker that best measures this because it is strongly correlated with mean blood levels over the previous 3 months,[843] but although it was nominated as the primary outcome in the trial protocol, at most, only 25% of participants had a result available for analysis, depending on the date range used. This highlights one of the risks of being too 'pragmatic' in trial design, where the goal of providing a clinical service and the goal of collecting data to help answer a causal research question may not be compatible.

The amount of missing data in this trial and the likelihood that, on average, those with data missing are different to those with data recorded, in ways that may relate to their measured or unmeasured outcome, suggests that the assumption of exchangeability will be hard to defend. Nevertheless, our ultimate aim was to extract as much information as we could manage about the causal effect of each intervention, while fully recognising and communicating the uncertainty that surrounds the results.

A significant waste of resources occurs when missing data, measurement error, or any other source of bias, leaves the results of a trial ignored following publication or, perhaps just as bad, leaves the results ignored even by the investigators. When it comes to missing data, many have emphasised the crucial role that study design plays and the ways that a good design can reduce the chance of data not being collected.

In the case of the blood pressure arm, an improved trial design might be one that required participants to provide baseline and final blood pressure measurements, for example by using 24-hour ambulatory blood pressure monitors, or requiring a clinic visit where multiple measurements are averaged. This action, however, might also modify behaviour enough to change their results. Hence, it would only be justified if the intervention, in practice, also included these as components of the intervention, perhaps as part of an initial agreement with participants whether in a formal trial or not. Otherwise, the results might not be transportable from the trial to normal clinical practice.[828] This would probably limit the number of members agreeing when offered the intervention, but the limited participation in this study suggests a more targeted approach, with additional steps to increase data for analysis, might be a worthwhile next step.

On the other hand, it is important to note that these participants were recruited from a chronic disease management program that operates without face-to-face contact, whereas all of the 24 telemonitoring trials referred to at the start of this chapter recruited patients from primary care or specialist medical clinics. Therefore, lower participation may have been unavoidable in this case, when compared to most other telemonitoring trials, and might remain so unless the chronic disease management program transformed into something quite different.

When there is missing data, use of the intervention to generate control group outcomes, or concerns over data accuracy, any interpretation from an analysis needs to be viewed with caution and possibly allow for considerable uncertainty. In the case of this trial, however, of clear value is that it revealed the limits of analysing data from a clinical telemonitoring service containing no face-to-face contact.

With no standardised protocol for collecting data, and an existing data collection routine that did not encourage sufficient data accumulation for accurate inferences to be made, any future pragmatic trials should be aware of the potential for this and try to ensure the problem will not be encountered in the analysis.

If a similar trial were to be conducted in the future, the most important lesson learned from this would be to ensure that (somehow) HbA1c and BP values were recorded for most participants, both at the beginning and at the end, each within a short time frame that was the same for both groups.

Chapter 8
Case study: Avoiding bias and communicating the uncertainty that remains

List of acronyms and synonyms

RCT	Randomised controlled trial
Telemonitoring group	Intervention group
TM	Telemonitoring
BP	Blood pressure
CI	Confidence interval
P	P-value from a statistical test
N	Number of participants
DAG	Directed acyclic graph
NHST	Null-hypothesis significance testing

8.1 Introduction

8.1.1 Overview

Following the preparation detailed in Chapter 7, in this chapter we present and discuss the results of analyses 1 to 7 that used multiple complete datasets generated with the use of multiple imputation.

Additionally, we:

- Explain how causal diagrams helped to determine the variables included in models
- Describe a variety of sensitivity analyses
- Explain some further steps taken to better understand and communicate the uncertainty that remained following the analysis
- Present an additional exploratory analysis that investigates whether the frequency of measurement made a difference to the outcome, a question that involves time-dependent confounding and for which we used the g-formula

And finally, as stated at the start of Chapter 7, we provide conclusions that we hope are more accurate and relevant than we might otherwise have delivered, along with a more accurate sense of the uncertainty that remained following the analysis.

8.1.2 Avoiding bias and weighing the uncertainty

In cohort studies, some form of regression model is usually involved when attempting to avoid the influence of confounding bias.[253] Simpler methods, such as stratification of effect estimates, are sometimes used but are often not practical as the number of variables in the model increases.[844] Also, if missing data is a problem, then methods such as multiple imputation may be used in the attempt to avoid bias, as done in this chapter.

Regardless of whether the estimates from the analysis are true, the inferences researchers make and communicate can still be biased, potentially leading other people to form biased inferences as well. For example, the estimated average effect size of -4 mmHg for an antihypertensive drug may happen to be the true effect size in a trial, but if the 95% confidence interval was (-10, 2), the researchers may incorrectly infer that there was "no evidence" of the drug having an effect. This example is sometimes of greater concern when the effect concerns a risk to health, such as a possible serious side effect of a drug, and investigators deem the drug safe simply because p > 0.05 or the confidence interval includes the null.[845]

Of course, we never do know what the true value is, and so an appropriate sense of uncertainty needs to be considered and conveyed with any estimate. This thesis has attempted to shed light on the sources of bias that are less well understood by researchers and statisticians, such as the cognitive biases that influence the inferences people make and express. These sources will continue to hamper progress in research unless there is improved understanding, not only of these sources of bias but also of methods that can reduce their influence. Some of these will be discussed in the final chapter, but causal diagrams have already been introduced and can serve multiple purposes. Apart from helping with the identification and selection of confounders to include in a model, the process of creating the diagram can also help improve our understanding of the uncertainty surrounding either an effect or the absence of an effect. This includes potential sources of confounding or selection bias that have not been controlled for in the analysis. A causal diagram can also make it easier to judge the plausibility and potential strength of such confounding when forming conclusions following the analysis. But without a deliberate effort to understand the causal structure that underlies the research question, researchers may remain ignorant of possibly important sources of bias.

8.1.3 Sensitivity analysis

Sensitivity analysis is another method that can reduce the chance of biased inferences. The idea of a *sensitivity analysis* dates back at least to the 1950s where the term can be found in articles from econometrics[846] and marketing.[847] The meaning they ascribed to the term was, in essence, to alter some aspect of the final model to see if possible variations would lead to different results. A few years later, the landmark paper by Cornfield et al. (1959)[190] which assessed the sensitivity of the evidence for a causal association between smoking and lung cancer, became what is widely held[848] to be the first example of a sensitivity analysis in health research.

Nowadays, the term *sensitivity analysis* refers to a variety of processes, however, there are two general meanings used in health research, depending on the context and the experience and statistical philosophy of the analyst. The first broad sense of the term is often used for an analysis that mimics the study's primary analysis but varies one or more of the assumptions that the primary analysis has made.[8] This may be a model-based assumption, such as using a random effects model where a generalised estimating equation model would be equally valid,[849] or it may involve adding or removing variables in the model when there is doubt about a variable's role as a confounder. In a similar way, if there is doubt about the best way to define the intervention (such as when it started being used, possibly influencing eligibility of participants or measurements), the outcome (when hard to define precisely, such as mean blood glucose when HbA1c is not available), or an important confounder, then a sensitivity analysis may use a different definition to see if the results change. In other words, this type of sensitivity analysis tests how sensitive the results are to changes in the assumptions that underlie the original results.

Sensitivity analysis can also refer to a process, often now called *bias analysis*[848] or *quantitative bias analysis*[850] in fields such as epidemiology, where an attempt is made to estimate, or quantify, an unmeasured or uncontrolled bias in terms of direction (whether it increased or decreased the main effect), magnitude, and uncertainty.[851] It was in this sense of the term that Cornfield et al. estimated that an unmeasured and unknown confounding variable would need to increase the risk of lung cancer 10-fold to be able to explain away the apparent association of smoking and lung cancer.[188] Such bias analysis can also be used for measurement error and selection bias,[848] and methods now exist to estimate the combined effect of multiple unmeasured confounding variables.[852]

In the context of missing data, it is recommended that a sensitivity analysis should make assumptions about the missing data mechanism that are different from the primary analysis. Hence, if the primary analysis uses a method that is valid only if the missingness mechanism is MCAR, such as complete-case analysis, then the sensitivity analysis should assume that missingness is MAR or MNAR.[778]

In the context of a randomised trial, Morris et al. (2014)[568] suggest that for an analysis to qualify as a sensitivity analysis, assuming that it targets the conclusions drawn from the primary analysis, then (a) it should address the same primary research question, (b) it should be possible for the analysis to reach a different conclusion, and if that happens, then (c) it should not be clear which conclusion should be believed. That is, if a different conclusion is reached, it should increase the uncertainty attached to the original conclusion rather than be easy to dismiss.

The criteria by Morris et al. are general enough to also apply to observational studies, and the publication of these criteria suggests that many analyses that are called a sensitivity analysis' do not, in fact, meet that aim. One example is the observation we made in the methodological review of Chapter 5 that quite a few 'sensitivity analyses' appeared to be no more than subgroup analyses. As such, they could not have challenged the original primary analysis conclusion. But because a sensitivity analysis can increase confidence in the original result if the new result agrees, a false 'sensitivity analysis' may in turn lead to false confidence in the study's finding because it is thought that a sensitivity analysis was done.[313] They can also be consciously or subconsciously manipulated until the result agrees with the original, and there is less pressure to publish additional sensitivity analyses, unlike the primary analysis. Thus, it is not surprising that not all prominent statisticians recommend sensitivity analyses. Frank Harrell, for example, recently wrote in an online forum:[††]

> I've always had trouble with sensitivity analysis. When the different approaches disagree it gives those who favor a certain answer an excuse to use the analysis that most closely provides that. Contrast that with a principled selection of 'the' analysis, which is the way I like to operate in most cases.

8.1.4 Alternative explanations

However, a core reason to conduct a sensitivity analysis is to assess the evidence for one or (preferably) more alternative explanations,[47] because it seems that unless there is a specific stimulus to consider alternative explanations, such as having to devise a sensitivity analysis,

[††] discourse.datamethods.org/t/many-analysts-one-data-set-many-conclusions/1051/12

or at least assessing how likely it is that an alternative exists, as suggested by a statistic like VanderWeele and Ding's E-value,[580] then people will tend to be influenced by the take-the-first heuristic[480] and fail to consider alternative explanations of a study's results.

Another way that a person may be stimulated to consider explanations for an association different from 'the intervention caused the outcome' is for that person to deliberately create a list of plausible alternative explanations.[109,308,528,853,854] While the motivation to routinely do this may need to come from, for example, journals, regulatory agencies or research funders, the establishment of pre-specification of statistical analysis plans as an expected standard for many randomised controlled trials[855] suggests that other standards of practice can happen with time if there is widespread agreement. And this can perhaps be done most easily with the aid of causal diagrams.[856]

8.1.5 Time-dependent confounding and the g-methods

Prior to this century, it was generally accepted that once an intervention had commenced, only data for the outcome should contribute to the analysis, with included covariates restricted to baseline values only. This is because when the intervention is being used, it might affect not only the outcome but also some of the covariates, and this can only occur in the group that receives the treatment, thereby introducing bias if those modified covariate values are conditioned on in the analysis.[857,858] But not including the covariate might also introduce bias, as Kalbfleisch and Prentice suggested in 1980 (discussed in Keiding and Clayton (2014)[859]). For example, if the covariate represented the severity of a particular symptom in a trial where the final outcome is all-cause mortality, and the symptom's severity often leads to an additional treatment that increases survival (and possibly censoring), then leaving the covariate out of the model ignores this source of confounding. On the other hand, the covariate might lie on the causal pathway between the treatment and the outcome; hence, conditioning on that covariate could remove some of the causal effect of the treatment on the outcome. Standard regression methods such as linear, logistic and Cox regression are unable to handle time-dependent confounding.

A solution to this problem was devised by James Robins in 1986[860] which he called the *g-computation algorithm*, with the 'g' referring to the 'generalised' nature of the algorithm where, subject to the standard assumptions of no unmeasured or uncontrolled confounding, no measurement error and no model misspecification, it has the ability to provide unbiased estimates of the causal effect of a 'hypothetical' intervention, providing that the intervention, outcome and all covariates are measured at each individual time point.[861] In the same original paper as the algorithm was the more compact *g-computation algorithm formula*, which by 1995 had been shortened to the *g-formula*.[862] In addition, it is also sometimes referred to as the *parametric g-formula* because although in simple cases the g-formula can be used without the aid of statistical models, in most realistic analyses, the g-formula algorithm will require parametric models and a Monte Carlo simulation.[861]

Within a few years, Robins developed an alternative method for time-dependent analysis with the semiparametric *g-estimation for structural nested models (SNMs)* (1989),[863] where the models include structural nested failure time models (1989, 1992)[69] and structural nested mean models (1989, 1994).[70] Finally, in 1998 Robins developed a third method for time-varying exposures that he called *marginal structural models* (MSMs) and which are commonly estimated using *inverse probability weighting* (IPW), which is usually *inverse probability of treatment weighting* (IPTW). Note, however, that MSMs can be estimated using g-computation or targeted maximum likelihood estimation (TMLE).[864] Taken together, the g-formula, inverse probability weighted marginal structural models, and g-estimation of structural nested models make up the group that Robins and Hernán calls the *g-methods*.[865,866]

When the intervention and covariates are all discrete, with only a few time points and the study is large, then estimates can be calculated non-parametrically because the models are fully saturated and in this case, all three g-methods will give the same answer.[867] In most cases, however, modelling assumptions are needed and these differ between the three methods.

The g-formula

Briefly, and based on examples by Robins and Hernán (2009)[865] and Daniel et al. (2013),[278] the simplest version of the g-formula is for the expectation of the mean outcome Y, given the intervention received $A = a$ (e.g., treatment or control), and a set L of baseline covariates, and is defined to be the weighted sum of the means of Y within each unique set l of covariate values or strata and for each intervention a. The weights equal the number of participants in each stratum and the sum is over all the different levels l of L in the study sample. In mathematical notation, the g-formula for $E(Y_a)$ can be expressed as:

$$E(Y^a) = \sum_l E(Y|A = a, L = l)\Pr(L = l)$$

If L contains continuous variables, then the sum becomes an integral:

$$E(Y^a) = \int E(Y|A = a, L = l)dF_L(l)$$

The estimates $E(Y_a)$ for each hypothetical intervention a can then be compared. And because the average is taken over the whole sample, we consider it to be *marginal* over all of the covariates, meaning that the estimated mean is in relation to an average of the measured covariates, as opposed to the results from a regression analysis which is conditional on specific values of the measured covariates. Hence, the g-formula is considered to be a generalisation of the technique of standardisation to enable the handling of time-varying treatments and confounders.

Generalising to a time-varying setting, for the period of a study up to and including time t (e.g., a follow-up visit or regular home measurement), we now set $\bar{A}_t = (A_0, \cdots, A_t)$ to denote the vector of treatment history up until that time and $\bar{L}_t = (L_0, \cdots, L_t)$ to denote the covariate history. The above formula for fixed settings now becomes:[278]

$$E(Y^{\bar{a}}) = \sum_{\bar{l} \in \bar{\mathcal{L}}} E\left(Y \middle| \bar{A} = \bar{a}, \bar{L} = \bar{l}\right) \prod_{t=0}^{T} f\left\{L_{t-1} = l_t \middle| \bar{A}_{t-1} = \bar{a}_{t-1}, \bar{L}_{t-1} = \bar{l}_{t-1}\right\}$$

The g-formula for time-varying exposures can provide estimates $E(Y^{\bar{a}})$ for each treatment by simulating the joint distribution of the intervention history (if received on multiple occasions), covariate history, and the outcome, such that the means of Y are estimated for each unique combination of the intervention and covariate history. In other words, $E(Y|A = a, L = l)$ is calculated for each combination of the treatment and covariates at each time point.

One potential downside of using the g-formula is what has been called the "g-null paradox", where a null hypothesis of no effect would tend to be rejected with a large study, even when it is true.[278] Hence, the g-formula is not recommended when interest lies in testing such a causal null hypothesis. Neither marginal structural models nor g-estimation exhibit this problem, however.[865]

Marginal structural models and inverse probability weighting

The most popular of the g-methods by a considerable margin, marginal structural models and their estimation by inverse probability of treatment weighting (IPTW)[864,868] are perhaps the easiest to understand and implement with available software, helping to explain their appeal.[865,869]

The weights for each individual are calculated as the inverse of the probability that they received treatment, conditional on the measured potential confounders. In other words, the inverse of the propensity score.[59] The weighting of a participant by the conditional probability that they are in the intervention group can remove confounding by creating a pseudo-population where participants with each unique combination of covariate values will have an equal number who received the treatment and an equal number who received the control. Further details are beyond the scope of this chapter, however, with the g-formula the only method that we make use of. Likewise, we do not feel that any understanding of the more complicated g-estimation can be achieved without an in-depth study, and hence we will not try to provide a simple description nor explore it any further here.

On a final note, while one or more of the g-methods are an obvious choice to deal with time-dependent confounding, some recent developments[288] suggest they are no longer the only option. Nevertheless, such methods will need to be understandable without exhaustive effort, if any but a small number of specialists are to use them.

8.2 Methods

8.2.1 Group comparisons

To assist the reader, Tables 3 and 4 from Chapter 7 have been combined (Table 8.1) to show the group comparison analyses that were devised following an assessment of missing data.

Table 8.1 Intervention, control and definitions of the sample used for group comparisons

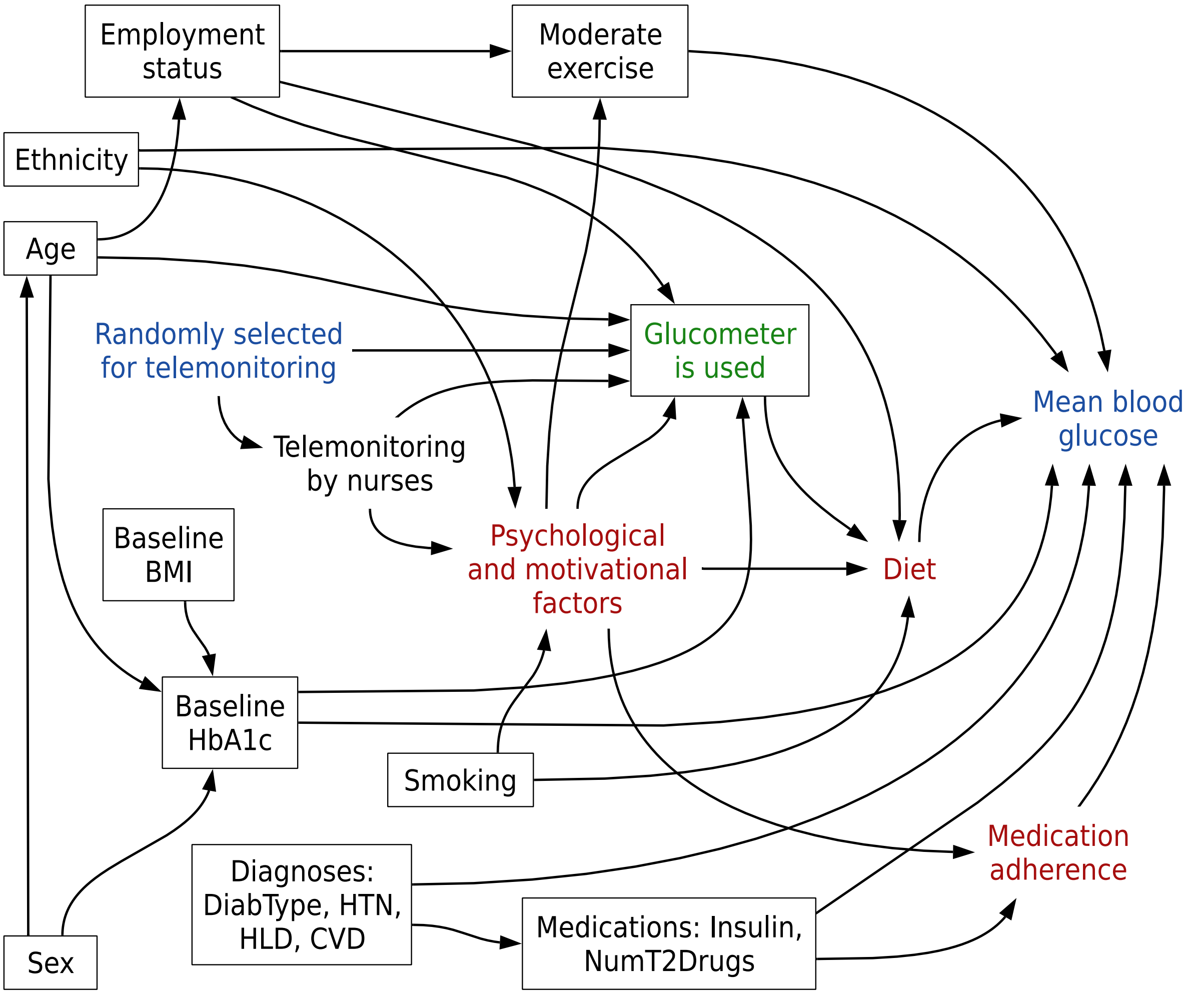
Glucose arm	Analysis 1	Analysis 2	Analysis 3	Analysis 4
Participants	All*	All*	All*	All*
Measurements used to create outcome[†]	All first-in-day glucose 1-Jul-15 to 30-Nov-15	First-in-day glucose 1-Jul-15 to 30-Nov-15 satisfying below	Any HbA1c 1-Apr-15 to 31-Dec-15	BMI from height and • Last weight 1-7-13 to 30-6-14 (baseline) • First weight 1-4-15 to 31-12-15
Telemonitoring outcome	Mean blood glucose	Mean of middle 5 measurements[♯]	HbA1c	Change in BMI from baseline
Control outcome	Mean blood glucose	Mean of first 5 measurements	HbA1c	Change in BMI from baseline

BP arm	Analysis 5	Analysis 6	Analysis 7
Participants	All	All	All who measured on at least 5 days, at least 208 days after enrolment
Blood pressure measurements used[†]	Any from 23-Nov-15 to 31-Jul-16	Any from 23-Nov-15 to 31-Jul-16 but daily averages used	
Telemonitoring outcome	Middle 5 BP measurements[‡]	BP measurements from middle 5 days[§]	
Control outcome	First 5 BP measurements[¶]	BP measurements from first 5 days**	

[†] before multiple imputation; * following multiple imputation; [♯] mean of 5 measurements closest to 31-Mar-16 (middle of 23-Nov-15 to 31-Jul-16) [‡] mean of 5 measurements closest to 31-Mar-16 (middle of 23-Nov-15 to 31-Jul-16); [§] mean of 5 days with measurements closest to 31-Mar-16; [¶] mean of first 5 measurements 23-Nov-15 to 31-Jul-16; ** mean of first 5 days with measurements 23-Nov-15 to 31-Jul-16

All outcomes were continuous, so a linear regression model was used to compare the Telemonitoring and Control groups. And although most participants recorded more than one blood glucose or blood pressure measurement, they were not recorded at the same time or with the same frequency, so a single summary measure was thought to be the best way to compare measurements instead of attempting a longitudinal model.

Figure 8.1 Causal diagram showing the causal structure for analyses 1 and 2



Based on the causal diagram for mean blood glucose as the outcome (Figure 8.1), a multivariable linear model was fitted to the imputed datasets using the SAS multiple imputation procedures PROC MI and PROC MIANALYZE, and PROC GLM, with the following baseline covariates:

```
Telemonitoring, Age, Sex, Ethnicity, baseline HbA1c, baseline BMI, Diabetes type,
Hypertension, Hyperlipidaemia, Cardiovascular Disease, Arthritis (any type), Back
Pain, Walking Pain, Eye Problem, Insulin or Analogue, Number of diabetes drugs,
Pain relief drug, Employment status, Self-employed, Moderate exercise, Smoking
history, Risk Level
```

The same linear model was used for each of the four outcomes.

In terms of forming inferences from the results, to reduce the chance of accidental bias through a subconscious process such as 'significance questing',[324] we followed the guidance of the American Statistical Association[383] and many prominent statisticians[294] by not using significance testing. That is, while p-values were calculated, we did not use a threshold such as 0.05 to declare support for an association, which we believe can easily mask the true level of uncertainty. Instead, p-values were used simply as a guide, along with confidence intervals and knowledge of potential confounding and selection bias, to help form judgements about the strength of any associations indicated by the data and model at hand.

Finally, only the estimates of effect for the intervention group were reported from the results of the multivariable models, rather than the estimates for all of the variables in the model, as is not uncommon. This was to avoid what has been termed the "table 2 fallacy",[870] where the estimates for the confounders in the same model tend to also be interpreted as effect estimates when presented in a table (often "Table 2" in health research articles), though the "confounders of the confounders"[678] are often going to be missing from the model, and some covariates are also likely to be mediators for the effect of other covariates. This last possibility can be seen in Figure 8.1 with, for example, the effect of Baseline BMI at least partly mediated, and thus diluted, by Baseline HbA1c.

8.2.2 Sensitivity analyses

Alternative definitions of the outcome

For an individual participant who uses the telemonitoring intervention, it is hard to define what they or their doctor might consider to be a successful outcome, other than the long-term goals of avoiding the health problems associated with poor glucose or blood pressure control. Hence, it was worth varying the definitions we used to see if the results were consistent. Especially if this may have avoided a possible source of bias and thus tested for, or presented evidence for, an alternative explanation of the results.

Measurement trajectories

While comparison with a control group is very important for good evidence, analysing individual and group trajectories of measurements can provide further information to help understand the effects of an intervention. For both the blood glucose and blood pressure arms, we examined whether there was evidence that the telemonitoring program led to lower values, on average, being recorded over time.

Examination of assumptions

Numerous assumptions were made implicitly during the analysis and we endeavoured to examine any that seemed questionable. These assumptions included that:

1. the delivery of the intervention did not change over time in a way that influenced the outcome (the *consistency* assumption)
2. the intervention had no effect on the initial 5 measurements
3. if the outcome had been assessed in a different way, a similar conclusion would have been reached
4. the difference in group outcomes was not sensitive to small changes in the dates on which they were compared
5. no measurement error existed sufficient to have changed conclusions

E-values

We note here that E-values were calculated for one of the BP arm analyses, however, it was decided that they were not readily interpretable within the context of this trial. As an explanation, VanderWeele and Ding[580] define an E-value as “the minimum strength of association, on the risk ratio scale, that an unmeasured confounder would need to have with both the treatment and outcome, conditional on the measured covariates, to fully explain away a specific treatment–outcome association.” However, they also state that for a "difference in continuous outcomes … an approximate E-value may be obtained by applying the approximation RR ≈ exp(0.91 × d) in the E-value formula". In other words, while an E-value could be calculated for the results in this trial, as a risk ratio, we could not clearly relate it to differences found in continuous measurements.

8.2.3 Alternative explanations

A deliberate attempt was made to think of alternative explanations for the results and to list them. In doing so, it was hoped that we would gain a more accurate sense of the uncertainty that remained following the analysis, and that we could communicate this more effectively to HCF and any other interested parties.

8.2.4 The effect of measurement frequency

Two clinical questions of interest were:

1. If a person used the telemonitoring device more often, was their mean blood glucose level more likely to be lower at the end of the trial, compared to no change or higher?
2. Is there likely to be any benefit from encouraging participants to take regular measurements?

To help answer these questions the following research question was created:

> Did the frequency of home blood glucose self-measurement, within this telemonitoring program, have a causal effect on the group's mean blood glucose level over time?

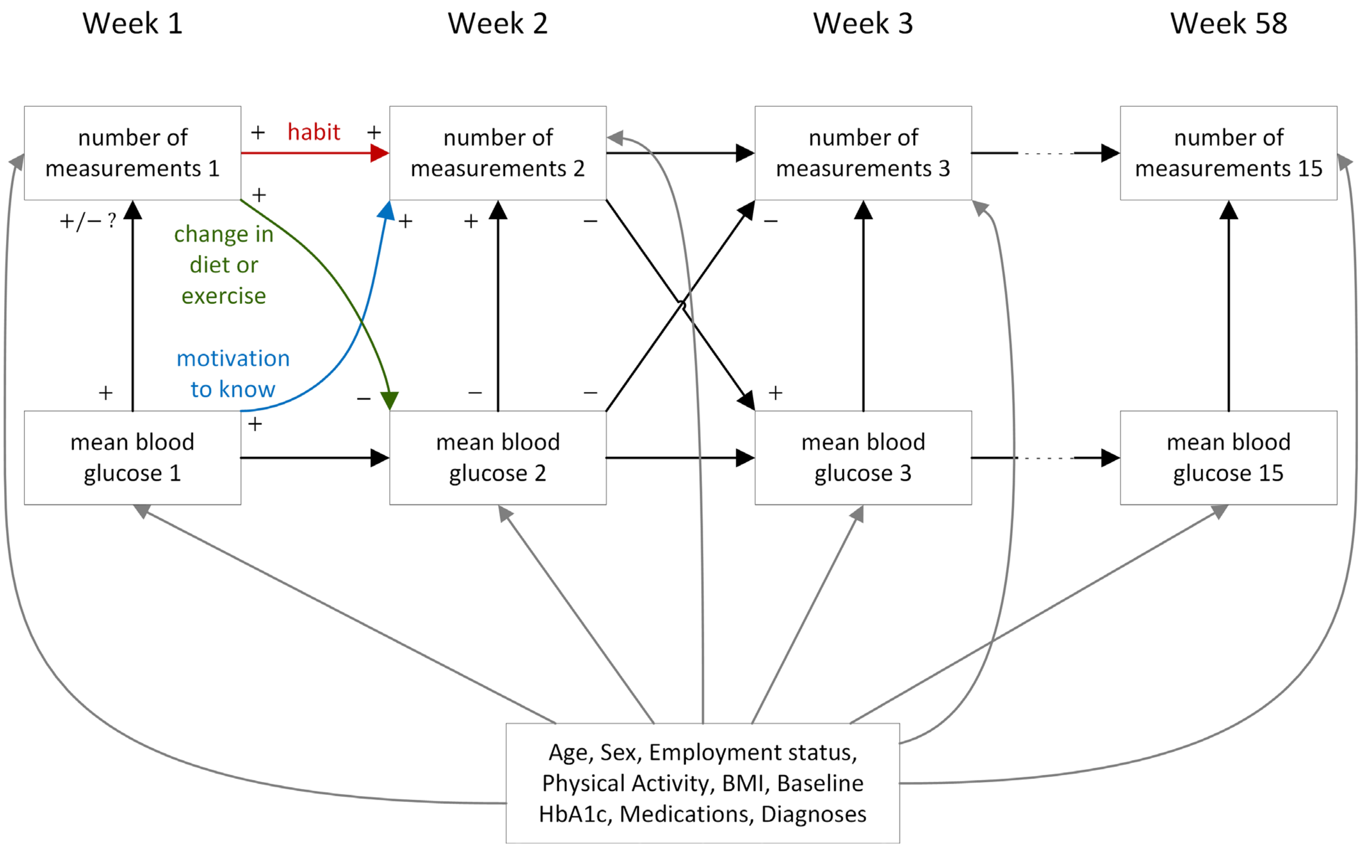
We created a directed acyclic graph (DAG) (Figure 8.2), initially based on a design from Daniel et al.,[871] but we added a few features that we thought might improve the ease and speed of understanding, with its intended use to explain the potential for time-dependent confounding to HCF researchers. These were simply labelling some arrows, using colours, and using + or – to denote whether the causal effect was expected to increase or decrease the variable being acted on.

The DAG shows a number of plausible causal associations. First, there is our initial hypothesis that the number of measurements each week might have an influence on the mean blood glucose of participants in the following week. This relates to the idea that measuring blood glucose will alert the participant to higher than expected values, if present, and hopefully prompt some action such as lowering sugars in the diet or increasing exercise. Second, the number of measurements in each successive week might be causally linked through the development of habits. However, third, it's also possible that worse than expected blood glucose levels in one week, might increase measuring frequency in the following week due to increased motivation to know what the blood glucose levels are. Alternatively, better than expected results might create less of a psychological need to know and so measurement frequency might decrease.

If any of these occurred, then a standard analysis of the data might find that higher measurement frequency was associated with higher blood glucose levels and lower measurement frequency with lower blood glucose levels. Hence, the relationship estimated using standard techniques might end up suggesting the opposite of what really occurred. To deal with this problem, we decided to use the parametric g-formula, partly because the 'treatment' was effectively continuous and there was some mention in the literature on marginal structural models that "one should be careful when using IP weighting for continuous treatments because the effect estimates may be exquisitely sensitive to the choice of the model for the conditional density".[59] As it turned out, however, we eventually did define a binary treatment of sorts.

Figure 8.2 Causal diagram showing time-dependent confounding potential (Glucose arm)

Causal effect of blood glucose measurement frequency on changes in mean blood glucose over time, where + or − denotes the expected increase or decrease of the variable being acted on



For this analysis, the intervention (or treatment) was initially defined vaguely as more frequent measuring of blood glucose at home, while the alternative intervention was defined as non-frequent measuring. In the trial, the frequency was controlled by the participant and each participant’s measurement frequency (by month) was calculated.

The outcome used in this analysis was the mean, each week, of all first-in-day blood glucose measurements between 1 Jul 2014 and 31 Dec 2015, with each participant’s data modified so that the week number for their first weekly or monthly measurement was set as occurring in week 1 (rather than week 5 or week 8, etc).

The eligible participants were those in the Telemonitoring group with at least 1 blood glucose self-measurement in each of the 14 months covered.

Note that the consistency assumption would imply that for measurement frequency, the blood glucose outcome should be the same whether the person was “required” to measure blood glucose daily or happened to do so on their own.

Before using the g-formula, we used a variety of techniques to examine whether time-dependent confounding may have occurred. But as with the g-formula, the process was more one of discovery than pre-planned, and hence the process will be described with the results.

8.3 Results and Discussion

8.3.1 Glucose arm group comparisons

A summary of the main statistical results for the Glucose arm is shown in Table 8.2.

Table 8.2 Glucose arm group comparison results

Analysis Outcome		N	N with outcome missing*	Results	
				Estimate† of difference between groups (95% CI)	P
1. Mean blood glucose	TM	549	278 (51%)	-0.38 mmol/L (-0.78, 0.02)	0.06
	Controls	299	36 (12%)		
2. Middle 5 vs first 5 glucose measurements	TM	549	278 (51%)	-0.59 mmol/L (-1.03, -0.14)	0.01
	Controls	299	36 (12%)		
3. HbA1c	TM	549	436 (79%)	-0.13 % (-0.49, 0.24)	0.50
	Controls	299	203 (68%)		
4. BMI change from baseline	TM	549	378 (69%)	-0.19 (-0.73, 0.35)	0.49
	Controls	299	148 (49%)		

* Before multiple imputation; † Mean difference in outcomes, adjusted for baseline values

Some evidence of an association is apparent between both of the glucose outcomes and the Telemonitoring intervention. But while it should be remembered that both outcomes relate to the same overall dataset, the small subset of the first's outcome data that the second analysis uses, is very different in important ways.

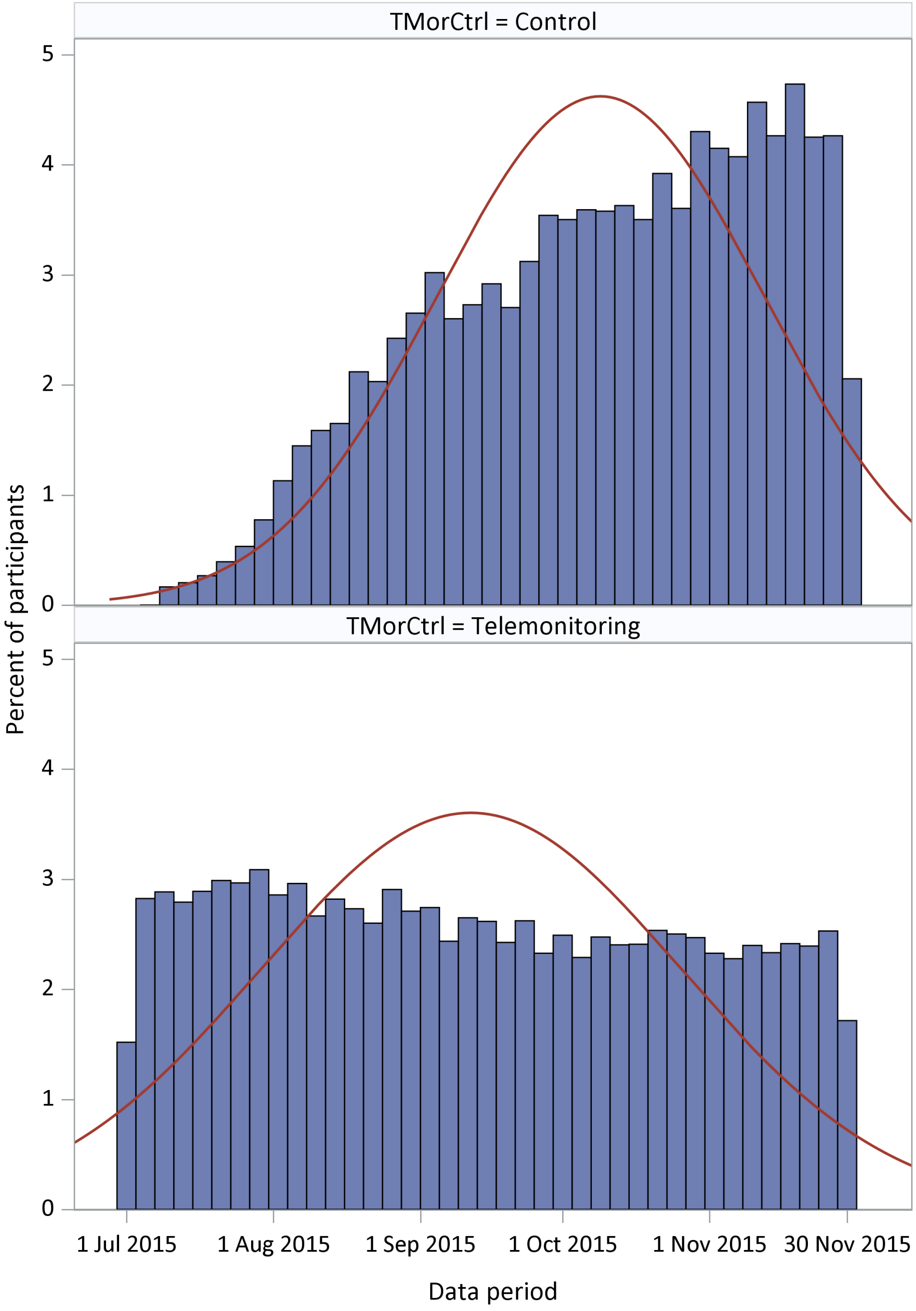
One other concern is that a difference of 0.38 or 0.59 mmol/L may not be clinically meaningful. There is considerable doubt about the accuracy of these estimates, however, due both to the amount of missing data, as well as other possible biases we will explore below. There was no evidence found, however, that would suggest changes in HbA1c or BMI to the program.

8.3.2 Glucose arm group comparison sensitivity analyses

For analysis 1

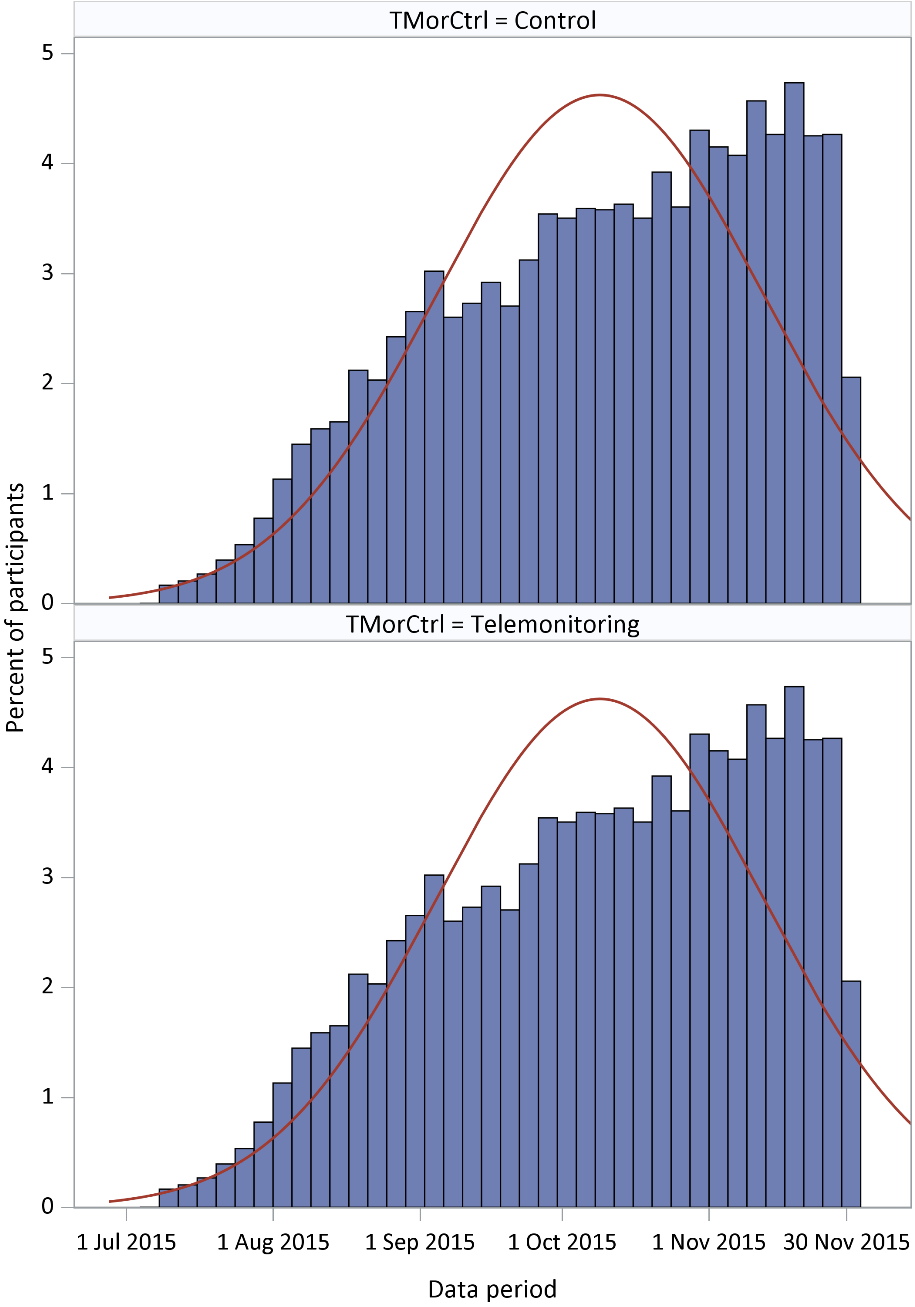
In Figure 8.3, we compare the distribution of the dates on which blood glucose measurements were taken for the two groups.

Figure 8.3 Glucose arm distribution of measurement dates for each group before matching



The difference can be explained by the fact that the Control group (top) started recruiting at the beginning of this period and so as enrolment progressed, we see the total number of participants measuring each week increasing. The Telemonitoring group participants (bottom), however, are those that have continued measuring for many months and, with most who have made it to this stage less likely to stop, the number of participants measuring each week remains relatively constant.

Figure 8.4 Glucose arm distribution of measurement dates for each group after matching



To rule out the possibility that changes in intervention delivery over time might have had an effect on the outcome, we matched participants measurements by date so that the distribution of dates that measurements were taken on became roughly equal between the two groups (Figure 8.4). While there was no specific reason to suspect that the intervention did change over time, it is plausible that as staff gained greater experience with the intervention or there were changes in personnel, then the telemonitoring component, or even the types of participant that agreed to participate, might have changed slightly.

Table 8.3 Glucose arm model results with and without date matched measurements

	Variable	Estimate (95% CI) *	P-value
With matching on date	Telemonitoring	-0.44 mmol/L (-0.85, -0.03)	0.03
No matching	Telemonitoring	-0.38 mmol/L (-0.78, 0.02)	0.06

* the estimate is for the mean difference in mean blood glucose, following multiple imputation and adjusted for baseline values

Table 8.3 compares the results and suggests little difference. Assuming that no $p < 0.05$ cut-off was used to designate the status of evidence, we would not alter any previous conclusions.

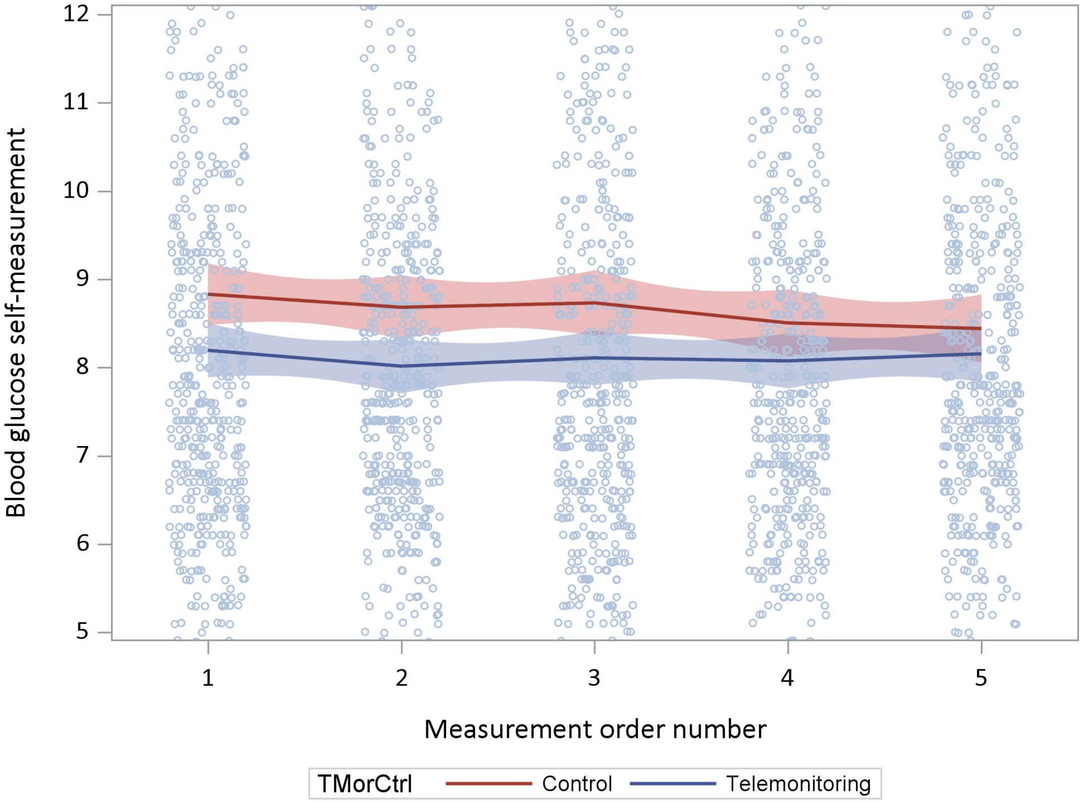
For analysis 2

One potential concern was that the intervention might have had an influence on participant behaviour as soon as they started measuring. If this were the case, the glucose measurements of the Control group might have rapidly dropped within the first 5 measurements that they recorded. Figure 8.5 tested this by plotting lines of best fit through the first 5 measurements from the Control group and also from the Telemonitoring group.

While a slight downward slope is suggested in the Control group curve, it is not present in the Telemonitoring curve, and the drop is too small to represent a concern. A simple linear regression model found that the slope had a p-value of 0.11 (for the Telemonitoring group P = 0.96). Thus, we can be somewhat confident that the first 5 measurements of the Control group are a reasonable approximation of the group’s mean blood sugar levels before the Telemonitoring program had a chance to influence it.

We added an interaction term for glucose measurement order number to the intervention term in a PROC MIXED model instead of using PROC GLM which did not appear to allow it, however, it only made a small difference to the Type 3 p-values.

Figure 8.5 Glucose arm LOESS lines for the first 5 (Controls) or middle 5 (Telemonitoring)



8.3.3 Visually judging line of best fit graphs

The curves in Figure 8.6, Figure 8.7 and Figure 8.8 appear to suggest a reduction in blood glucose, on average, over the trial period in the Telemonitoring group members who used the glucometer: 454/549 participants (83%). There also appears to be an initial sharp drop in blood glucose levels, on average, that lasts about 2 months. Following this, the level appears to decline slowly until a rebound appears after 7 months and peaks at around 1 year. After this, the level starts to drop again, though by this point more than half of the participants who started recording their blood glucose have stopped (see Table 8.5) so the characteristics of these remaining participants are probably a little different to those no longer recording.

Figure 8.6 LOESS line for mean monthly blood glucose of Telemonitoring participants

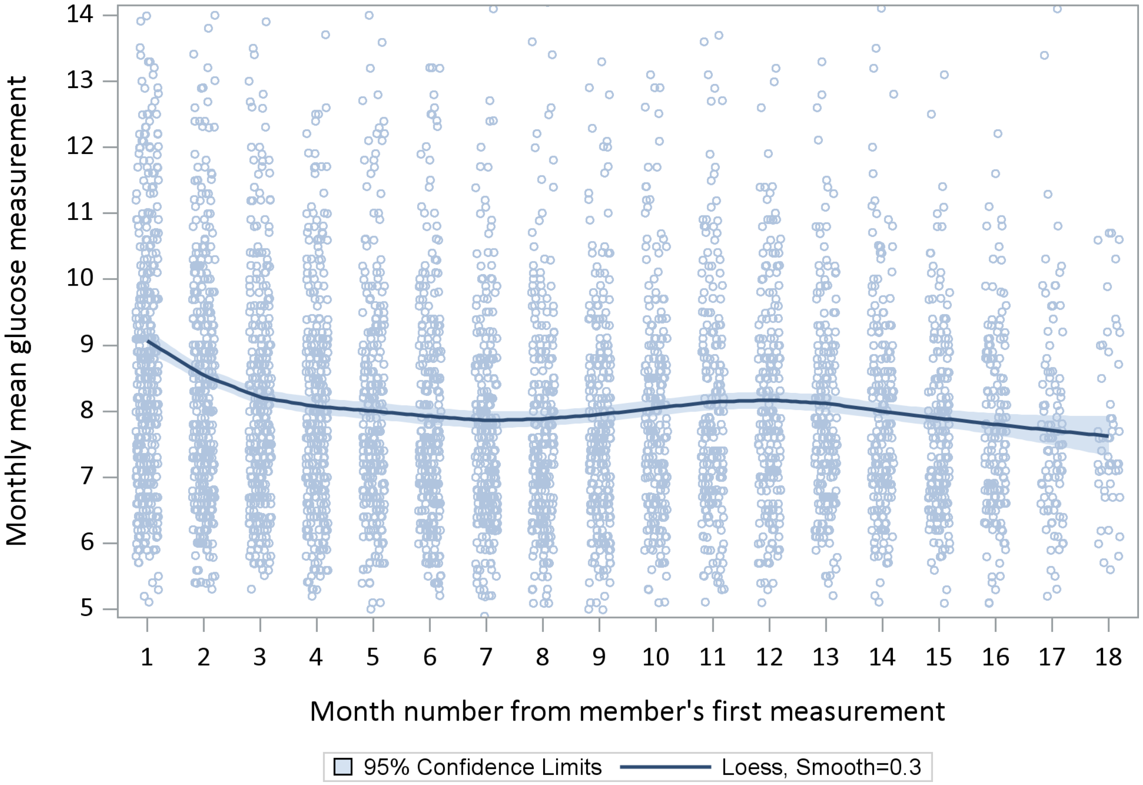


Figure 8.7 LOESS line for mean monthly blood glucose of Control participants

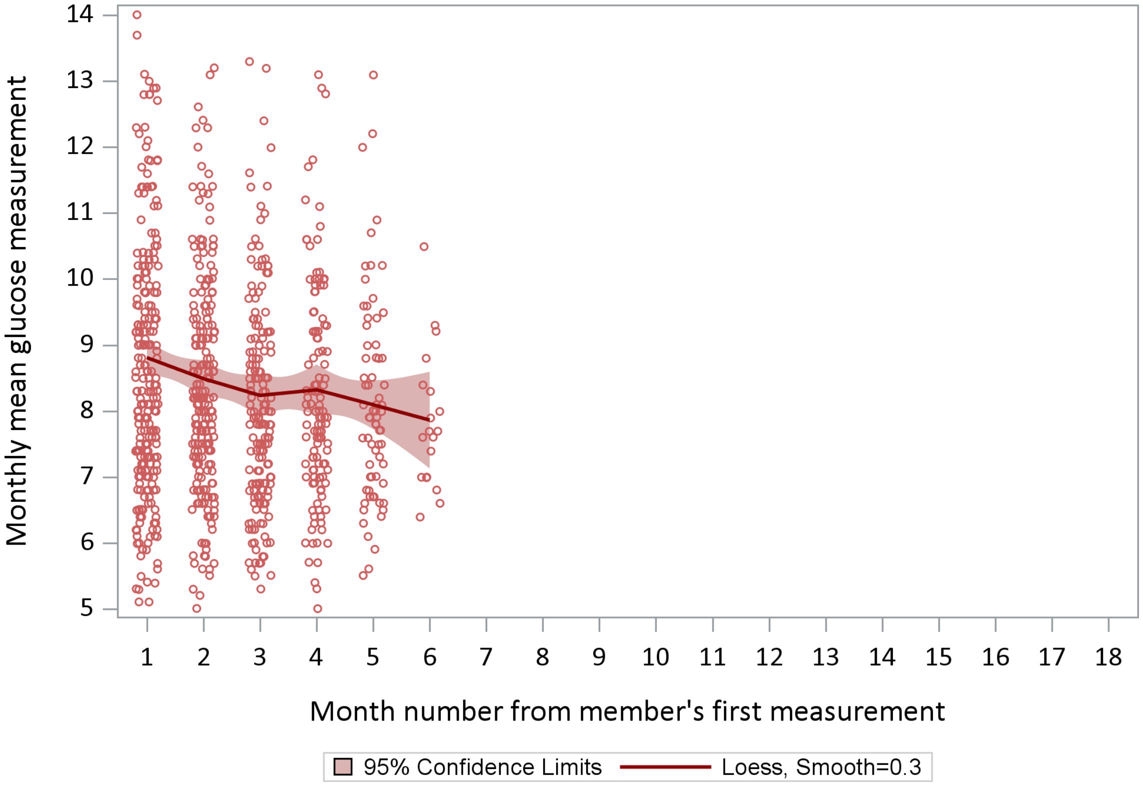
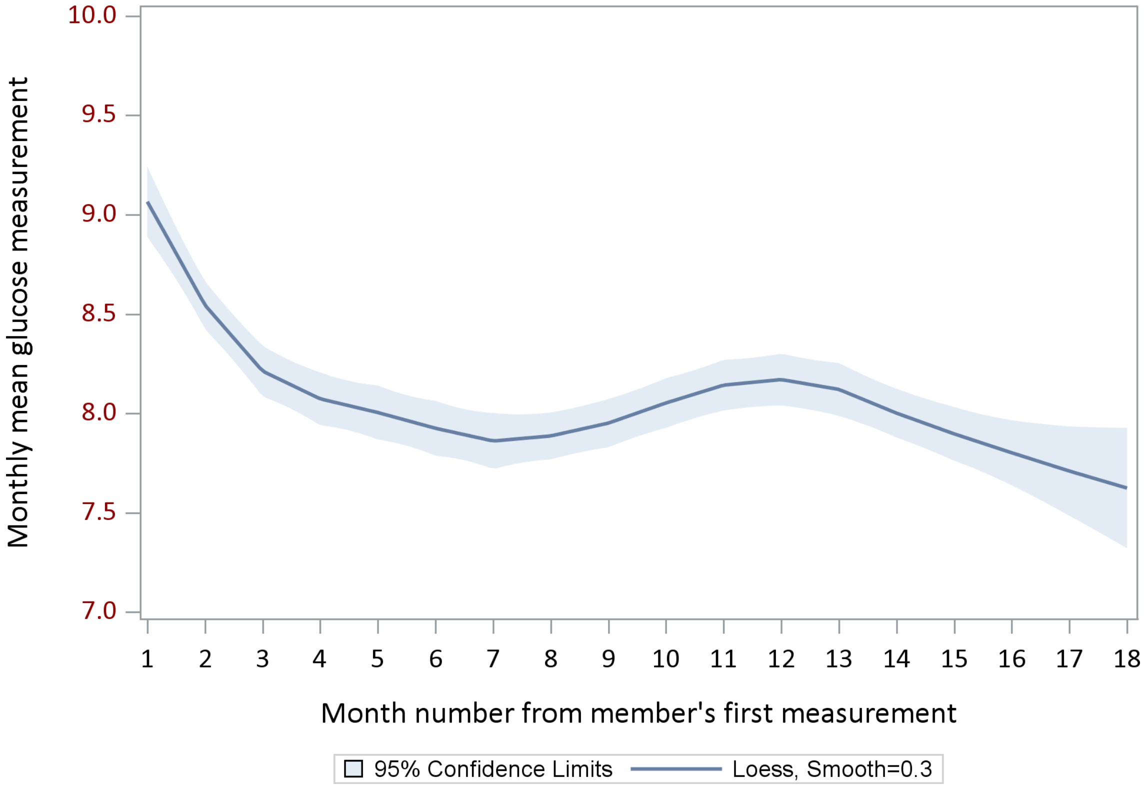


Figure 8.8 LOESS line for mean monthly glucose of Telemonitoring participants zoomed-in



8.3.4 Glucose measurement trajectories

For this analysis, the outcome was the weekly or (separately) monthly mean of all first-in-day blood glucose measurements. The included participants were the Telemonitoring group members with at least 2 glucose measurements (minimum for a trajectory line) between 1 July 2014 and 31 December 2015. This resulted in 324/512 (59%) participants, and the research question was:

> Does the self-measurement telemonitoring program cause at least some participants to make changes to their lifestyle, diet, medication adherence or other factors, that result in a lower mean blood sugar level over time?

The methods we used were: visually judging line of best fit graphs; using time series regression models to estimate the mean slope of blood glucose over time; and comparing the proportion of participants whose mean blood glucose increased or decreased over time.

The main trigger for looking at the trajectories of the Telemonitoring glucose group's values was the lack of baseline values with which to compare changes between the two groups.

The LOESS procedure that produced the graphs above assumes that the blood glucose measurements are independent. This is unlikely to be true for each individual participant, however, because values of blood glucose close in time will often be correlated through day-to-day similarities in lifestyle, diet, medication and other factors that will be less similar with time gaps more distant, such as one year later. This phenomenon is termed autocorrelation and will lead to inflated standard errors if autoregressive terms are not included in regression models.[872] However, autocorrelation does not bias the estimated slopes of the fitted regression lines,[873] so for the purposes of calculating a mean of the slope estimates, autocorrelation can be safely ignored.

For each participant, we calculated an estimate of their blood-glucose-over-time linear slope, and then calculated the overall mean slope for all participants with at least 10 weeks' worth of measurements, as well as 30 or more weeks' worth (Table 8.4). Shown also is the result of a one sample t-test.

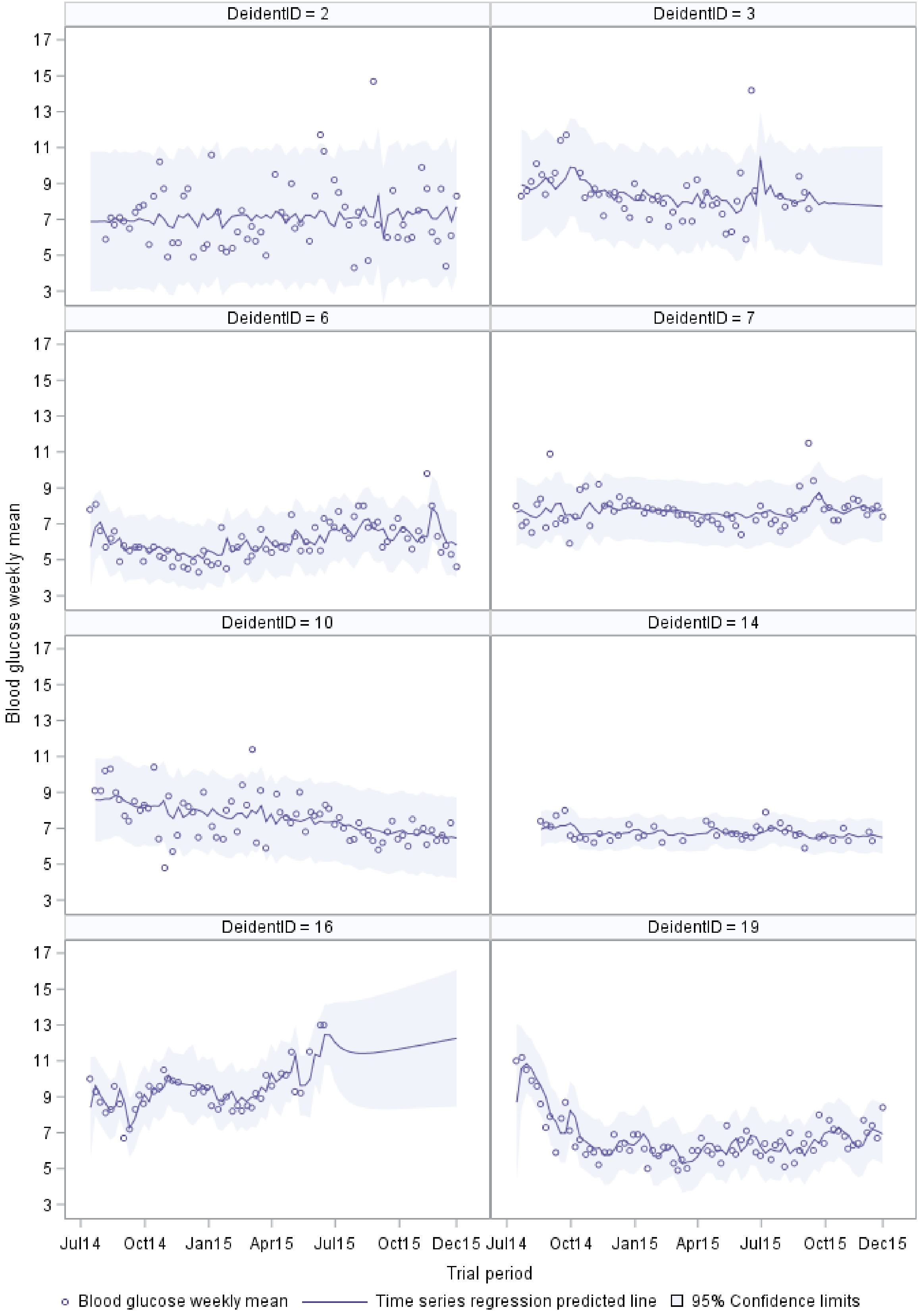
Table 8.4 Average slope of weekly mean blood glucose for Telemonitoring arm

Participants measured for	N	Min	Max	Mean	95% CI for Mean		P-value
At least 10 weeks	314	-0.372	0.192	-0.0067	-0.0121	-0.0013	0.015
At least 30 weeks	243	-0.102	0.155	-0.0017	-0.0053	0.0019	0.344

A one sample t-test was used to determine, for the distribution of blood glucose slopes with a minimum number of weeks with measurements set to 10 or 30, the chance that it is sampled from a population of blood glucose slopes with a mean of 0.

If this mean change in blood glucose were maintained for one year, the drop in a participant's mean blood glucose would equal 52 x -0.0067 = -0.35 mmol/L. But it is worth noting that the individual trajectories were quite variable (see Figure 8.9 for some examples), with some seeming to decrease and some increase.

Figure 8.9 Some examples of individual participant's weekly mean blood glucose series with time series predicted regression lines



An illustration of the variability of individual trajectories can be seen in Figure 8.10 and Figure 8.11, where the graphs of participants' weekly mean glucose trajectory slopes appears to be an approximate normal distribution about a mean that is close to zero.

Figure 8.10 Distribution of weekly mean blood glucose trajectory line slopes 10+ weeks

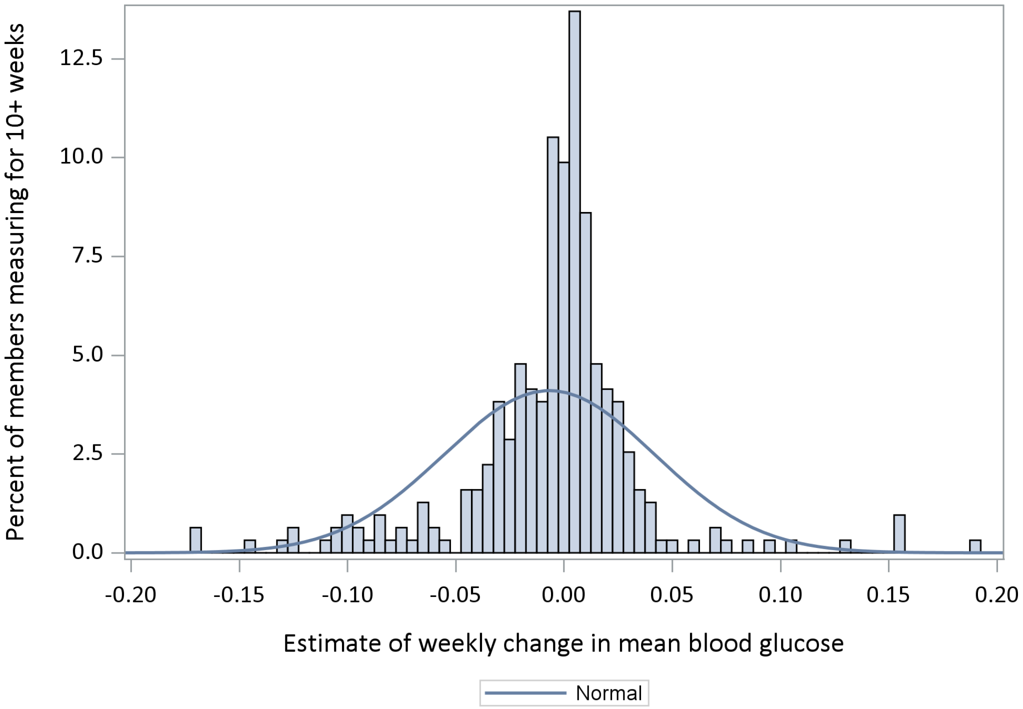
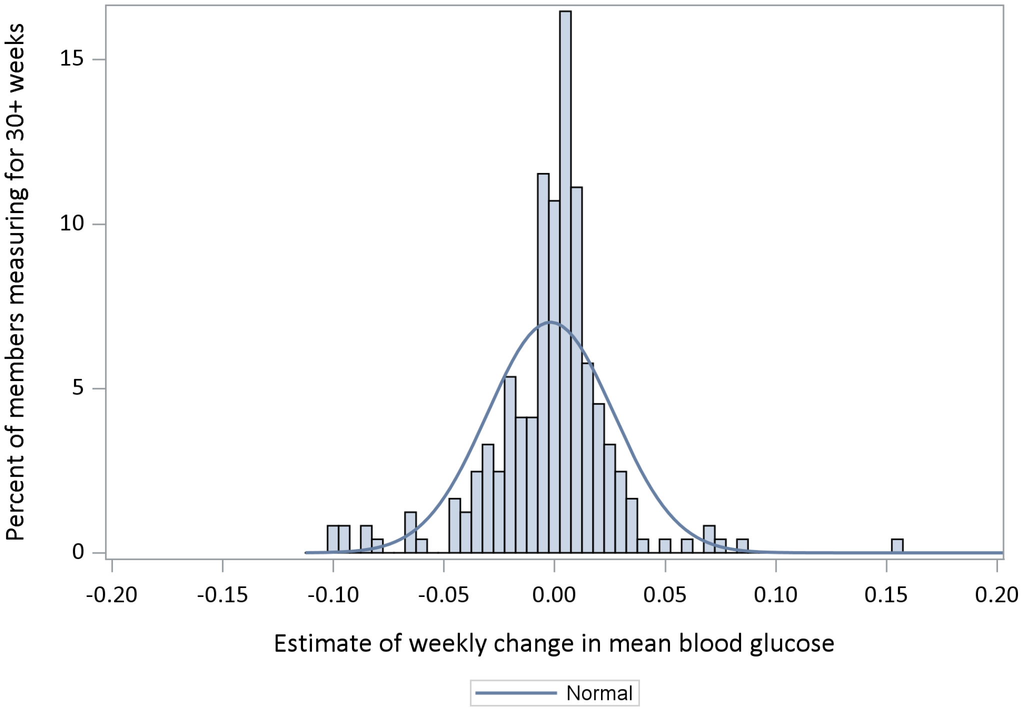


Figure 8.11 Distribution of weekly mean blood glucose trajectory line slopes 30+ weeks



Supporting this perspective, Table 8.5 shows that a roughly equal number of participants had a positive or negative trajectory, though the subgroup that measured the longest appeared to favour negative trajectories where glucose levels rose slightly. This suggests that the program did not result in an overall reduction in mean blood glucose levels, at least not to any meaningful degree. There is still great uncertainty, however, because of the quantity of missing data, and in this case, the number that either did not use the glucometer or who did not use it for more than a few months.

Table 8.5 Participants with a positive or negative glucose slope estimate

	Number of participants				
Minimum weeks with measurements	Mean glucose went down		Mean glucose went up		Total participants
10	158	50%	156	50%	314
15	147	49%	152	51%	299
20	134	48%	144	52%	278
25	124	48%	135	52%	259
30	113	47%	130	53%	243
35	104	45%	126	55%	230
40	94	45%	116	55%	210
45	83	44%	107	56%	190
50	67	42%	91	58%	158
54	57	43%	77	57%	134
58	49	43%	64	57%	113

8.3.5 BP arm group comparisons

Comparison of the baseline variables between the Telemonitoring group and the Control group, detailed in Chapter 7, suggested some differences between the groups, at least in terms of the data available. Analyses 5 and 6 used all of the BP arm participants, with missing values handled by multiple imputation before comparing the groups. Analysis 7, on the other hand, involved restricting the participants to those who measured for at least 208 days to try to make the Telemonitoring and Control groups more alike, though in the end, this did not appear to have been achieved when we looked at the baseline tables.

At first glance, the results in Table 8.6 appear to provide some support for the idea that the telemonitoring program caused an average drop in blood pressure in the intervention group. However, it is uncertain whether these are of sufficient magnitude to be regarded as clinically meaningful. For example, a recent meta-analysis[874] focused on a 10 mmHg reduction in SBP as their criteria, while another study[875] suggested a reduction of at least 20 mmHg in SBP or 10 mmHg in DBP could be regarded as clinically meaningful.

Part of the precision that is indicated by the p-values comes from the reasonably large sample size. It is nevertheless a stronger result in support of the intervention than we saw for the Glucose arm, though in common with that arm there remains much uncertainty.

Table 8.6 BP arm comparisons between the Telemonitoring and Control groups

				Results for both groups		
Analysis	**Group**	**N**	**N with outcome missing***	**BP type**	**Estimate† of difference between groups (95% CI)**	**P**
5.	TM	1,429	541 (38%)	SBP	-8.0 mmHg (-9.3, -6.7)	<0.0001
	Controls	1,259	198 (16%)	DBP	-4.1 mmHg (-4.9, -3.3)	<0.0001
6.	TM	1,429	541 (38%)	SBP	-7.7 mmHg (-8.9, -6.5)	<0.0001
	Controls	1,259	198 (16%)	DBP	-4.0 mmHg (-4.7, -3.3)	<0.0001
7.	TM	773	0	SBP	-6.6 mmHg (-8.0, -5.3)	<0.0001
	Controls	617	0	DBP	-3.1 mmHg (-3.9, -2.3)	<0.0001

* Before multiple imputation; † Mean difference in outcomes, adjusted for baseline values

8.3.6 BP arm sensitivity analyses

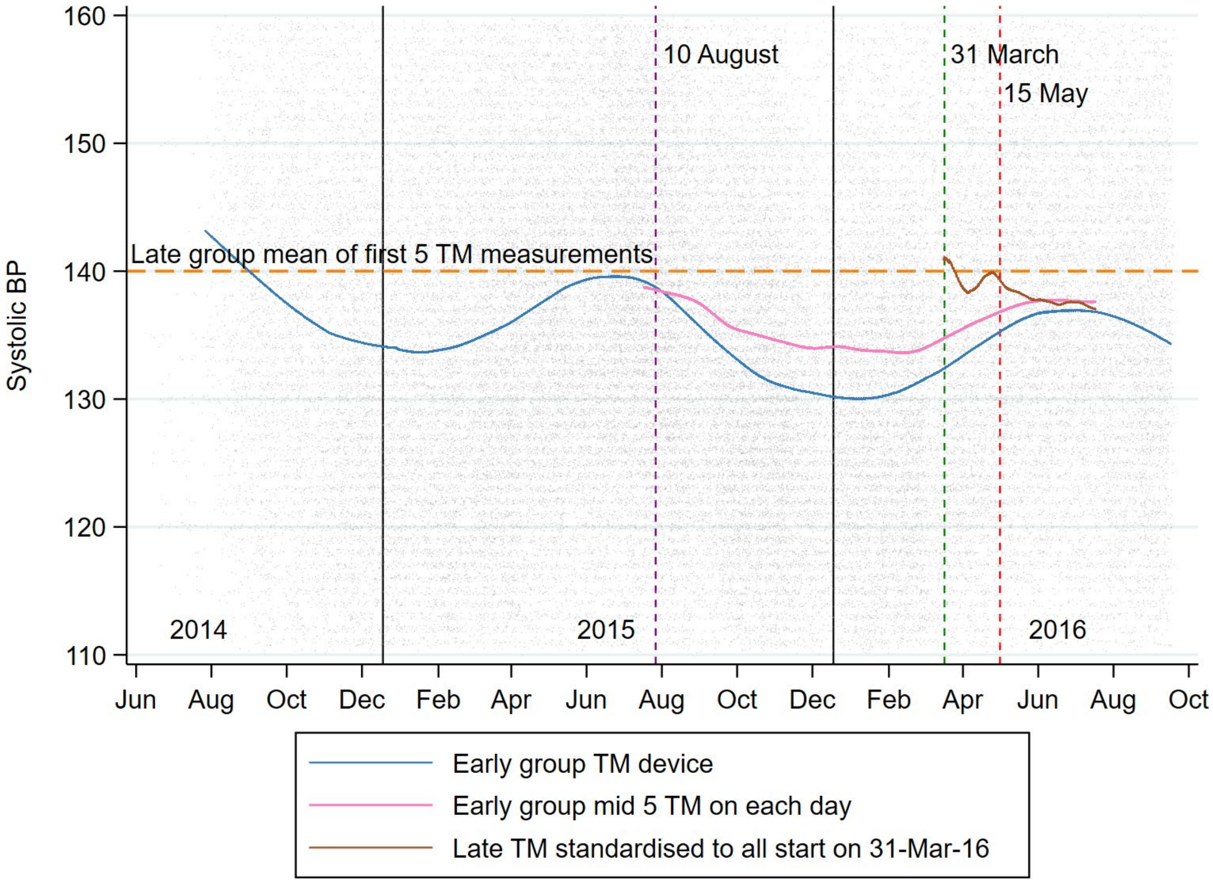
Instability of gap between group outcomes over time

The Lowess lines in Figure 8.12 suggest that:

1) The mean of the device measurements in the Telemonitoring group (blue line) varied considerably over the two years. Possible reasons for this include the effect on mean blood pressure of certain types of people dropping out; and participants becoming used to the intervention so that it no longer prompted lifestyle changes or other factors that might have affected their blood pressure.
2) If the analysis period had, by chance, been different, for example 23 Nov 2015 to 31 Oct 2016 to include some people's initial Telemonitoring group measurements that occurred after 31 July, the mid-point would then be 15 May 2016 (vertical red dashed line) instead of 31 March (vertical green dashed line), and the gap between the first 5 measurements of the Control group and the Telemonitoring group's mid-5 values (between the horizontal orange dashed line and the pink line) would be noticeably smaller. Likewise, if the Control group had, by chance, commenced enrolling in August 2015, one year after the Telemonitoring group's first device measurements, and the comparison was made at this time point (vertical purple dashed line), the gap would also have been considerably smaller.

Hence, the size of the gap estimated by the primary analysis and reported in Table 8.6, is probably larger than it might have been because of when it happened to occur by chance.

Figure 8.12 Lowess lines highlighting variability of gap between BP arm groups



Early group = Telemonitoring group; Late group = Control group; TM = telemonitoring device measurements

The possibility of pre-TM BP measurements as Control group outcome

Unlike in the Glucose arm, where no pre-Telemonitoring trial blood glucose measurements were available, in the BP arm, some pre-trial blood pressure measurements were occasionally reported by participants during phone calls with Healthways nurses, as part of the My Health Guardian program. Referring to these as 'Reported' measurements, we explored the possibility that these values would serve as better Control group measurements than the first 5 intervention measurements. To remove the potential impact that staggered enrolment might have had on average BP measurements, we first standardised the Telemonitoring group's device measurements so that all measurements were shifted back in time with the effect that the first measurement occurred on 1 July 2014. The Control group's Reported values were left with the same date. To avoid confusion, the week starting 1 July 2014 was then called week 0. We used a total of 72 weeks as this was the number of full weeks from 1 July 2014 to 23 November 2015.

We must first make the assumptions that

1. The Control group participants with Reported values were exactly the same type of people as those in the Telemonitoring group
2. All BP measurements were accurately reported, both from the Telemonitoring device and reported by participants over the phone

Using kernel-weighted local polynomial regression to provide a line of best fit through weekly means of each group's BP values, Figure 8.13 shows some overlap of the lines. We included 99.9% confidence intervals because the amount of missing data suggests that significant unmeasured confounding might exist that the random-error-only 95% confidence intervals do not take into account.

If the Telemonitoring device measurements and Reported measurements were of similar accuracy, we would expect both lines in Figure 13 to start at the same point. And if the intervention caused better blood pressure over time compared to not using the intervention, we would expect the blue Telemonitoring curve to slope down more steeply, or at least stay below the Control group's curve, with the gap getting wider as the weeks progress. But because the two types of measurements did not start with the same mean blood pressure, the trajectories are difficult to interpret. We could potentially shift the Control group's measurements higher so that the starting values are the same, and then compare the trajectories. However, the small number of participants (examined below) suggests this would probably give a biased result because it assumes that all participants' Reported measurements were in error by the same amount.

Figure 8.14 examines the number of participants contributing measurements to Figure 8.13. Comparing the Y-axis scales of each bar graph suggests that only a small number provided Reported measurements compared to the number contributing Telemonitoring device measurements each week. In total, there were 1303/1429 (91%) Telemonitoring device measurements and 481/1259 (38%) Reported measurements in this timeframe.

Potential bias from measurement error

Jump from last Reported BP to first Telemonitoring device BP

The Lowess curves in Figure 8.15 and Figure 8.16 suggest that the initial Telemonitoring device measurements (the beginning of the line on the right in each graph) were higher, on average, than the participant's pre-trial Reported measurements. We compared the last few Reported measurements (5 or less) with the first few Telemonitoring device measurements (5 or less) (Table 8.7) and there is a clear difference, suggesting that the Reported values might have been lower than in reality – a possibility because the measurements are reported over the phone to Healthways nurses and a reporting bias has been reported in the literature for home monitored BP.[876] Alternatively, initial measurements taken with the Telemonitoring device might have been higher on average, than in reality, perhaps reflecting a 'white coat' type of effect.[877] It is not uncommon to discard measurements from the first day of home blood pressure monitoring because it is believed they are often higher than a patients' normal BP.[878,879] Or both may be at work in producing this difference. These possibilities provide one alternative explanation for part or all of the effect of the intervention.

Figure 8.13 TM (Early group) device and Control (Late group) reported weekly means

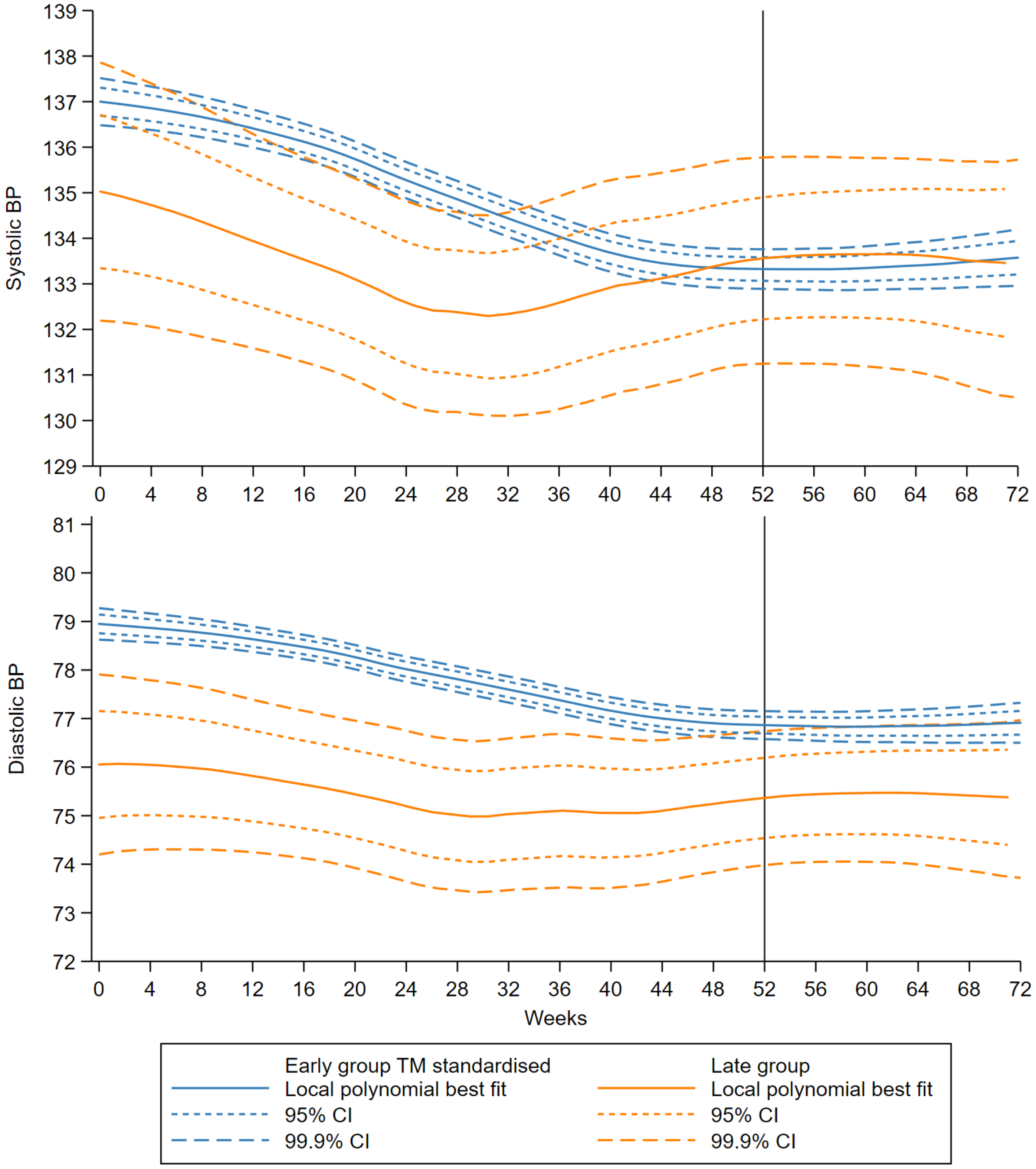


Figure 8.14 Participants contributing to the weekly means above (note Y axis scales)

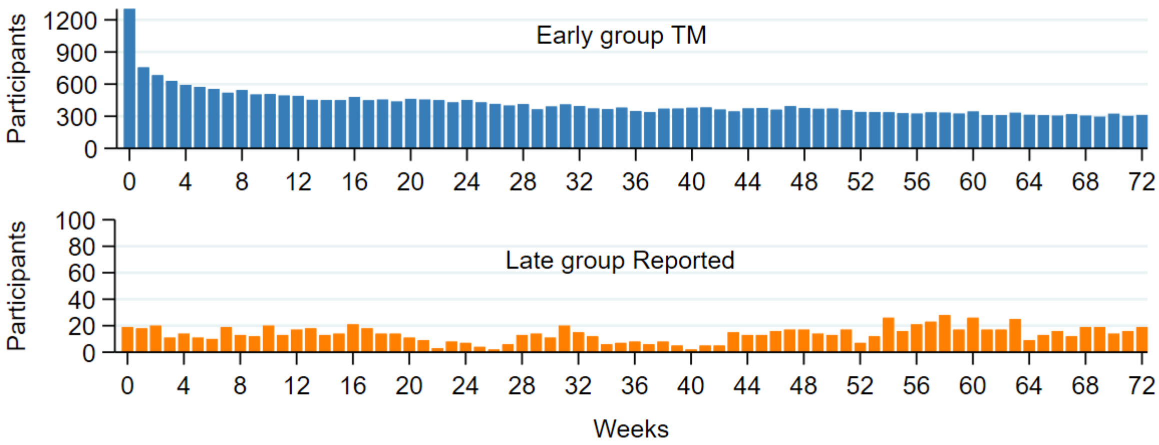
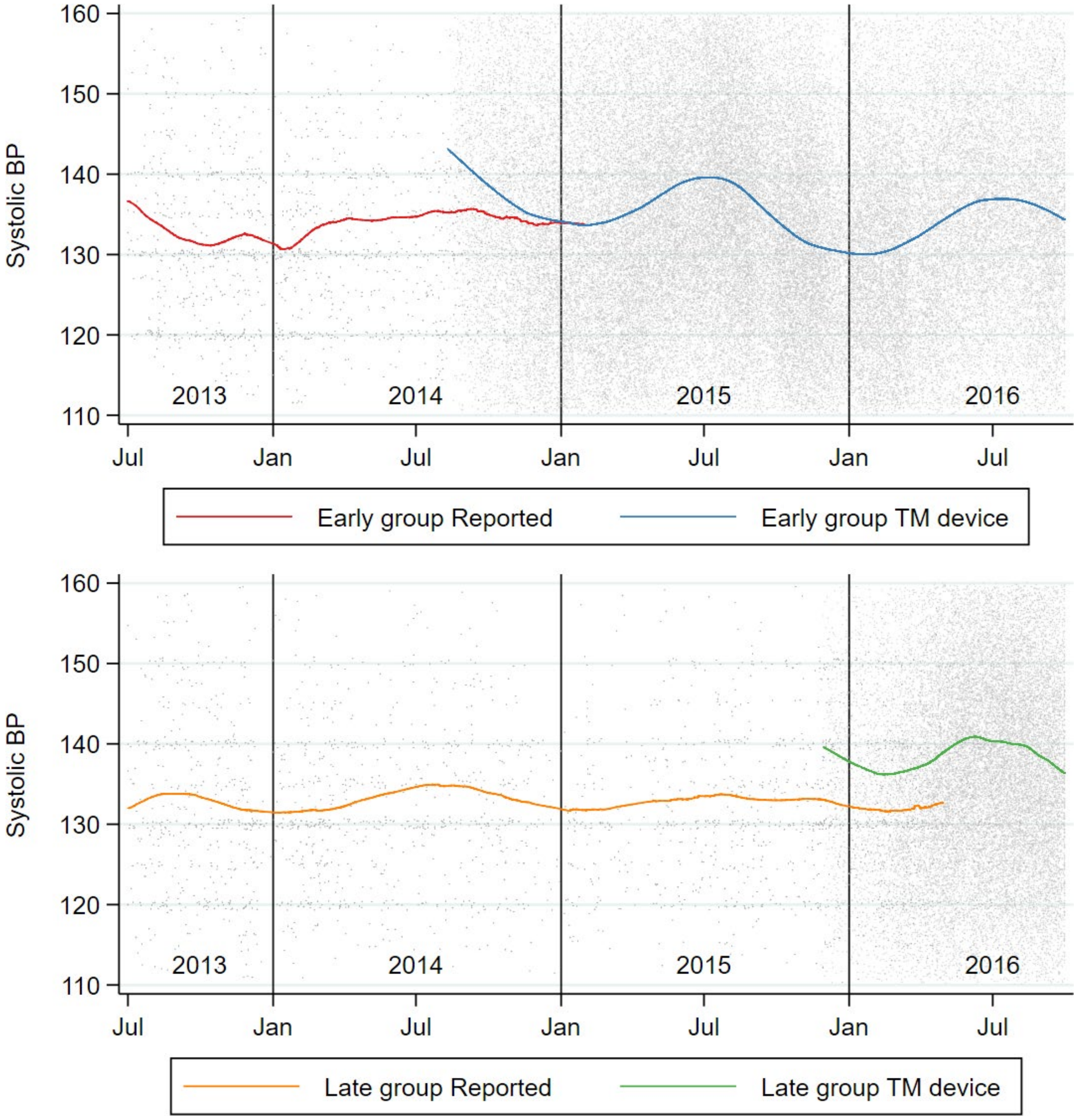


Figure 8.15 Systolic BP Lowess lines of best fit for Reported and Telemonitoring values



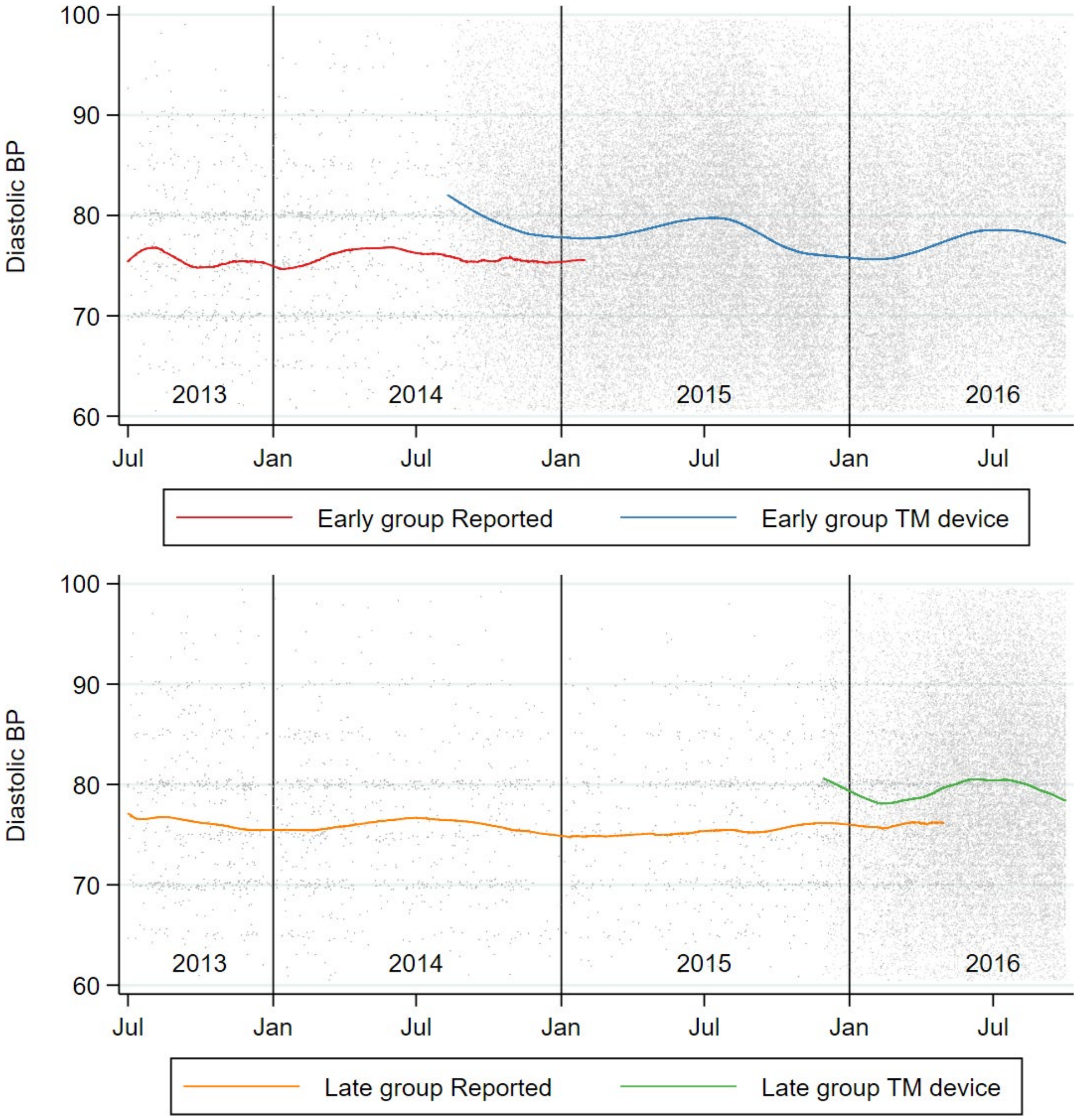
Early group = Telemonitoring group; Late group = Control group

Table 8.7 Last 5 Reported compared to first 5 Telemonitoring measurements

	Telemonitoring group		Control group	
	Mean SBP	Mean DBP	Mean SBP	Mean DBP
Last 5† Reported measurements	133	76	133	76
First 5† Telemonitoring measurements	139	80	140	81

† less if 5 not available

Figure 8.16 Diastolic BP Lowess lines of best fit for Reported and Telemonitoring values



Early group = Telemonitoring group; Late group = Control group

We examined the initial Telemonitoring device measurements in Table 8.8 to see if they suggest an initial spike before quickly settling down, presuming that the participants quickly became comfortable in using the device. However, the measurements appear reasonably consistent with each other. The mean number of days between each successive measurement of all participants is shown in Table 8.9 and it suggests that quite a few participants measured with gaps of a week or more between each use. It is possible that this may have prevented complete relaxation developing when using the device.

Table 8.8 Initial Telemonitoring device BP mean measurements

Telemonitoring device measurements	Telemonitoring group		Control group	
	SBP	DBP	SBP	DBP
All participants†				
Mean of 2nd to 6th	*139.4*	*80.1*	*140.1*	*81.2*
1st	138.6	80.7	140.1	81.7
2nd	139.3	80.4	140.5	81.3
3rd	139.5	80.4	139.5	81.1
4th	138.7	79.7	139.5	80.9
5th	139.4	79.7	139.3	80.9
6th	139.0	79.9	138.8	80.3
Analysis 3 & 5 participants‡				
Mean of 2nd to 6th	*137.6*	*79.6*	*138.3*	*80.3*
1st	136.5	79.7	137.5	80.5
2nd	137.6	79.8	139.0	80.5
3rd	138.1	80.1	138.1	80.2
4th	137.4	79.3	138.2	80.4
5th	137.6	79.3	138.3	80.5
6th	137.2	79.3	137.7	79.8

† 1241 participants had at least 2 Telemonitoring device measurements; 188 had 0 or 1

‡ 772 participants had at least 2 Telemonitoring device measurements

Table 8.9 Days between first 6 Telemonitoring device measurements (all participants)

Telemonitoring device measurement interval	Days between measurements			
	Telemonitoring group		Controls	
	Mean	Median	Mean	Median
1st to 2nd	13 days	5 days	8 days	5 days
2nd to 3rd	13 days	6 days	8 days	5 days
3rd to 4th	15 days	6 days	9 days	5 days
4th to 5th	15 days	6 days	8 days	5 days
5th to 6th	14 days	6 days	7 days	4 days

If the initial measurements of both groups were found to be higher than the participants mean BP was in reality, then the differences found in the group comparison analyses might have overstated the effect because we compared the non-initial measurements of the Telemonitoring group with the initial measurements of the Control group.

Accuracy of Reported measurements

With concerns over the accuracy of measurements, we next examined the Reported and Telemonitoring device values after first standardising all participants' measurements so that it was as if every participant recorded their first Telemonitoring device measurement in the same week (Week 0). Lowess lines of best fit are shown in Figure 8.17 and Figure 8.18 using these values. The Reported values appear to trend higher in the weeks before the first Telemonitoring device measurement. It is unclear why this might have happened, though the measurement number is small. It does, however, increase doubt over the accuracy of the Reported measurements.

Once Telemonitoring had started, some Reported measurements were still recorded, and the curves of those measurements largely follow the Telemonitoring curves which suggests a mixing of Reported and Telemonitoring values. To check on this, Table 8.10 shows the proportion of Reported measurements that were the same as the previous Reported or Telemonitoring device measurement and suggests that more than half of the Reported measurements after Telemonitoring had started are, in fact, the most recent Telemonitoring measurement. This is perhaps not surprising, however, as Telemonitoring measurements were those readily available to participants during phone calls to nurses.

Table 8.10 Proportion of Reported measurements that were the same as the previous Reported or Telemonitoring device measurement

	Reported BP Date before Telemonitoring Start	Reported BP Date after Telemonitoring Start
Telemonitoring group	123 / 1,692 (7%)	1,671 / 2,653 (63%)
Control group	660 / 2,851 (23%)	799 / 1,338 (60%)

Figure 8.17 Lowess lines of best fit comparing standardised Telemonitoring group TM device and Reported values

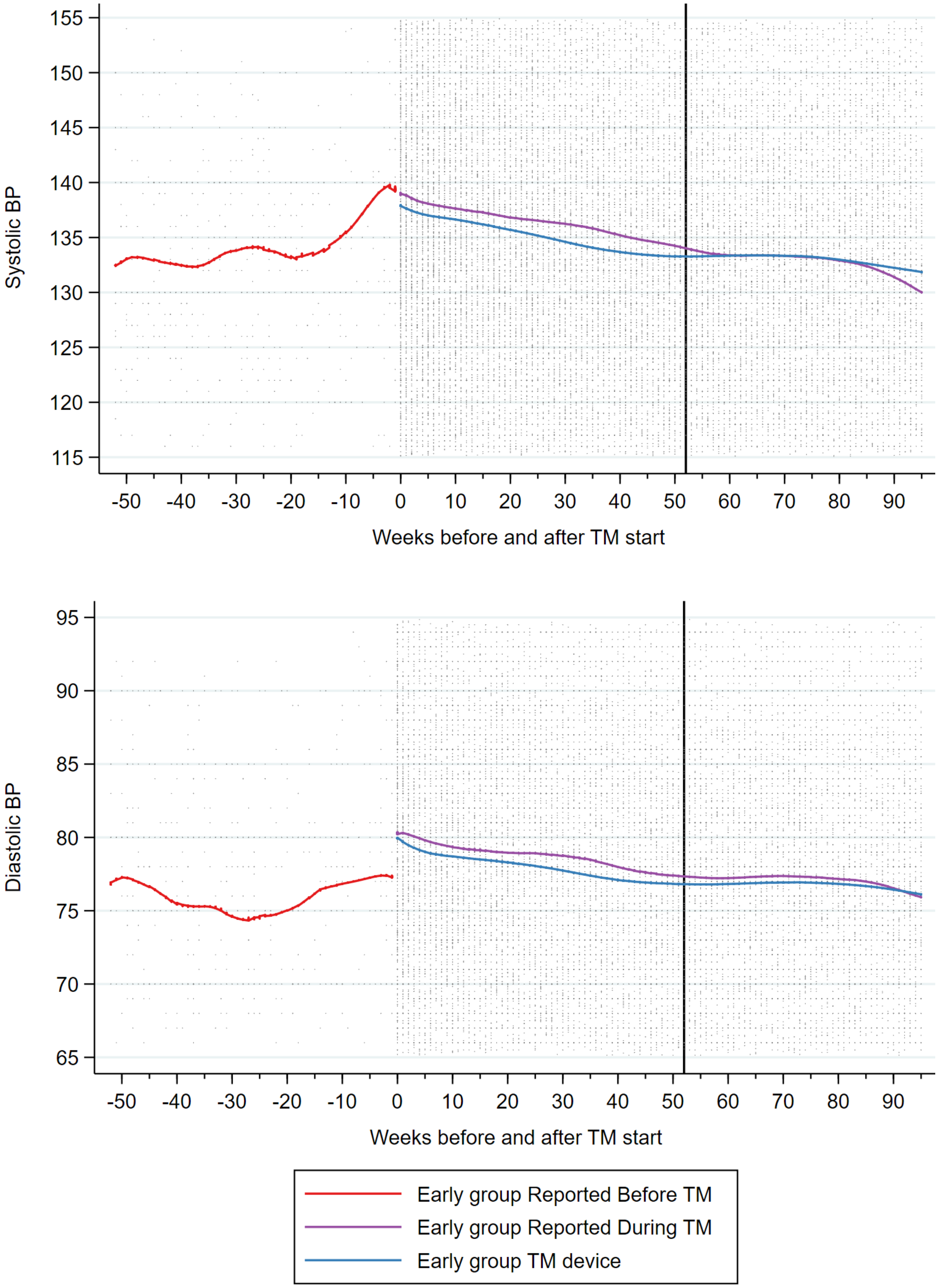
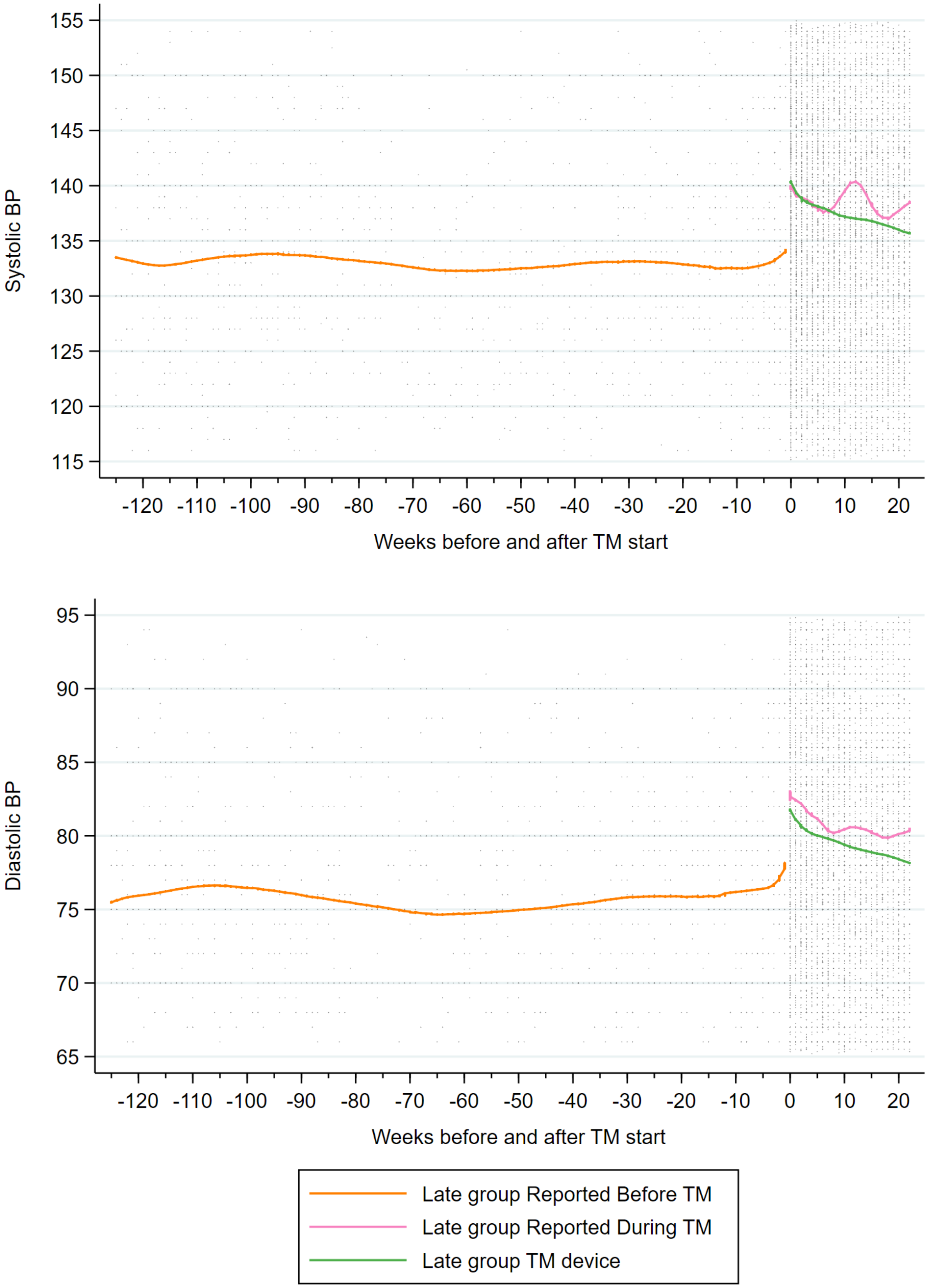


Figure 8.18 Lowess lines of best fit comparing standardised Control group Telemonitoring and Reported values



However, on inspection of the data, almost a quarter of the Reported measurements before Telemonitoring started were also repeats. This suggests that using initial Telemonitoring device measurements as baseline values for the Control group might provide more valid estimates than using the Reported values.

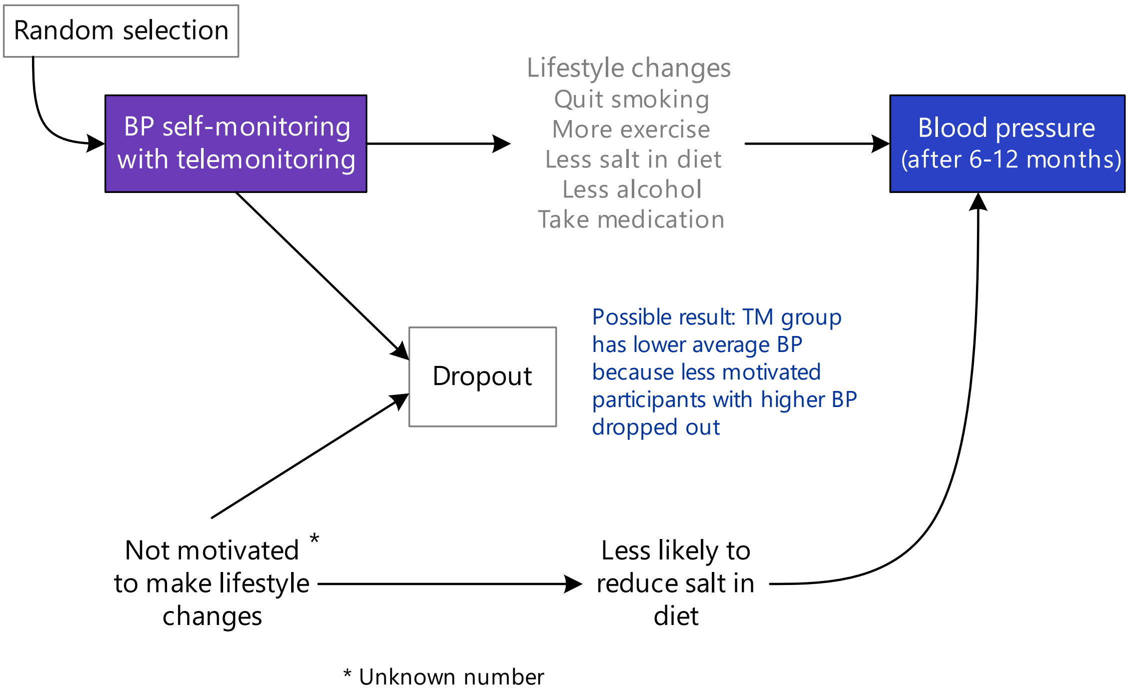
8.3.7 Alternative explanations

One of our aims was to help those involved in the trial develop a more complete understanding of the information the trial can provide and the uncertainty that needs to be taken into account. To fully understand the level of uncertainty that exists around research findings, it is essential to consider any plausible alternative explanations for part or all of the observed effects of the intervention being investigated. For example:

- Participants who were less likely to modify their lifestyle to lower their blood glucose or blood pressure may have been more likely to drop out of the Telemonitoring group because they lacked the motivation to self-measure.
    - ↳ Those remaining in the Telemonitoring group would have been more willing to make the necessary lifestyle changes and so their measurements improved. Similarly motivated participants in the Control group provided only their initial measurements so later dropout was not a problem.
    - ↳ This suggests that the Telemonitoring group participants who provided outcome data for the analysis might have been more motivated to make lifestyle changes than Control group participants, regardless of the effect of the intervention, and this might partly or wholly explain any difference in the outcomes observed between the two groups.
        - ▪ The causal diagram in Figure 8.19 illustrates one specific example.
- One assumption made in both analyses was that the Control group’s outcome of mean blood glucose or blood sugar could be approximated by the first few measurements taken by using the intervention and hence, we assumed that the intervention had no effect in that initial period of time. If it did, then the bias would be toward the null and

the real effect of the intervention was greater than that measured. While the BP arm had this problem from the beginning, the Glucose arm was originally designed with HbA1c as the outcome which would have avoided this possibility. While not an alternative explanation as such, because it suggests the effect might be greater, it is important to identify possible biases that might influence the results in either direction.

Figure 8.19 Causal diagram showing one alternative explanation for the BP arm results


Random selection
BP self-monitoring with telemonitoring
Lifestyle changes
Quit smoking
More exercise
Less salt in diet
Less alcohol
Take medication
Blood pressure (after 6-12 months)
Dropout
Possible result: TM group has lower average BP because less motivated participants with higher BP dropped out
Not motivated * to make lifestyle changes
Less likely to reduce salt in diet
* Unknown number

- It is also possible that the intervention's largest effect was, in fact, to raise blood pressure and possibly blood sugar when it was first used. The 'white coat' effect is fairly well known with respect to blood pressure, where the mild stress experienced by some patients when they visit their doctor causes a small increase in their blood pressure.[877]

  ↳ However, such an effect has also been reported for blood glucose.[880,881] Physiologically, this is plausible given that the hormones released during the stress response stimulate the liver to raise blood sugar.[882]

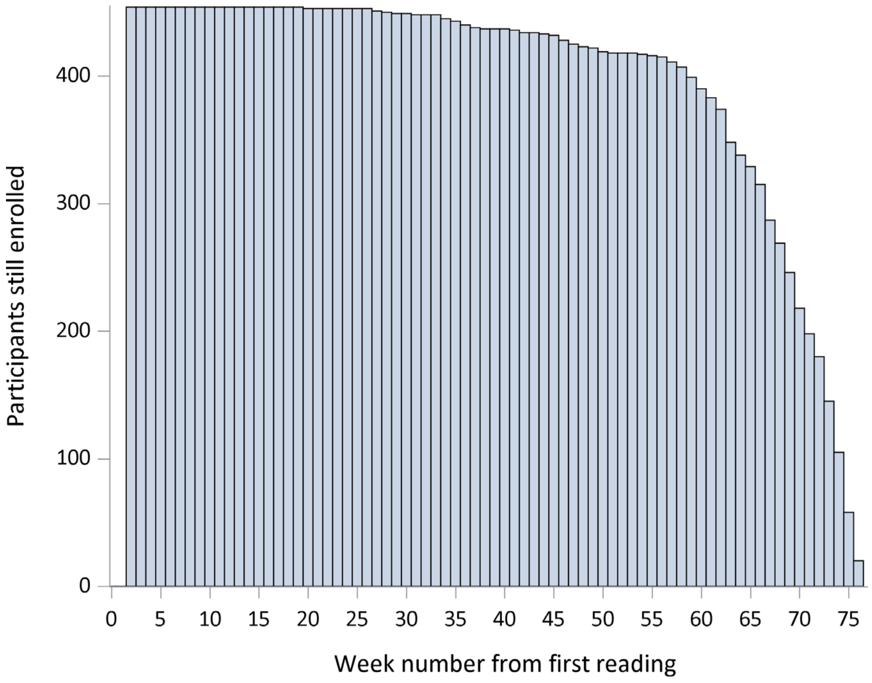
  ↳ With initial intervention measurements used as the Control group's outcome in both arms, while the Telemonitoring group's measurements were from a period long after the start, it is plausible that an initial increase may partly or wholly explain the difference observed between the two groups.

- Another assumption is that the treatment outcome at 12 months was not greatly different to the outcome a few months earlier or a few months later. This assumption appears not to have been met, however. Yet it is important because if such variation in the outcome is ignored, with just the difference at 12 months reported, many will have the impression that the difference was relatively stable.

8.3.8 Parametric g-formula and the possible effect of measurement frequency

Some participants enrolled months later in the trial than others, so the first step was to determine the maximum number of weeks we might use where every participant could potentially have measurements for that many weeks. In Figure 20, a sharp drop in the number of participants still enrolled appears to begin around the 58-week mark, or approximately 14 months.

Figure 8.20 Participants still enrolled each week after first glucometer use (TM group)



Evidence for time-dependent confounding potential

Before using the g-formula to handle possible time-dependent confounding, we first tried to determine if such confounding might be present. Figure 8.21 and Figure 8.22 use linear regression to plot the relationship between measurement frequency and mean blood glucose with one from the week before, and then the reverse. These causal relationships are represented by the green and blue arrows in the causal diagram. Both plots suggest that time-dependent confounding is possibly small enough to be ignorable.

Figure 8.21 Mean glucose previous week and change in measurement frequency this week

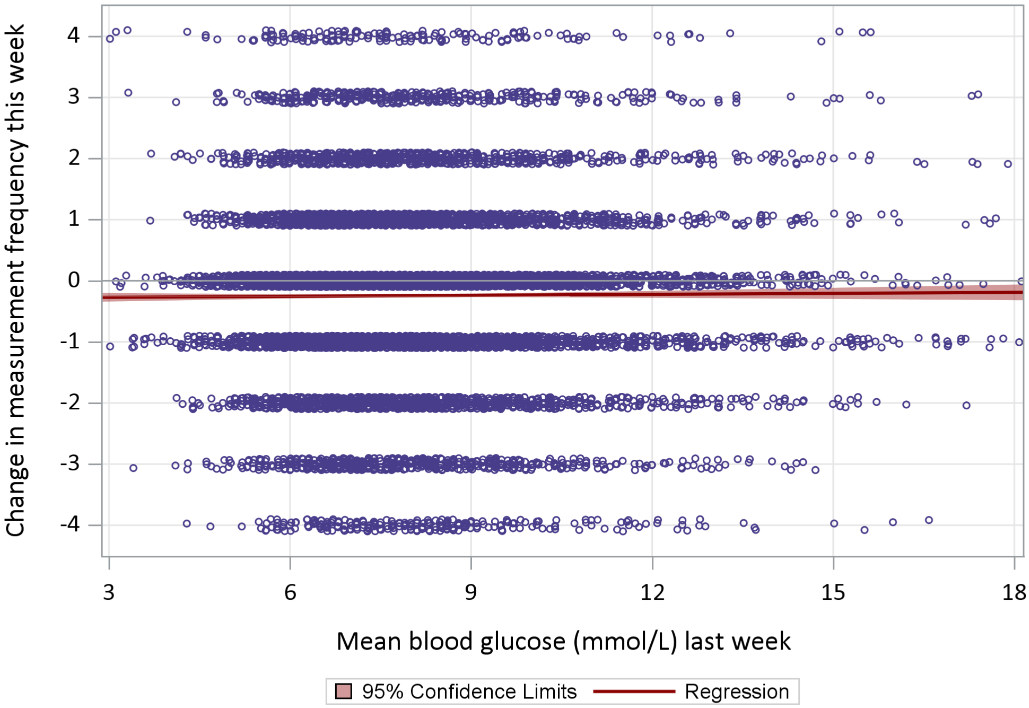
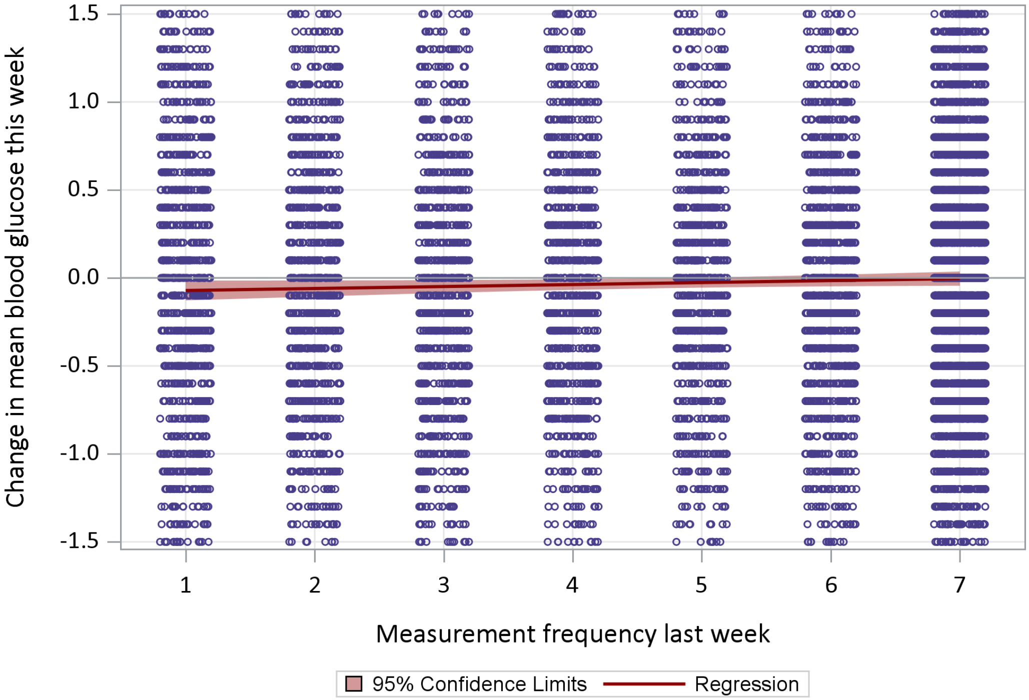


Figure 8.22 Measurement frequency previous week and change in mean glucose this week



The estimates and p-values (Table 11) from the predicted linear regression lines in Figure 8.21 and Figure 8.22 in both cases suggests that the relationship is weak.

Table 8.11 Linear relationship between meas. frequency and glucose with one lagged 1 week

Previous week	Current week	Estimate	95% Confidence Limits		P-value
Mean glucose	M. frequency	0.005	-0.007	0.018	0.40
M. frequency	Mean glucose	0.011	-0.002	0.024	0.09

G-formula used to adjust for possible time-dependent confounding

The g-formula analysis was limited to the 113 (22% of 512) participants with blood glucose values in every one of the 14 months that were covered. And similarly, only complete covariates were included in the model. The included participants were from the Telemonitoring group only because the Control group had no more than a few months' worth of data at the time of the analysis. Naturally, any inference on whether measurement frequency makes a difference is only relevant to those who continue to measure. Unfortunately, this was not many.

The g-formula compares the predicted outcome after several 'intervention' variations are implemented. Using the GFORMULA SAS macro[883] we compared the predicted mean blood glucose level between a measurement frequency of 30 days per month on which measurements occurred, against only 1 day per month. In other words, we compared the effect of measuring blood glucose every day for 14 months against measuring only once per month. These could also be loosely described as two hypothetical treatment strategies; defined for the purpose of estimating the effect of frequent or non-frequent measuring. The variables used are listed in Table 8.12 with the code that makes use of the SAS macro shown below the table.

One last feature of the dataset is that the first month during which the first measurement took place was not included. This is because the date that the first measurement occurred might be the 1st, and so the whole month potentially had measurements, or it might be later in the month and so the total reading count would be incomplete.

Table 8.12 Variables used with the GFORMULA SAS macro

Variable	Role in model	Description
id	ID	ID number of participant
time	Time	Month number from first measurement
Age	Fixed	Age on 1 July 2014
Sex	Fixed	Sex
DiabType	Fixed	Diabetes type
HTN	Fixed	Hypertension
HLD	Fixed	Hyperlipidemia
CVD	Fixed	Cardiovascular disease
Arthritis	Fixed	Arthritis (any type)
BackPain	Fixed	Back pain
WalkPain	Fixed	Walking pain
EyeProb	Fixed	Eye problem
Insulin	Fixed	Insulin
NumT2Drugs	Fixed	Number of diabetes drugs
PainDrugs	Fixed	Pain relief drugs
RLBase	Fixed	Baseline risk level
RL	Time-varying	Risk level after 1st measurement
RL_l1	Lag1 time-varying	Risk level 1 month before RL month
RL_l2	Lag2 time-varying	Risk level 2 months before RL month
RL_l3	Lag3 time-varying	Risk level 3 months before RL month
meascount	Time-varying	Measurement frequency by month
meascount_l1	Lag1 time-varying	M. frequency 1 month before meascount month
meascount_l2	Lag2 time-varying	M. frequency 2 months before meascount month
meascount_l3	Lag3 time-varying	M. frequency 3 months before meascount month
glucofinal	Fixed	Final mean glucose after 14 months
glucomean	Time-varying	Mean glucose of each month
glucomean_l1	Lag1 time-varying	Mean glucose 1 month before glucomean month
glucomean_l2	Lag2 time-varying	Mean glucose 2 months before glucomean month
glucomean_l3	Lag3 time-varying	Mean glucose 3 months before glucomean month

SAS code used to call GFORMULA macro:

```
%let interv1 =
  intno = 1,        /* intervention number */
  nintvar = 1,      /* number of intervened on variables. */
  intlabel = 'meascount min of 30 per month',       /* for output */
  intvar1 = meascount,    /* variable undergoing intervention */
  inttype1 = 2,     /* 1=static, 2=threshold, 3=fixed, 4=prev val, -1=user def */
  intmin1 = 30,     /* min value for inttype=2 if interv value is below this */
  inttimes1 = 0 1 2 3 4 5 6 7 8 9 10 11 12 13; /* times intvar# intervened on */

%let interv2 =
  intno = 2,
  nintvar = 1,
  intlabel = 'meascount set to 1 per month',
  intvar1 = meascount,
  inttype1 = 1,
  intvalue1 = 1,
  inttimes1 = 0 1 2 3 4 5 6 7 8 9 10 11 12 13;

%gformula(
  data=monthlyrcountwithlag,
  id=id,
  time=time,        /* time point variable (must begin at 0) */
  timepoints=14,    /* number of time points */
  timeptype=concat,       /* choices: conbin, concat, conqdc, concub, conspl */
  timeknots = 1 2 3 4 5 6 7 8 9 10 11 12 13,

  outc=glucofinal,        /* outcome variable */
  outctype=conteofu,      /* outcome type: binsurv (time-varying failure),
      bineofu (binary end of follow-up), conteofu (contin. end of follow-up). */
  fixedcov=RLBase Age Sex DiabType HTN HLD CVD Arthritis BackPain WalkPain
  EyeProb Insulin NumT2Drugs PainDrugs,       /* predictors not predicted */

  ncov=3,    /* number of (time-varying) covariates to be estimated */
  cov1=meascount,   /* covariate 1 name */
  cov1otype = 3,    /* defines cov1 outcome type for regression procedure */
  cov1ptype = lag3bin,    /* how cov1 history will be incl. in regress. models */
  cov2=glucomean,
  cov2otype = 3,
  cov2ptype = lag3bin,
  cov3=RL,
  cov3otype = 5,
  cov3ptype = lag3cat,

  numint=4,         /* number of interventions called by the INTERV macro*/
  seed= 9458        /* random numbers seed */
  );
```

After running the macro in SAS 9.4, the results (Table 8.13) show that there is considerable overlap in the 95% confidence intervals and thus they support the previous results that suggest no meaningful relationship exists between measurement frequency and mean blood glucose level.

Table 8.13 Predicted mean final blood glucose level under two possible interventions

Observed mean= 7.78

Measurement frequency (Days per month)	Estimate of final mean blood glucose level (95% CI) (mmol/L)
1	7.44 (6.78, 7.94)
30	7.35 (7.26, 8.16)

8.4 Final thoughts

Using blood glucose as the outcome, some evidence was found that suggests the Telemonitoring program resulted in lower mean blood glucose levels in participants who continued to use the glucometer. It could be argued that this is not surprising and not particularly relevant that the statistical results do not relate to the kind of people who dropped out of this program early or did not even start measuring. But one potential scenario is that some of the members with certain characteristics who did not make use of this program following enrolment, might have measured if it had been set up a little differently, and hence they might still use a service that is very similar at another time. If they also have characteristics that mean they do not respond to the program with lowered mean blood sugar, or instead, respond better than the average response that was recorded by members who did measure for a sufficiently long time, then either way there is the possibility of bias in the estimates with respect to the population they are thought to relate to.

Conclusions provided to HCF

Glucose:

The balance of the available evidence seems to weigh on the side of supporting the program, though the effect on blood glucose levels is probably small. The lack of available data has added much uncertainty to any conclusions arising from these results. If a similar trial were to be conducted in the future the most important lesson learned from this one would be to ensure (somehow) that most participants had HbA1c values recorded, both at the beginning and at the end, each within a short time frame that is the same for both groups.

BP:

Evidence from this trial suggests the Telemonitoring program may have reduced the mean blood pressure of the Telemonitoring group participants compared to the Control group. However, this interpretation needs to be viewed with caution and allow for considerable uncertainty because of the level of missing data, use of the intervention to generate the Control group outcomes, and concerns over data accuracy.

Chapter 9
Strategies and resources for less biased causal inference

9.1 Understanding causal inference, bias and uncertainty

For many health problems today where a cure is the ultimate goal like the common cold and permanent paralysis, no interventions exist that can help to restore full health. For many other conditions, more typical is simply an improvement, or none at all, with individual responses varying widely. And where interventions do have some success, they often come with side-effects so that new interventions are ever desirable. To get better health interventions sooner requires research that provides accurate answers to causal questions, where an example might be 'does intervention X reduce illness A' or 'can intervention Y cause side-effect B'. Thus, progress in health intervention research depends on the accuracy of causal inferences. And that is determined by how we handle the many possible sources of bias that can shift such inferences away from the underlying truth; as well as the biases that lead us to perceive more certainty than is really justified, both in the information we make those inferences from, and in the inferences themselves.

Like all areas of science, however, research in health has struggled to reduce the level of bias and improve the standard of the causal inferences researchers make. Despite regular criticism over the way researchers deal with bias and uncertainty, especially with regard to statistics, it is not clear that progress has been made over the last 40 years. This thesis sought to understand the reasons behind this enduring problem, with the hope that a deeper understanding of causal inference, bias and uncertainty—the concepts these words refer to, their nature, and how they are dealt with — can clarify the strategies to follow for greater improvement in health intervention research.

To enhance this understanding, an important component of this thesis is the drawing together of relevant knowledge from not only the discipline of health, but also of statistics, philosophy, linguistics, and cognitive psychology. This has made clearer the many factors that influence, and potentially bias a causal inference, itself a cognitive process. It also seemed that a more precise and useful definition of 'a causal inference' was needed, which we proposed as: 'a conclusion that the evidence available supports either the existence, or the non-existence, of a causal effect'. And while not part of this definition, some sense of the uncertainty that surrounds the inference would tend to be part of this process, especially when in relation to research.

We explored various frameworks that can be used when considering causal questions, with the use of more than one appearing optimal, depending on the perspective. For example, the potential outcomes framework is very useful when considering an analysis methodology, but the sufficient-component cause model might help when constructing a causal diagram.

To understand why causal inference, bias and uncertainty are approached and handled as they currently are, we examined the evolution of these concepts in health research over the last few centuries, with their history contributing to an overall understanding of how problems have developed. Karl Pearson's disapproval of talking about the 'causes' of phenomena, the use of the word 'bias' by mathematical statistics for an idealised situation, and the development of what some call null-hypothesis significance testing, have greatly influenced how causal inferences are often made in health research.

The importance of using a classification system for bias was examined, with the suggested benefits relating to ease of recall, learning and communication. But with the unlikely chance that a consensus would ever be reached on terminology, the structural classification of bias that uses causal diagrams was promoted as possibly the only way out of this problem.

The first three chapters provided many of the details that are needed for an understanding of causal inference, bias and uncertainty, as covered by the literature in health and statistics. Chapter Four began with an overview of the evidence that bias is a problem in health research, because the motivation for change will not exist if statisticians and researchers are not aware of the problem.

9.2 The importance of how people think

The rest of Chapter Four discussed findings in cognitive psychology that have considerable relevance to the problem of ongoing bias in health research. Within this context, the content of Chapter Two and Chapter Three can also be better understood, such as the influence that our built-in desire for cognitive ease has likely had on the development of null-hypothesis significance testing, where decisions are considerably easier with a binary significance cutoff.

In fact, the ‘principle of least effort’, one of many names by which the desire for cognitive ease is known, leads to many recommendations that are often understood in some fashion, but where knowledge of its fundamental role in how people think adds new weight. For example, to change the way researchers use statistics, new ways of analysing data need to be sufficiently easy to learn and use, else they will only be used by a small minority of people such as the mathematical statisticians who are familiar with them. Likewise, although the use of p-values is often criticised, any suggested replacement would probably need to be easy to understand. Hence, statistics like the likelihood ratio seem unlikely to take hold, with efforts to improve research perhaps more likely to work if they focus on how p-values are used, a statistic that almost all researchers are at least familiar with.

Underlining the relevance of this information from cognitive psychology is that these cognitive biases occur mostly below conscious awareness. Likewise, the fact that everyone is susceptible to these biases, with higher intelligence only providing a little protection, and then only for some biases. But a few of these biases make combating them in research difficult, such as a tendency for people to accept the first explanation that comes to mind (take-the-first heuristic), or a bias blind spot for our own cognitive biases but not for the biases of other people, or a tendency to seek arguments for our beliefs rather than the objective truth (myside bias), and there is also our bias for causal explanations (causality heuristic).

9.2 The importance of how people think

One of the few debiasing techniques with some evidence of a benefit is anything that promotes thinking about alternative explanations for the results. This is especially important when forming conclusions at the end of a study. It also highlights one of the primary benefits of constructing causal diagrams.

Chapter Five gave us a snapshot of the statistical methods being used to control for bias in health intervention cohort studies. The use of propensity scores by a third of the studies suggests that there is a widespread awareness of the need to improve the methodology used for causal inference, even though propensity scores may or may not have been the better approach in each case. But the other more recently developed methods that focus specifically on causal inference, including causal diagrams, were seldom used. Also, the underreporting of how missing data was handled suggests that many still lack a full appreciation for the potential bias that missing values can produce.

The communication of causal inferences was examined in Chapter Six, beginning with a discussion on why we think all conclusions from health intervention studies can be usefully classed as causal inferences. We also discussed some of underlying drives that influence our choice of language, such as the desire for respect and social status, and how this helps to motivate spin. Also contributing is our built-in bias for causal explanations.

One of the main findings from our review of causal language is that creating an algorithm that can automatically rate the strength of causal language no longer seems to be an achievable goal. This is because language depends more heavily on context than we had previously realised and the number of possible contexts of words in a conclusion is very high. We also found that articles using either multivariable regression, propensity score methods or a sensitivity analysis were more likely to be cautious in their use of causal language in conclusions. Partly, this may result from such methods helping to bring alternative explanations to mind when considering their interpretation of the results, leading in turn to more caution when judging causal effects.

9.3 Strategies and resources

From the point of view of this thesis, one the main findings from conducting the case study is that methods like multiple imputation and the g-formula are not at all easy to do properly. While this will not hinder a determined statistician, it seems likely that many less well-trained and less experienced researchers would either avoid such methods, preferring much easier though often biased techniques like complete-case analysis, or they would use such methods but with a greater risk of making unintended and undetected mistakes. Other than demonstrating the advantage of including a statistician in the research team, this also highlights the benefit of easy to follow guides and additions to software.

Another feature that became apparent when conducting the case study was the initial difficulty faced when constructing a causal diagram. With no examples to follow that bore any relation to that which was to be created, considerable effort was required to overcome the many unknowns. These included which software to try, which variables to include, how to start drawing each variable and the initial shape of the diagram. This level of effort may help to explain why causal diagrams are still used by only a minority of researchers.

A resource that may be of some assistance would be an online searchable website containing examples of causal diagrams. It would hopefully expand over time and encourage the use of causal diagrams through learning by example and perhaps, by providing example DAGs (or other types of causal diagrams) with some similarity to a researcher's own study. By acting as a mental starting point for their own DAGs it would lower the effort required to give DAGs a try. However, avoiding the accidental promotion of DAGs that are missing important sources of confounding or selection bias is one potential problem. Another would be copyright concerns.

9.4 Adversarial collaborations

A research design that has appeared in psychology but is yet to be taken up by more than a handful of researchers has been called an 'adversarial collaboration'.[884,885] It is where a collaborative research project involves two opposing research groups with hypotheses that conflict. By conducting a combined research project that seeks to resolve the dispute, they are more likely to recognise the limitations of the claims they make.[884] An alternative design that is somewhat different yet might also be called an 'adversarial collaboration', does not target hypotheses in dispute but instead targets the cognitive biases of an opposing research group and making use of the fact that people recognise the cognitive biases, or the product of such biases, much more easily in other people than in themselves. The word adversarial seems apt, in the sense of a courtroom or even a sporting contest, though the idea still lacks details. It would involve competing groups, where group 1 would design the study, group 2 would perform it, group 1 would analyse it and so on, with some kind of third party umpire, and where it is in each group's interest to publicly criticise any shortcomings of the other group's work in some pre-arranged way.

An obvious downside of this design is that many researchers might not like the adversarial nature of it. On the other hand, if it could be made to work it would seem likely to produce better research, and the public and research funders alike would probably prefer this.

Nevertheless, for research to overcome some of the biases that currently seem to prevent progress in combating bias, ideas need to be proposed and possibly tested. In time, solutions will be found that have some success, and the biases that influence causal inference and our perception of uncertainty can be better controlled.

References

1. Chalmers I, Bracken MB, Djulbegovic B, et al. How to increase value and reduce waste when research priorities are set. *The Lancet*. 2014;383:156-165. doi:10.1016/S0140-6736(13)62229-1.
2. Ioannidis JPA, Greenland S, Hlatky Ma, et al. Increasing value and reducing waste in research design, conduct, and analysis. *The Lancet*. 2014;383(9912):166-175. doi:10.1016/S0140-6736(13)62227-8.
3. Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. *The Lancet*. 2009;374(9683):86-89. doi:10.1016/S0140-6736(09)60329-9.
4. Yordanov Y, Dechartres A, Porcher R, Boutron I, Altman DG, Ravaud P. Avoidable waste of research related to inadequate methods in clinical trials. *BMJ*. 2015;350(mar24 20):h809-h809. doi:10.1136/bmj.h809.
5. World Health Organisation. International Classification of Health Interventions (ICHI). http://www.who.int/classifications/ichi/en/. Accessed April 3, 2018.
6. Vandenbroucke JP. Those who were wrong. *American Journal of Epidemiology*. 1989;130(1):3-5. doi:10.1093/oxfordjournals.aje.a115320.
7. Wasserman L. Comment. *Journal of the American Statistical Association*. 1999;94(447):704-706. doi:10.1080/01621459.1999.10474171.
8. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Lippincott, Williams & Wilkins; 2008.
9. Newman G. The Bias Toward Cause and Effect. In: Stone GWM, Sarah J, eds. *Psychology of Bias*; 2012:69-82.
10. VanderWeele TJ. *Explanation in causal inference: Methods for mediation and interaction*. New York: Oxford University Press; 2015.
11. Hulswit M. *From cause to causation: A Peircean perspective*. Dordrecht: Springer Science; 2002.
12. Rothman KJ. Inferring Causal Connections - Habit, Faith or Logic? In: Rothman KJ, ed. *Causal Inference*. Chester Hill, MA: Epidemiology Resources Inc; 1988:3-12.

13. Susser M. Rational Science versus a System of Logic. In: Rothman KJ, ed. *Causal Inference*. Chester Hill, MA: Epidemiology Resources Inc; 1988:189–199.

14. Glass Ta, Goodman SN, Hernán MA, Samet JM. Causal Inference in Public Health. *Annual Review of Public Health*. 2013;34(1):61-75. doi:10.1146/annurev-publhealth-031811-124606.

15. Shadish WR, Cook TD, Campbell DT. *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin; 2002.

16. Holland PW. Statistics and Causal Inference. *Journal of the American Statistical Association*. 1986;81(396):945. doi:10.2307/2289064.

17. Oxford Dictionaries. Oxford English Dictionary. https://en.oxforddictionaries.com/. Accessed April 19, 2018.

18. Briggs W. *Uncertainty*. Cham: Springer International Publishing; 2016.

19. Matute H, Blanco F, Yarritu I, Diaz-Lago M, Vadillo MA, Barberia I. Illusions of causality: how they bias our everyday thinking and how they could be reduced. *Frontiers in Psychology*. 2015;6:888. doi:10.3389/fpsyg.2015.00888.

20. Greenland S. Induction versus Popper: Substance versus semantics. *International Journal of Epidemiology*. 1998;27:543-548. doi:10.1093/ije/27.4.543.

21. Horton R. Common sense and figures: the rhetoric of validity in medicine (Bradford Hill Memorial Lecture 1999). *Statistics in Medicine*. 2000;19(23):3149-3164. doi:10.1002/1097-0258(20001215)19:23<3149:AID-SIM617>3.0.CO;2-E.

22. Wikipedia. Abductive reasoning. https://en.wikipedia.org/wiki/Abductive_reasoning. Accessed April 19, 2018.

23. Senn SJ. Falsificationism and clinical trials. *Statistics in Medicine*. 1991;10(11):1679-1692. doi:10.1002/sim.4780101106.

24. Bunge M. *Doing science: In the light of philosophy*. New Jersey: World Scientific; 2017.

25. Pearce N. White Swans, Black Ravens, and Lame Ducks: Necessary and Sufficient Causes in Epidemiology. *Epidemiology*. 1990;1(1):47-50. doi:10.1097/00001648-199001000-00011.

26. Susser M. Falsification, Verification and Causal Inference in Epidemiology: Reconsideration in the Light of Sir Karl Popper's Philosophy. In: Rothman KJ, ed. *Causal Inference*. Chester Hill, MA: Epidemiology Resources Inc; 1988:33-57.

27. Munafò MR, Davey Smith G. Robust research needs many lines of evidence. *Nature*. 2018;553(7689):399-401. doi:10.1038/d41586-018-01023-3.

28. Koch E, Otarola A, Kirschbaum A. A landmark for popperian epidemiology: refutation of the randomised Aldactone evaluation study. *Journal of Epidemiology and Community Health*. 2005;59(11):1000-1006. doi:10.1136/jech.2004.031633.

29. Munafò MR, Davey Smith G. Philosophy of science isn't pointless chin-stroking – it makes us better scientists. *The Guardian*. Updated February 1, 2018.

30. Pearce N, Lawlor DA. Causal inference-so much more than statistics. *International Journal of Epidemiology*. 2016;45(6):1895-1903. doi:10.1093/ije/dyw328.

31. Broadbent A, Vandenbroucke JP, Pearce N. Response: Formalism or pluralism? A reply to commentaries on 'Causality and causal inference in epidemiology'. *International Journal of Epidemiology*. 2016;45(6):1841-1851. doi:10.1093/ije/dyw298.

32. Cox DR, Mayo DG. Objectivity and Conditionality in Frequentist Inference. In: Mayo DG, Spanos A, eds. *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability and the Objectivity and Rationality of Science*. Cambridge: Cambridge University Press; 2010:276-304.

33. Gelman A. Induction and deduction in Bayesian data analysis. *Rationality, Markets and Morals*. 2011;2:67-78.

34. Stone R. The Assumptions on which Causal Inference Rests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 1993;55(2):455-466.

35. Hill AB. The Environment and Disease: Association or Causation? *Proceedings of the Royal Society of Medicine*. 1965;58(2):295-300. doi:10.1016/j.tourman.2009.12.005.

36. Ioannidis JPA. Exposure-wide epidemiology: revisiting Bradford Hill. *Statistics in Medicine*. 2016;35(11):1749-1762. doi:10.1002/sim.6825.

37. Fedak KM, Bernal A, Capshaw ZA, Gross S. Applying the Bradford Hill criteria in the 21st century: how data integration has changed causal inference in molecular epidemiology. *Emerging Themes in Epidemiology*. 2015;12(1):14. doi:10.1186/s12982-015-0037-4.

38. Wakeford R. Association and causation in epidemiology - half a century since the publication of Bradford Hill's interpretational guidance. *Journal of the Royal Society of Medicine*. 2015;108(1):4-6. doi:10.1177/0141076814562713.

39. Rothman KJ. Causes. *American Journal of Epidemiology*. 1976;104(6):587-592.

40. Mackie JL. Causes and Conditions. *American Philosophical Quarterly*. 1965;2(4):245-264.

41. Cayley A. XXXVII. Note on a question in the theory of probabilities. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*. 1853;6(39):259. doi:10.1080/14786445308647360.

42. VanderWeele TJ. The Sufficient Cause Framework in Statistics, Philosophy and the Biomedical and Social Sciences. In: Berzuini C, Dawid AP, Bernardinelli L, eds. *Causality: Statistical Perspectives and Applications*. First. Chichester, UK: John Wiley & Sons, Ltd; 2012:180-191.

43. Greenland S, Brumback BA. An overview of relations among causal modelling methods. *International Journal of Epidemiology*. 2002;31(5):1030-1037. doi:10.1093/ije/31.5.1030.

44. Suzuki E, Tsuda T, Yamamoto E. Covariate balance for no confounding in the sufficient-cause model. *Annals of Epidemiology*. 2018;28(1):48-53.e2. doi:10.1016/j.annepidem.2017.11.005.

45. Olsen J. What characterises a useful concept of causation in epidemiology? *Journal of Epidemiology and Community Health*. 2003;57(2):86-88. doi:10.1136/jech.57.2.86.

46. VanderWeele TJ. Invited Commentary: The Continuing Need for the Sufficient Cause Model Today. *American Journal of Epidemiology*. 2017;185(11):1041-1043. doi:10.1093/aje/kwx083.

47. Daniel RM, Stavola BL de, Vansteelandt S. Commentary: The formal approach to quantitative causal inference in epidemiology: misguided or misrepresented? *International Journal of Epidemiology*. 2016;45(6):1817-1829. doi:10.1093/ije/dyw227.

48. Casella G, Schwartz SR. Comment. *Journal of the American Statistical Association*. 2000;95(450):425-427. doi:10.1080/01621459.2000.10474212.

49. Shadish WR. Campbell and Rubin: A primer and comparison of their approaches to causal inference in field settings. *Psychological Methods*. 2010;15(1):3-17. doi:10.1037/a0015916.

50. Hernán MA. A definition of causal effect for epidemiological research. *Journal of Epidemiology and Community Health*. 2004;58(4):265-271. doi:10.1136/jech.2002.006361.

51. van der Laan, Mark J., Rose S. *Targeted Learning*. New York, NY: Springer New York; 2011.

52. Speed TP. Introductory Remarks on Neyman (1923). *Statistical Science*. 1990;5(4):463-464.

53. Splawa-Neyman J, Dabrowska DM, Speed TP. On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. *Statistical Science*. 1990;5(4):465-472.

54. Rubin DB. Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies. *Statistical Science*. 1990;5(4):472-480.

55. Fisher RA. *Statistical Methods for Research Workers.* 1st. Edinburgh: Oliver and Boyd; 1925.

56. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*. 1974;66. doi:10.1037/h0037350.

57. Rubin DB. Assignment to Treatment Group on the Basis of a Covariate. *Journal of Educational Statistics*. 1977;2(1):1-26. doi:10.3102/10769986002001001.

58. Rubin DB. Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics*. 1978;6. doi:10.1214/aos/1176344064.

59. Hernán MA, Robins JM. *Causal Inference*. Boca Raton: Chapman & Hall/CRC, forthcoming; 2018.

60. Copas JB. Randomization models for the matched and unmatched 2 × 2 tables. *Biometrika*. 1973;60(3):467-476. doi:10.1093/biomet/60.3.467.

61. Maldonado G, Greenland S. Estimating causal effects. *International Journal of Epidemiology*. 2002;31(2):422-438. doi:10.1093/ije/31.2.422.

62. Rubin DB. Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. *Journal of the American Statistical Association*. 2005;100(469):322-331. doi:10.1198/016214504000001880.

63. Hernán MA. Beyond exchangeability: The other conditions for causal inference in medical research. *Statistical Methods in Medical Research*. 2012;21(1):3-5. doi:10.1177/0962280211398037.

64. Westreich DJ, Cole SR. Invited Commentary: Positivity in Practice. *American Journal of Epidemiology*. 2010;171(6):674-677. doi:10.1093/aje/kwp436.

65. Greenland S, Mansournia MA. Limitations of individual causal models, causal graphs, and ignorability assumptions, as illustrated by random confounding and design unfaithfulness. *European Journal of Epidemiology*. 2015;30(10):1101-1110. doi:10.1007/s10654-015-9995-7.

66. Cole SR, Frangakis CE. The consistency statement in causal inference: a definition or an assumption? *Epidemiology*. 2009;20(1):3-5. doi:10.1097/EDE.0b013e31818ef366.

67. Cole SR, Hudgens MG, Edwards JK. A Fundamental Equivalence between Randomized Experiments and Observational Studies. *Epidemiologic Methods*. 2016;5(1):113-117. doi:10.1515/em-2015-0029.

68. Hernán MA, Taubman SL. Does obesity shorten life? The importance of well-defined interventions to answer causal questions. *International Journal of Obesity*. 2008;32:S8-S14. doi:10.1038/ijo.2008.82.

69. Robins JM, Blevins D, Ritter G, Wulfsohn M. G-Estimation of the Effect of Prophylaxis Therapy for Pneumocystis carinii Pneumonia on the Survival of AIDS Patients. *Epidemiology*. 1992;3(4):319-336. doi:10.1097/00001648-199207000-00007.

70. Robins JM. Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics - Theory and Methods*. 1994;23(8):2379-2412. doi:10.1080/03610929408831393.

71. Rubin DB. Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment. *Journal of the American Statistical Association*. 1980;75(371):591. doi:10.2307/2287653.

72. VanderWeele TJ, Tchetgen Tchetgen EJ, Halloran ME. Interference and Sensitivity Analysis. *Statistical Science*. 2014;29(4):687-706. doi:10.1214/14-STS479.

73. Vandenbroucke JP, Broadbent A, Pearce N. Causality and causal inference in epidemiology: the need for a pluralistic approach. *International Journal of Epidemiology*. 2016;45(6):1776-1786. doi:10.1093/ije/dyv341.

74. Broadbent A, Vandenbroucke J, Pearce N. Authors' Reply to: VanderWeele et al., Chiolero, and Schooling et al. *International Journal of Epidemiology*. 2016;45(6):2203-2205. doi:10.1093/ije/dyw163.

75. Ebrahim S, Ferrie JE, Davey Smith G. The future of epidemiology: methods or matter? *International Journal of Epidemiology*. 2016;45(6):1699-1716. doi:10.1093/ije/dyx032.

76. Krieger N, Davey Smith G. Response: FACEing reality: productive tensions between our epidemiological questions, methods and mission. *International Journal of Epidemiology*. 2016;45(6):1852-1865. doi:10.1093/ije/dyw330.

77. Krieger N, 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):1787-1808. doi:10.1093/ije/dyw114.

78. Krieger N, Smith GD. Reply to Naimi. *International Journal of Epidemiology*. 2017;46(4):1342. doi:10.1093/ije/dyx087.

79. Schwartz S, Gatto NM, Campbell UB. Causal identification: a charge of epidemiology in danger of marginalization. *Annals of Epidemiology*. 2016;26(10):669-673. doi:10.1016/j.annepidem.2016.03.013.

80. Schwartz S, Prins SJ, Campbell UB, Gatto NM. Is the "well-defined intervention assumption" politically conservative? *Social Science & Medicine*. 2016;166:254-257. doi:10.1016/j.socscimed.2015.10.054.

81. Schwartz S, Gatto NM, Campbell UB. Heeding the call for less casual causal inferences: the utility of realized (quantitative) causal effects. *Annals of Epidemiology*. 2017;27(6):402-405. doi:10.1016/j.annepidem.2017.05.012.

82. Hernán MA. Does water kill? A call for less casual causal inferences. *Annals of Epidemiology*. 2016;26(10):674-680. doi:10.1016/j.annepidem.2016.08.016.

83. Kaufman JS. There is no virtue in vagueness. *Annals of Epidemiology*. 2016;26(10):683-684. doi:10.1016/j.annepidem.2016.08.018.

84. Naimi AI. On wagging tales about causal inference. *International Journal of Epidemiology*. 2017;46(4):1340-1342. doi:10.1093/ije/dyx086.

85. VanderWeele TJ. Commentary: On Causes, Causal Inference, and Potential Outcomes. *International Journal of Epidemiology*. 2016;45(6):1809-1816. doi:10.1093/ije/dyw230.

86. Greenland S. For and Against Methodologies: Some Perspectives on Recent Causal and Statistical Inference Debates. *European Journal of Epidemiology*. 2017;32(1):3-20. doi:10.1007/s10654-017-0230-6.

87. Robins JM, Weissman MB. Counterfactual causation and streetlamps: what is to be done? *International Journal of Epidemiology*. 2017;x(x):dyw231. doi:10.1093/ije/dyw231.

88. Lipton P. *Inference to the best explanation*. 2nd ed. London: Routledge; 2004.

89. Lawlor DA, Tilling K, Davey Smith G. Triangulation in aetiological epidemiology. *International Journal of Epidemiology*. 2016;45(6):1866-1886. doi:10.1093/ije/dyw314.

90. Glymour MM, Hamad R. Causal Thinking as a Critical Tool for Eliminating Social Inequalities in Health. *American Journal of Public Health*. 2018;108(5):623. doi:10.2105/AJPH.2018.304383.

91. Pearl J. *Causality: Models, reasoning, and inference*. 2nd ed. Cambridge: Cambridge University Press; 2009.

92. Kline RB. *Principles and practice of structural equation modeling*. 3rd ed. New York, London: Guilford; 2011.

93. Mansiaux Y, Salez N, Lapidus N, et al. Causal analysis of H1N1pdm09 influenza infection risk in a household cohort. *Journal of Epidemiology and Community Health*. 2015;69(3):272-277. doi:10.1136/jech-2014-204678.

94. Sobel ME. An Introduction to Causal Inference. *Sociological Methods & Research*. 1996;24(3):353-379. doi:10.1177/0049124196024003004.

95. Wilkinson L, & the Task Force on Statistical Inference. Statistical Methods in Psychology Journals: Guidelines and Explanations. *American Psychologist*. 1999;54(8):594-604.

96. VanderWeele TJ. Invited Commentary: Structural Equation Models and Epidemiologic Analysis. *American Journal of Epidemiology*. 2012;176(7):608-612. doi:10.1093/aje/kws213.

97. Bollen KA, Pearl J. Eight Myths About Causality and Structural Equation Models. In: Morgan SL, ed. *Handbook of causal analysis for social research*. Dordrecht: Springer; 2013:301-328. *Handbooks of sociolgy and social research*.

98. Dawid AP. Conditional Independence in Statistical Theory. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 1979;41(1):1-31.

99. Dawid AP. Statistical Causality from a Decision-Theoretic Perspective. *Annual Review of Statistics and Its Application*. 2015;2(1):273-303. doi:10.1146/annurev-statistics-010814-020105.

100. Dawid AP. Causal diagrams for empirical research: Discussion of 'Causal diagrams for empirical research' by J. Pearl. *Biometrika*. 1995;82(4):689-690. doi:10.1093/biomet/82.4.689.

101. Dawid AP. Causal Inference without Counterfactuals. *Journal of the American Statistical Association*. 2000;95(450):407-424. doi:10.1080/01621459.2000.10474210.

102. Greenland S, Robins JM, Pearl J. Confounding and Collapsibility in Causal Inference. *Statistical Science*. 1999;14(1):29-46. doi:10.1214/ss/1009211805.

103. Dawid AP. The Decision-Theoretic Approach to Causal Inference. In: *Causality: Statistical Perspectives and Applications*; 2012:25-42.

104. Robins JM, Greenland S. Comment. *Journal of the American Statistical Association*. 2000;95(450):431-435. doi:10.1080/01621459.2000.10474214.

105. Vandenbroucke JP. Clinical epidemiology: A daydream? *European Journal of Epidemiology*. 2017;32(2):95-101. doi:10.1007/s10654-017-0226-2.

106. West SG, Thoemmes FJ. Campbell's and Rubin's perspectives on causal inference. *Psychological Methods*. 2010;15(1):18-37. doi:10.1037/a0015917.

107. Smith HL. Research Design: Toward a Realistic Role for Causal Analysis. In: Morgan SL, ed. *Handbook of causal analysis for social research*. Dordrecht: Springer; 2013:45-73. *Handbooks of sociolgy and social research*.

108. Henrion M, Fischhoff B. Assessing Uncertainty in Physical Constants. In: Gilovich T, Griffin D, Kahneman D, eds. *Heuristics and biases: The psychology of intuitive judgement*. Cambridge: Cambridge University Press; 2002:666-677.

109. Lash TL. Heuristic Thinking and Inference From Observational Epidemiology. *Epidemiology*. 2007;18(1):67-72. doi:10.1097/01.ede.0000249522.75868.16.

110. Nester MR. An Applied Statistician's Creed. *Applied Statistics*. 1996;45(4):401. doi:10.2307/2986064.

111. Tukey JW. The Philosophy of Multiple Comparisons. *Statistical Science*. 1991;6(1):100-116. doi:10.1214/ss/1177011945.

112. Little RJ. Comment. *The American Statistician*. 2016;70(suppl).

113. McShane BB, Gal D. Blinding Us to the Obvious? The Effect of Statistical Training on the Evaluation of Evidence. *Management Science*. 2016;62(6):1707-1718. doi:10.1287/mnsc.2015.2212.

114. Gelman A, Carlin JB. Some Natural Solutions to the p -Value Communication Problem—and Why They Won't Work. *Journal of the American Statistical Association*. 2017;112(519):899-901. doi:10.1080/01621459.2017.1311263.

115. Anscombe FJ. The summarizing of clinical experiments by significance levels. *Statistics in Medicine*. 1990;9(6):703-708. doi:10.1002/sim.4780090617.

116. Morabia A. History of the modern epidemiological concept of confounding. *Journal of Epidemiology and Community Health*. 2011;65(4):297-300. doi:10.1136/jech.2010.112565.

117. Vineis P. History of bias. *Sozial- und Prventivmedizin/Social and Preventive Medicine*. 2002;47(3):156-161. doi:10.1007/BF01591887.

118. Porta MS, Greenland S, Hernán MA, dos Santos Silva I, Last JM. *A Dictionary of Epidemiology.* Six edition. Oxford: Oxford University Press; 2014.

119. Oxford Dictionaries. Definition of bias. https://en.oxforddictionaries.com/definition/bias. Accessed June 15, 2018.

120. Aronson J. A Word About Evidence 4. Bias - etymology and usage. https://catalogofbias.org/2018/04/10/a-word-about-evidence-4-bias-etymology-and-usag/. Accessed May 12, 2018.

121. Emerson G. Medical Statistics: being a Series of Tables, showing the Mortality in Philadelphia, and its immediate Causes, during a period of twenty years. *American Journal of the Medical Sciences*. 1827;1(1):116.

122. Jastrow J. Some Peculiarities in the Age Statistics of the United States. *Science*. 1885;5(122):461-464.

123. Bowley AL. *Elements of statistics*. London: P. S. King & Son; 1901.

124. Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*. 1900;50(302):157-175. doi:10.1080/14786440009463897.

125. Wishart J. Some Aspects of the Teaching of Statistics. *Journal of the Royal Statistical Society*. 1939;102(4):532-564.

126. Upton GJG, Cook I. *A dictionary of statistics.* 3rd ed. Oxford: Oxford University Press; 2014.

127. Devore JL, Berk KN. *Modern Mathematical Statistics with Applications*. New York, NY: Springer New York; 2012.

128. Greenland S, Pearce N. Statistical Foundations for Model-Based Adjustments. *Annual Review of Public Health*. 2015;36(1):89-108. doi:10.1146/annurev-publhealth-031914-122559.

129. Miller JB, Gelman A. Laplace's Theories of Cognitive Illusions, Heuristics, and Biases. *SSRN Electronic Journal*. 2018. doi:10.2139/ssrn.3149224.

130. Freedman D. From Association to Causation: Some Remarks on the History of Statistics. *Statistical Science*. 1999;14(3):243-258. doi:10.2307/2676760.

131. Pinker S. *The Sense of Style: The Thinking Person's Guide to Writing in the 21st Century*. New York, NY: Penguin; 2014.

132. Saeed JI. *Semantics.* Fourth edition. Chichester, West Sussex: Wiley Blackwell; 2016.

133. Mansournia MA, Hernán MA. The Authors Respond. *Epidemiology*. 2017;28(4):e41. doi:10.1097/EDE.0000000000000662.

134. Everitt BS, Skrondal A. *The Cambridge Dictionary of Statistics.* 4th ed. Cambridge, UK, New York: Cambridge University Press; 2010.

135. Starmans, Richard J. C. M. Models, Inference, and Truth: Probabilistic Reasoning in the Information Era. In: *Targeted Learning*. New York, NY: Springer New York; 2011:li-lxxi. *Springer Series in Statistics*. http://link.springer.com/10.1007/978-1-4419-9782-1.

136. Pearson K. *The Grammar of Science.* 2nd. London: Black; 1900.

137. Diggle PJ. Statistics: a data science for the 21st century. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 2015;178(4):793-813. doi:10.1111/rssa.12132.

138. Fisher RA. *The Design of Experiments*. Edinburgh: Oliver and Boyd; 1935.

139. Matthews JR. *Quantification and the Quest for Medical Certainty*. Princeton: Princeton University Press; 1995.

140. Shrier I. Structural Approach to Bias in Meta-analyses. *Research Synthesis Methods*. 2011;2(4):223-237. doi:10.1002/jrsm.52.

141. Ioannidis JPA. Randomized controlled trials: Often flawed, mostly useless, clearly indispensable: A commentary on Deaton and Cartwright. *Social Science & Medicine*. 2018;210:53-56. doi:10.1016/j.socscimed.2018.04.029.

142. Armitage P, Berry G, Matthews JNS. *Statistical Methods in Medical Research.* 4th ed. Oxford: Blackwell Science; 2002.

143. Gelman A, Hennig C. Beyond subjective and objective in statistics. *Journal of the Royal Statistical Society. Series A (General)*. 2017;180(4):967-1033. doi:10.1111/rssa.12276.

144. Hall P, Selinger B. Statistical Significance: Balancing Evidence Against Doubt. *Australian Journal of Statistics*. 1986;28(3):354-370. doi:10.1111/j.1467-842X.1986.tb00708.x.

145. Sterne JAC, Davey Smith G. Sifting the evidence - what's wrong with significance tests? *BMJ*. 2001;322(7280):226-231. doi:10.1136/bmj.322.7280.226.

146. Neyman J, Pearson ES. On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions. Series A: Mathematical, Physical, and Engineering Sciences*. 1933;231(694-706):289-337. doi:10.1098/rsta.1933.0009.

147. Goodman SN. p Values, Hypothesis Tests, and Likelihood: Implications for Epidemiology of a Neglected Historical Debate. *American Journal of Epidemiology*. 1993;137(5):485-496. doi:10.1093/oxfordjournals.aje.a116700.

148. Lehmann EL. *Fisher, Neyman, and the Creation of Classical Statistics*. New York NY: Springer; 2011.

149. Lash TL. The Harm Done to Reproducibility by the Culture of Null Hypothesis Significance Testing. *American Journal of Epidemiology*. 2017;186(6):627-635. doi:10.1093/aje/kwx261.

150. Mansournia MA, Higgins JPT, Sterne JAC, Hernán MA. Biases in Randomized Trials: A Conversation Between Trialists and Epidemiologists. *Epidemiology*. 2017;28(1):54-59. doi:10.1097/EDE.0000000000000564.

151. Chavalarias D, Ioannidis JPA. Science mapping analysis characterizes 235 biases in biomedical research. *Journal of Clinical Epidemiology*. 2010;63(11):1205-1215. doi:10.1016/j.jclinepi.2009.12.011.

152. Bussmann H, Trauth G, Kazzazi K. *Routledge dictionary of language and linguistics*. London: Routledge; 1996.

153. Hofstadter DR, Sander E. *Surfaces and essences: Analogy as the fuel and fire of thinking*. New York: Basic Books; 2013.

154. Bailey KD. *Typologies and taxonomies: An introduction to classification techniques*. Thousand Oaks, Calif., London: SAGE Publications; 1994.

155. Higgins JPT, Altman DG, Gøtzsche PC, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ*. 2011;343(oct18 2):d5928-d5928. doi:10.1136/bmj.d5928.

156. Higgins JPT, Sterne JAC, Savović J, et al. A revised tool for assessing risk of bias in randomized trials. In: Chandler J, McKenzie JE, Boutron I, Welch V, eds. *Cochrane Methods. Cochrane Database of Systematic Reviews 2016*. Issue 10 (Suppl 1); 2016.

157. Jarde A, Losilla J-M, Vives J, Rodrigo MF. Q-Coh: A tool to screen the methodological quality of cohort studies in systematic reviews and meta-analyses. *International Journal of Clinical and Health Psychology*. 2013;13(2):138-146. doi:10.1016/S1697-2600(13)70017-6.

158. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*. 2016;355:i4919. doi:10.1136/bmj.i4919.

159. Gawande A. *The checklist manifesto: How to get things right*. London: Profile Books; 2010.

160. Hales BM, Pronovost PJ. The checklist--a tool for error management and performance improvement. *Journal of Critical Care*. 2006;21(3):231-235. doi:10.1016/j.jcrc.2006.06.002.

161. Borgerson K. Valuing evidence: bias and the evidence hierarchy of evidence-based medicine. *Perspectives in Biology and Medicine*. 2009;52(2):218-233. doi:10.1353/pbm.0.0086.

162. Schooling CM, Cowling BJ. Type of Question Could Inform the Taxonomy of Bias. *Epidemiology*. 2015;26(4):e48. doi:10.1097/EDE.0000000000000308.

163. Sica GT. Bias in research studies. *Radiology*. 2006;238(3):780-789. doi:10.1148/radiol.2383041109.

164. Suzuki E, Mitsuhashi T, Tsuda T, Yamamoto E. A typology of four notions of confounding in epidemiology. *Journal of Epidemiology*. 2017;27(2):49-55. doi:10.1016/j.je.2016.09.003.

165. Weisberg HI. *Bias and causation: Models and judgment for valid comparisons*. Hoboken, New Jersey: John Wiley & Sons, Inc; 2011.

166. Kass PH. Converging toward a "Unified Field Theory" of Epidemiology. *Epidemiology*. 1992;3(6):473-475. doi:10.1097/00001648-199211000-00001.

167. Meinert CL. *Clinical trials dictionary: Terminology and usage recommendations / Curtis L. Meinert*. 2nd ed. Hoboken, N.J.: Wiley; 2012.

168. Wikipedia. Epidemiology. https://en.wikipedia.org/wiki/Epidemiology. Accessed July 13, 2018.

169. Nickerson RS. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*. 1998;2(2):175-220. doi:10.1037//1089-2680.2.2.175.

170. Matthews JR. History of biostatistics. *Medical Writing*. 2016;25(3):8-11.

171. Greenwood M. Is the statistical method of any value in medical research? *The Lancet*. 1924;204(5265):153-158. doi:10.1016/S0140-6736(01)35847-6.

172. Armitage P. Reflections on statistics at the London School of Hygiene and Tropical Medicine 30 years ago. *Statistics in Medicine*. 2000;19(23):3165-3170. doi:10.1002/1097-0258(20001215)19:23<3165:AID-SIM618>3.0.CO;2-L.

173. Streptomycin Treatment of Pulmonary Tuberculosis: A Medical Research Council Investigation. *BMJ*. 1948;2(4582):769-782. doi:10.1136/bmj.2.4582.769.

174. Bothwell LE, Greene JA, Podolsky SH, Jones DS, Malina D. Assessing the Gold Standard — Lessons from the History of RCTs. *New England Journal of Medicine*. 2016;374(22):2175-2181. doi:10.1056/NEJMms1604593.

175. Kaptchuk TJ. Intentional Ignorance: A History of Blind Assessment and Placebo Controls in Medicine. *Bulletin of the History of Medicine*. 1998;72(3):389-433. doi:10.1353/bhm.1998.0159.

176. Lindner MD. Clinical attrition due to biased preclinical assessments of potential efficacy. *Pharmacology & Therapeutics*. 2007;115(1):148-175. doi:10.1016/j.pharmthera.2007.05.002.

177. Egger M, Davey Smith G, O'Rourke K. Introduction: Rationale, Potentials, and Promise of Systematic Reviews. In: Egger M, Davey Smith G, Altman DG, eds. *Systematic Reviews in Health Care: Meta-Analysis in Context*. 2nd ed. London: BMJ Books; 2001:1-19.

178. Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: An annotated bibliography of scales and checklists. *Controlled Clinical Trials*. 1995;16(1):62-73. doi:10.1016/0197-2456(94)00031-W.

179. Jüni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA*. 1999;282(11):1054-1060. doi:10.1001/jama.282.11.1054.

180. Higgins JPT, Altman DG. Assessing Risk of Bias in Included Studies. In: *Cochrane Handbook for Systematic Reviews of Interventions*. Chichester, UK: John Wiley & Sons, Ltd; 2008:187-241.

181. Pearce N. Traditional epidemiology, modern epidemiology, and public health. *American Journal of Public Health*. 1996;86(5):678-683. doi:10.2105/AJPH.86.5.678.

182. Morabia A. Has Epidemiology Become Infatuated With Methods? A Historical Perspective on the Place of Methods During the Classical (1945–1965) Phase of

Epidemiology. *Annual Review of Public Health*. 2015;36(1):69-88. doi:10.1146/annurev-publhealth-031914-122403.

183. Morabia A. Epidemiology: An epistemological perspective. In: Morabia A, ed. *A History of Epidemiologic Methods and Concepts*. Basel: Birkhäuser Basel; 2004:3-125.
184. Kleinbaum DG, Morgenstern HAL, Kupper L. Selection Bias in Epidemiologic Studies. *American Journal of Epidemiology*. 1981;113(4):452-463. doi:10.1093/oxfordjournals.aje.a113113.
185. Delgado-Rodríguez M, Llorca J. Bias. *Journal of Epidemiology and Community Health*. 2004;58:635-641. doi:10.1136/jech.2003.008466.
186. Greenland S, Morgenstern H. Confounding in Health Research. *Annual Review of Public Health*. 2001;22(1):189-212. doi:10.1146/annurev.publhealth.22.1.189.
187. Greenland S. Confounding and Confounder Control. In: Lovric M, ed. *International Encyclopedia of Statistical Science*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2011.
188. Vandenbroucke JP. The history of confounding. *Sozial- und Prventivmedizin/Social and Preventive Medicine*. 2002;47(4):216-224. doi:10.1007/BF01326402.
189. Yule GU. Notes on the Theory of Association of Attributes in Statistics. *Biometrika*. 1903;2(2):121-134.
190. Cornfield J, Haenszel W, Hammond EC, Lilienfeld AM, Shimkin MB, Wynder EL. Smoking and Lung Cancer: Recent Evidence and a Discussion of Some Questions. *Journal of the National Cancer Institute*. 1959;22(1):173-203. doi:10.1093/jnci/22.1.173.
191. Miettinen OS. Matching and design efficiency in retrospective studies. *American Journal of Epidemiology*. 1970;91(2):111-118. doi:10.1093/oxfordjournals.aje.a121118.
192. Miettinen OS. *A Conversation with Olli Miettinen*. 2011. http://bcooltv.mcgill.ca/ListRecordings.aspx?CourseID=6485. Updated October 5, 2011. Accessed July 25, 2018.
193. Rothman KJ. A pictorial representation of confounding in epidemiologic studies. *Journal of Chronic Diseases*. 1975;28(2):101-108. doi:10.1016/0021-9681(75)90066-1.
194. Sackett DL. Bias in analytic research. *Journal of Chronic Diseases*. 1979;32(1-2):51-63. doi:10.1016/0021-9681(79)90012-2.

195. Greenland S, Neutra R. Control of Confounding in the Assessment of Medical Technology. *International Journal of Epidemiology*. 1980;9(4):361-367. doi:10.1093/ije/9.4.361.

196. Miettinen OS, Cook EF. Confounding: essence and detection. *American Journal of Epidemiology*. 1981;114(4):593-603.

197. Greenland S, Robins JM. Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology*. 1986;15(3):413-419. doi:10.1093/ije/15.3.413.

198. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *Journal of Epidemiology and Community Health*. 2006;60(7):578-586. doi:10.1136/jech.2004.029496.

199. Vandenbroucke JP, von Elm E, Altman DG, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and elaboration. *PLOS Medicine*. 2007;4:1628-1654. doi:10.1371/journal.pmed.0040297.

200. Berkson J. Limitations of the Application of Fourfold Table Analysis to Hospital Data. *Biometrics Bulletin*. 1946;2:47-53.

201. Snoep JD, Morabia A, Hernández-Díaz S, Hernán MA, Vandenbroucke JP. Commentary: A structural approach to Berkson's fallacy and a guide to a history of opinions about it. *International Journal of Epidemiology*. 2014;43(2):515-521. doi:10.1093/ije/dyu026.

202. Google Scholar. Search for Berkson's in the title of articles published since 2014. https://scholar.google.com.au/scholar?as_q=Berkson%27s&as_epq=&as_oq=&as_eq=&as_occt=title&as_sauthors=&as_publication=&as_ylo=2014&as_yhi=&hl=en&as_sdt=1%2C5. Accessed July 15, 2018.

203. Greenland S. Response and Follow-Up Bias in Cohort Studies. *American Journal of Epidemiology*. 1977;106(3):184-187.

204. Hernán MA. Invited Commentary: Selection Bias Without Colliders. *American Journal of Epidemiology*. 2017;185(11):1048-1050. doi:10.1093/aje/kwx077.

205. Bross IDJ. Misclassification in 2 X 2 Tables. *Biometrics*. 1954;10(4):478-486.

206. Murphy EA. *The Logic of Medicine*. Baltimore, Md: The Johns Hopkins University Press; 1976.

207. Feinstein AR. *Clinical Judgment*. Huntington: Krieger; 1967.

208. Catalogue of Bias Collaboration. About the Catalogue of Bias. https://catalogofbias.org/about/. Accessed May 26, 2018.

209. Vandenbroucke JP. Alvan Feinstein and the art of consulting. *Journal of Clinical Epidemiology*. 2002;55(12):1176-1177. doi:10.1016/S0895-4356(02)00523-1.

210. Feinstein AR. Epidemiologic analyses of causation: the unlearned scientific lessons of randomized trials. *Journal of Clinical Epidemiology*. 1989;42(6):481-489. doi:10.1016/0895-4356(89)90142-X.

211. Feinstein AR. Para-analysis, Faute de mieux, and the perils of riding on a data barge. *Journal of Clinical Epidemiology*. 1989;42(10):929-935. doi:10.1016/0895-4356(89)90157-1.

212. Choi BCK, Noseworthy aL. Classification, Direction, and Prevention of Bias in Epidemiologic Research. *Journal of Occupational and Environmental Medicine*. 1992;34(3):265-271. doi:10.1097/00043764-199203000-00010.

213. Steineck G, Ahlbom A. A definition of bias founded on the concept of the study base. *Epidemiology*. 1992;3(6):477-482. doi:10.1097/00001648-199211000-00003.

214. Miettinen OS. On Progress in Epidemiologic Academia. *European Journal of Epidemiology*. 2017;32(3):173-179. doi:10.1007/s10654-017-0227-1.

215. Maclure M, Schneeweiss S. Causation of Bias: The Episcope. *Epidemiology*. 2001;12(1):114-122. doi:10.1097/00001648-200101000-00019.

216. Howe CJ, Cain LE, Hogan JW. Are All Biases Missing Data Problems? *Current Epidemiology Reports*. 2015;2(3):162-171. doi:10.1007/s40471-015-0050-8.

217. Schwartz S, Campbell UB, Gatto NM, Gordon K. Toward a Clarification of the Taxonomy of "Bias" in Epidemiology Textbooks. *Epidemiology*. 2015;26(2):216-222. doi:10.1097/EDE.0000000000000224.

218. Rothman KJ. *Modern Epidemiology*. Boston: Little, Brown & Co; 1986.

219. Porta M, Fernandez E, Puigdomènech E. Book citations: influence of epidemiologic thought in the academic community. *Revista de Saúde Pública*. 2006;40(spe):50-56. doi:10.1590/S0034-89102006000400008.

220. Feinstein AR. Methodologic problems and standards in case-control research. *Journal of Chronic Diseases*. 1979;32(1-2):35-41. doi:10.1016/0021-9681(79)90009-2.

221. Feinstein AR. Scientific Standards in Epidemological Studies of the Menace of Daily Life. *Science*. 1988;242(4883):1257-1263. doi:10.1126/science.3057627.

222. Savitz DA, Greenland S, Stolley PD, Kelsey JL. Scientific Standards of Criticism : A Reaction to " Scientific Standards in Epidemiologic Studies of the Menace of Daily Life ," by. *Epidemiology*. 1990;1(5):78-83.

223. Morabia A. The controversial controversy of a passionate controversialist. *Journal of Clinical Epidemiology*. 2002;55(12):1207-1213. doi:10.1016/S0895-4356(02)00526-7.

224. Greenland S, Pearl J, Robins JM. Causal Diagrams for Epidemiologic Research. *Epidemiology*. 1999;10(1):37-48.

225. Elwert F. Graphical Causal Models. In: Morgan SL, ed. *Handbook of causal analysis for social research*. Dordrecht: Springer; 2013:245-273. *Handbooks of sociolgy and social research*.

226. Pearl J, Glymour M, Jewell NP. *Causal inference in statistics: A primer.* 1st ed. Hoboken, New Jersey: John Wiley & Sons; 2016.

227. Semin GR, Garrido MV, Farias AR. How Many Processes Does It Take to Ground a Concept? In: Sherman JW, Gawronski B, Trope Y, eds. *Dual-process theories of the social mind*. New York: The Guilford Press; 2014:542-559.

228. Kahneman D. *Thinking, Fast and Slow*. New York, NY: Straus & Giroux; 2011.

229. Hastie R. Causal Thinking in Judgments. In: Keren G, Wu G, eds. *The Wiley Blackwell Handbook of Judgment and Decision Making*. Vol. 54. Chichester, UK: John Wiley & Sons, Ltd; 2015:590-628.

230. Sloman SA, Lagnado D. Causality in Thought. *Annual Review of Psychology*. 2015;66(1):223-247. doi:10.1146/annurev-psych-010814-015135.

231. Greenland S, Pearl J. Causal Diagrams. In: Lovric M, ed. *International Encyclopedia of Statistical Science*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2011:208-216.

232. Pearl J. *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, Calif.: Morgan Kaufmann Publishers; 1988.

233. Wright S. Correlation and causation. *Journal of Agricultural Research*. 1921;20(7):557-585.

234. Greenland S. Overthrowing the tyranny of null hypothesis hidden in causal diagrams. In: Dechter R, Geffner H, Halpern JY, eds. *Heuristics, probability and causality: A tribute to Judea Pearl*. London: College Press; 2010:365-382.

235. Spirtes P, Glymour C, Scheines R. *Causation, Prediction, and Search*. New York: Springer Verlag; 1993.

236. Pearl J. Causal diagrams for empirical research. *Biometrika*. 1995;82(4):669-688. doi:10.1093/biomet/82.4.669.

237. Pearl J. *Causality: Models, reasoning, and inference.* 1st ed. Cambridge: Cambridge University Press; 2000.

238. Pearl J. The Deductive Approach to Causal Inference. *Journal of Causal Inference.* 2014;2(2):115-129. doi:10.1515/jci-2014-0016.

239. VanderWeele TJ, Robins JM. Four Types of Effect Modification: A Classification Based on Directed Acyclic Graphs. *Epidemiology*. 2007;18(5):561-568. doi:10.1097/EDE.0b013e318127181b.

240. Rubin DB. Direct and Indirect Causal Effects via Potential Outcomes. *Scandinavian Journal of Statistics*. 2004;31(June 2002):161-170.

241. Porta M, Vineis P, Bolúmar F. The current deconstruction of paradoxes: one sign of the ongoing methodological "revolution". *European Journal of Epidemiology*. 2015;30(10):1079-1087. doi:10.1007/s10654-015-0068-8.

242. Blakely T, Lynch J, Bentley R. Commentary: DAGs and the restricted potential outcomes approach are tools, not theories of causation. *International Journal of Epidemiology*. 2016;45(6):1835-1837. doi:10.1093/ije/dyw228.

243. Bandoli G, Palmsten K, Flores KF, Chambers CD. Constructing Causal Diagrams for Common Perinatal Outcomes: Benefits, Limitations and Motivating Examples with Maternal Antidepressant Use in Pregnancy. *Paediatric and Perinatal Epidemiology*. 2016;30(5):521-528. doi:10.1111/ppe.12302.

244. Daniel RM, Kenward MG, Cousens SN, Stavola BL de. Using causal diagrams to guide analysis in missing data problems. *Statistical Methods in Medical Research*. 2012;21(3):243-256. doi:10.1177/0962280210394469.

245. Howards PP, Schisterman EF, Poole C, Kaufman JS, Weinberg CR. "Toward a Clearer Definition of Confounding" Revisited With Directed Acyclic Graphs. *American Journal of Epidemiology*. 2012;176(6):506-511. doi:10.1093/aje/kws127.

246. Moffa G, Catone G, Kuipers J, et al. Using Directed Acyclic Graphs in Epidemiological Research in Psychosis: An Analysis of the Role of Bullying in Psychosis. *Schizophrenia Bulletin*. 2017;43(6):1273-1279. doi:10.1093/schbul/sbx013.

247. Röhrig N, Strobl R, Müller M, et al. Directed acyclic graphs helped to identify confounding in the association of disability and electrocardiographic findings: Results from the KORA-Age study. *Journal of Clinical Epidemiology*. 2014;67(2):199-206. doi:10.1016/j.jclinepi.2013.08.012.

248. Savitz DA, Wellenius GA. *Interpreting epidemiologic evidence: Connecting research to applications*. 2nd edition. Oxford, New York: Oxford University Press; 2016.

249. Snowden JM, Klebanoff MA. Applying causal diagrams in pediatrics to improve research, communication, and practice. *Pediatric research*. 2018;84(4):485-486. doi:10.1038/s41390-018-0109-6.

250. Staplin N, Herrington WG, Judge PK, et al. Use of Causal Diagrams to Inform the Design and Interpretation of Observational Studies: An Example from the Study of Heart and Renal Protection (SHARP). *Clinical Journal of the American Society of Nephrology*. 2017;12(3):546-552. doi:10.2215/CJN.02430316.

251. Suttorp MM, Siegerink B, Jager KJ, Zoccali C, Dekker FW. Graphical presentation of confounding in directed acyclic graphs. *Nephrology Dialysis Transplantation*. 2015;30(9):1418-1423. doi:10.1093/ndt/gfu325.

252. Williams TC, Bach CC, Matthiesen NB, Henriksen TB, Gagliardi L. Directed acyclic graphs: a tool for causal studies in paediatrics. *Pediatric research*. 2018;84(4):487-493. doi:10.1038/s41390-018-0071-3.

253. Williamson EJ, Aitken Z, Lawrie J, Dharmage SC, Burgess JA, Forbes AB. Introduction to causal diagrams for confounder selection. *Respirology*. 2014;19(3):303-311. doi:10.1111/resp.12238.

254. Piantadosi ST, Tily H, Gibson E. The communicative function of ambiguity in language. *Cognition*. 2012;122(3):280-291. doi:10.1016/j.cognition.2011.10.004.

255. Hernán MA, Hernández-Díaz S, Werler MM, Mitchell AA. Causal Knowledge as a Prerequisite for Confounding Evaluation: An Application to Birth Defects Epidemiology. *American Journal of Epidemiology*. 2002;155(2):176-184. doi:10.1093/aje/155.2.176.

256. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology*. 2004;15(5):615-625. doi:10.1097/01.ede.0000135174.63482.43.

257. Hernán MA, Cole SR. Invited Commentary: Causal Diagrams and Measurement Bias. *American Journal of Epidemiology*. 2009;170(8):959-962. doi:10.1093/aje/kwp293.

258. Vander Weele TJ. Confounding and effect modification: distribution and measure. *Epidemiologic Methods*. 2012;1(1):55-82. doi:10.1515/2161-962X.1004.

259. VanderWeele TJ. On the Distinction Between Interaction and Effect Modification. *Epidemiology (Cambridge, Mass.)*. 2009;20(6):863-871. doi:10.1097/EDE.0b013e3181ba333c.

260. Weinberg CR. Can DAGs Clarify Effect Modification? *Epidemiology (Cambridge, Mass.)*. 2007;18(5):569-572. doi:10.1097/EDE.0b013e318126c11d.

261. Keeble C, Thwaites PA, Baxter PD, Barber S, Parslow RC, Law GR. Learning Through Chain Event Graphs: The Role of Maternal Factors in Childhood Type 1 Diabetes. *American Journal of Epidemiology*. 2017;186(10):1204-1208. doi:10.1093/aje/kwx171.

262. Ackley SF, Mayeda ER, Worden L, Enanoria WTA, Glymour MM, Porco TC. Compartmental Model Diagrams as Causal Representations in Relation to DAGs. *Epidemiologic Methods*. 2017;6(1):1-23. doi:10.1515/em-2016-0007.

263. Rehfuess EA, Best N, Briggs DJ, Joffe M. Diagram-based Analysis of Causal Systems (DACS): elucidating inter-relationships between determinants of acute lower respiratory infections among children in sub-Saharan Africa. *Emerging Themes in Epidemiology*. 2013;10(1):13. doi:10.1186/1742-7622-10-13.

264. Singer RS, Williams-Nguyen J. Human health impacts of antibiotic use in agriculture: A push for improved causal inference. *Current Opinion in Microbiology*. 2014;19(1):1-8. doi:10.1016/j.mib.2014.05.014.

265. Glymour C, Scheines R. Causal modeling with the TETRAD program. *Synthese*. 1986;68(1):37-63.

266. Scheines R, Spirtes P, Glymour C, Meek C, Richardson TS. The TETRAD Project: Constraint Based Aids to Causal Model Specification. *Multivariate Behavioral Research*. 1998;33(1):65-117. doi:10.1207/s15327906mbr3301_3.

267. Textor J, Hardt J, Knüppel S. DAGitty: A Graphical Tool for Analyzing Causal Diagrams. *Epidemiology*. 2011;22(5):745. doi:10.1097/EDE.0b013e318225c2be.

268. Textor J, van der Zander B, Gilthorpe MS, Liskiewicz M, Ellison GT. Robust causal inference using directed acyclic graphs: the R package 'dagitty'. *International Journal of Epidemiology*. 2016;45(6):1887-1894. doi:10.1093/ije/dyw341.

269. Fekete J-D, van Wijk JJ, Stasko JT, North C. The Value of Information Visualization. In: Kerren A, ed. *Information visualization: Human-centered issues and perspectives / Andreas Kerren ... [et al.] (eds.)*. Berlin: Springer; 2008:1-18. *Lecture notes in computer science, State-of-the-art survey 0302-9743*; 4950.

270. Larkin JH, Simon HA. Why a Diagram is (Sometimes) Worth Ten Thousand Words. *Cognitive Science*. 1987;11(1):65-100. doi:10.1016/S0364-0213(87)80026-5.

271. Bauer MI, Johnson-Laird PN. How Diagrams Can Improve Reasoning. *Psychological Science*. 1993;4(6):372-378.

272. McCrudden MT, Schraw G, Lehman S, Poliquin A. The effect of causal diagrams on text learning. *Contemporary Educational Psychology*. 2007;32(3):367-388. doi:10.1016/j.cedpsych.2005.11.002.

273. Brewer LE, Wright JM, Rice G, Neas L, Teuschler L. Causal inference in cumulative risk assessment: The roles of directed acyclic graphs. *Environment international*. 2017;102:30-41. doi:10.1016/j.envint.2016.12.005.

274. Glymour MM. When Is Baseline Adjustment Useful in Analyses of Change? An Example with Education and Cognitive Change. *American Journal of Epidemiology*. 2005;162(3):267-278. doi:10.1093/aje/kwi187.

275. Ding P, Miratrix LW. To Adjust or Not to Adjust? Sensitivity Analysis of M-Bias and Butterfly-Bias. *Journal of Causal Inference*. 2015;3(1):41-57. doi:10.1515/jci-2013-0021.

276. Jansen JP, Schmid CH, Salanti G. Directed acyclic graphs can help understand bias in indirect and mixed treatment comparisons. *Journal of Clinical Epidemiology*. 2012;65(7):798-807. doi:10.1016/j.jclinepi.2012.01.002.

277. Pang M, Kaufman JS, Platt RW. Studying noncollapsibility of the odds ratio with marginal structural and logistic regression models. *Statistical Methods in Medical Research*. 2016;25(5):1925-1937. doi:10.1177/0962280213505804.

278. Daniel RM, Cousens SN, Stavola BL de, Kenward MG, Sterne JAC. Methods for dealing with time-dependent confounding. *Statistics in Medicine*. 2013;32(9):1584-1618. doi:10.1002/sim.5686.

279. Howe CJ, Cole SR, Lau B, Napravnik S, Eron JJ. Selection Bias Due to Loss to Follow Up in Cohort Studies. *Epidemiology*. 2016;27(1):91-97. doi:10.1097/EDE.0000000000000409.

280. Moreno-Betancur M, Lee KJ, Leacy FP, White IR, Simpson JA, Carlin JB. Canonical Causal Diagrams to Guide the Treatment of Missing Data in Epidemiologic Studies. *American Journal of Epidemiology*. 2018;187(12):2705-2715. doi:10.1093/aje/kwy173.

281. Cinelli C, Judea P. RE: A Practical Example Demonstrating the Utility of Single-world Intervention Graphs. *Epidemiology*. 2018;29(6):e50-e51. doi:10.1097/EDE.0000000000000896.

282. Hernán MA, Clayton D, Keiding N. The Simpson’s paradox unraveled. *International Journal of Epidemiology*. 2011;40(March):780-785. doi:10.1093/ije/dyr041.

283. Hernández-Díaz S, Schisterman EF, Hernán MA. The birth weight "paradox" uncovered? *American Journal of Epidemiology*. 2006;164(11):1115-1120. doi:10.1093/aje/kwj275.

284. Banack HR, Kaufman JS. Does selection bias explain the obesity paradox among individuals with cardiovascular disease? *Annals of Epidemiology*. 2015;25(5):342-349. doi:10.1016/j.annepidem.2015.02.008.

285. Vansteelandt S. Asking Too Much of Epidemiologic Studies: The Problem of Collider Bias and the Obesity Paradox. *Epidemiology*. 2017;28(5):e47-e49. doi:10.1097/EDE.0000000000000693.

286. Schooling CM, Au Yeung SL. "Selection bias by death" and other ways collider bias may cause the obesity paradox. *Epidemiology*. 2017;28(2):1-6. doi:10.1097/EDE.0000000000000591.

287. Robins JM. Data, Design, and Background Knowledge in Etiologic Inference. *Epidemiology*. 2001;12(3):313-320. doi:10.1097/00001648-200105000-00011.

288. Keogh RH, Daniel RM, VanderWeele TJ, Vansteelandt S. Analysis of Longitudinal Studies With Repeated Outcome Measures: Adjusting for Time-Dependent Confounding Using Conventional Methods. *American Journal of Epidemiology*. 2018;187(5):1085-1092. doi:10.1093/aje/kwx311.

289. Berkson J. Tests of Significance Considered as Evidence. *Journal of the American Statistical Association*. 1942;37(219):325-335.

290. Schor S, Karten I. Statistical Evaluation of Medical Journal Manuscripts. *JAMA*. 1966;195(13):1123. doi:10.1001/jama.1966.03100130097026.

291. Salsburg DS. The Religion of Statistics as Practiced in Medical Journals. *The American Statistician*. 1985;39(3):220. doi:10.2307/2683942.

292. Altman DG. The Scandal of Poor Medical Research. *BMJ*. 1994;308(6924):283-284. doi:10.1136/bmj.308.6924.283.

293. Ioannidis JPA. Contradicted and Initially Stronger Effects in Highly Cited Clinical Research. *JAMA*. 2005;294(2):218-228. doi:10.1001/jama.294.2.218.

294. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*. 2016;31(4):337-350. doi:10.1007/s10654-016-0149-3.

295. Box GEP. Some Problems of Statistics and Everyday Life. *Journal of the American Statistical Association*. 1979;74(365):1-4. doi:10.1080/01621459.1979.10481600.

296. Hosmer DW, Taber S, Lemeshow S. The importance of assessing the fit of logistic regression models: a case study. *American Journal of Public Health*. 1991;81(12):1630-1635. doi:10.2105/AJPH.81.12.1630.

297. Bagley SC, White H, Golomb BA. Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain. *Journal of Clinical Epidemiology*. 2001;54(10):979-985. doi:10.1016/S0895-4356(01)00372-9.

298. Ottenbacher KJ, Ottenbacher HR, Tooth L, Ostir GV. A review of two journals found that articles using multivariable logistic regression frequently did not report commonly recommended assumptions. *Journal of Clinical Epidemiology*. 2004;57(11):1147-1152. doi:10.1016/j.jclinepi.2003.05.003.

299. Steyerberg E. Stepwise Selection in Small Data Sets A Simulation Study of Bias in Logistic Regression Analysis. *Journal of Clinical Epidemiology*. 1999;52(10):935-942. doi:10.1016/S0895-4356(99)00103-1.

300. Windish DM, Huot SJ, Green ML. Medicine Residents' Understanding of the Biostatistics and Results in the Medical Literature. *JAMA*. 2007;298(9):1010. doi:10.1001/jama.298.9.1010.

301. Hershberger PJ, Part HM, Markert RJ, Cohen SM, Finger WW. Development of a test of cognitive bias in medical decision making. *Academic Medicine*. 1994;92(10):839-842.

302. Redelmeier DA. Medical Decision Making in Situations That Offer Multiple Alternatives. *JAMA*. 1995;273(4):302. doi:10.1001/jama.1995.03520280048038.

303. Klein JG. Five pitfalls in decisions about diagnosis and prescribing. *BMJ*. 2005;330(7494):781-783. doi:10.1136/bmj.330.7494.781.

304. Croskerry P. From Mindless to Mindful Practice — Cognitive Bias and Clinical Decision Making. *New England Journal of Medicine*. 2013;368(26):2445-2448. doi:10.1056/NEJMp1303712.

305. Saposnik G, Redelmeier DA, Ruff CC, Tobler PN. Cognitive biases associated with medical decisions: a systematic review. *BMC Medical Informatics and Decision Making*. 2016;16(1):138. doi:10.1186/s12911-016-0377-1.

306. Scott IA, Soon J, Elshaug AG, Lindner R. Countering cognitive biases in minimising low value care. *Medical Journal of Australia*. 2017;206(9):407-411. doi:10.5694/mja16.00999.

307. Reilly JB, Ogdie AR, Feldt JM von, Myers JS. Teaching about how doctors think: a longitudinal curriculum in cognitive bias and diagnostic error for residents. *BMJ Quality & Safety*. 2013;22(12):1044-1050. doi:10.1136/bmjqs-2013-001987.

308. Nuzzo R. How scientists fool themselves - and how they can stop. *Nature*. 2015;526(7572):182-185. doi:10.1038/526182a.

309. Rose M. Let's think about cognitive bias. *Nature*. 2015;526(7572):163. doi:10.1038/526163a.

310. Kaptchuk TJ. Effect of interpretive bias on research evidence. *BMJ*. 2003;326(7404):1453-1455. doi:10.1136/bmj.326.7404.1453.

311. Lomangino KM. Countering Cognitive Bias: Tips for Recognizing the Impact of Potential Bias on Research. *Journal of the Academy of Nutrition and Dietetics*. 2016;116(2):204-207. doi:10.1016/j.jand.2015.07.014.

312. Molony DA. Cognitive Bias and the Creation and Translation of Evidence Into Clinical Practice. *Advances in Chronic Kidney Disease*. 2016;23(6):346-350. doi:10.1053/j.ackd.2016.11.018.

313. Greenland S. The Need for Cognitive Science in Methodology. *American Journal of Epidemiology*. 2017;186(6):639-645. doi:10.1093/aje/kwx259.

314. Saini V, Garcia-Armesto S, Klemperer D, et al. Drivers of poor medical care. *The Lancet*. 2017;6736(16). doi:10.1016/S0140-6736(16)30947-3.

315. Mercier H, Heintz C. Scientists' Argumentative Reasoning. *Topoi*. 2014;33(2):513-524. doi:10.1007/s11245-013-9217-4.

316. Fanelli D. How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLOS ONE*. 2009;4(5):e5738. doi:10.1371/journal.pone.0005738.

317. Steen RG. Misinformation in the medical literature: What role do error and fraud play? *Journal of Medical Ethics*. 2011;37(8):498-503. doi:10.1136/jme.2010.041830.

318. Greenland S. Accounting for uncertainty about investigator bias: disclosure is informative. *Journal of Epidemiology and Community Health*. 2009;63(8):593-598. doi:10.1136/jech.2008.084913.

319. Greenland S. Transparency and disclosure, neutrality and balance: shared values or just shared words? *Journal of Epidemiology and Community Health*. 2012;66(11):967-970. doi:10.1136/jech-2011-200459.

320. MacCoun RJ. Biases In The Interpretation And Use Of Research Results. *Annual Review of Psychology*. 1998;49(1):259-287. doi:10.1146/annurev.psych.49.1.259.

321. Amrhein V, Korner-Nievergelt F, Roth T. The earth is flat (p > 0.05): significance thresholds and the crisis of unreplicable research. *PeerJ*. 2017;5:e3544. doi:10.7717/peerj.3544.

322. Michels KB, Rosner BA. Data trawling: to fish or not to fish. *The Lancet*. 1996;348(9035):1152-1153. doi:10.1016/S0140-6736(96)05418-9.

323. Bruns SB, Ioannidis JPA, Marinazzo D. p-Curve and p-Hacking in Observational Research. *PLOS ONE*. 2016;11(2):e0149144-e0149144. doi:10.1371/journal.pone.0149144.

324. Rothman KJ. Significance questing. *Annals of Internal Medicine*. 1986;105(3):445-447.

325. Dwan K, Altman DG, Arnaiz JA, et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLOS ONE*. 2008;3(8). doi:10.1371/journal.pone.0003081.

326. McCoy MS, Emanuel EJ. Why There Are No "Potential" Conflicts of Interest. *JAMA*. 2017;317(17):1721. doi:10.1001/jama.2017.2308.

327. Stark PB, Saltelli A. Cargo-cult statistics and scientific crisis. *Significance*. 2018;15(4):40-43. doi:10.1111/j.1740-9713.2018.01174.x.

328. Redelmeier DA, Shafir E. Why even good physicians do not wash their hands. *BMJ Quality & Safety*. 2015;24(12):744-747. doi:10.1136/bmjqs-2015-004319.

329. Chalmers TC, Matta RJ, Smith H, Kunzler AM. Evidence favoring the use of anticoagulants in the hospital phase of acute myocardial infarction. *New England Journal of Medicine*. 1977;297(20):1091-1096. doi:10.1056/NEJM197711172972004.

330. Sacks H, Chalmers TC, Smith H. Randomized versus historical controls for clinical trials. *The American Journal of Medicine*. 1982;72(2):233-240. doi:10.1016/0002-9343(82)90815-4.

331. Chalmers TC, Celano P, Sacks HS, Smith H. Bias in treatment assignment in controlled clinical trials. *New England Journal of Medicine*. 1983;309(22):1358-1361. doi:10.1056/NEJM198312013092204.

332. Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *BMJ*. 1998;317(7167):1185-1190.

333. Ioannidis JPA. Comparison of Evidence of Treatment Effects in Randomized and Nonrandomized Studies. *JAMA*. 2001;286(7):821. doi:10.1001/jama.286.7.821.

334. Tzoulaki I, Siontis KCM, Ioannidis JPA. Prognostic effect size of cardiovascular biomarkers in datasets from observational studies versus randomised trials: meta-epidemiology study. *BMJ*. 2011;343:d6829.

335. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *New England Journal of Medicine*. 2000;342:1878-1886.

336. Concato J, Shah N, Horwitz RI. Randomized, Controlled Trials, Observational Studies, and the Hierarchy of Research Designs. *New England Journal of Medicine*. 2000;342(25):1887-1892. doi:10.1056/NEJM200006223422507.

337. Dahabreh IJ, Sheldrick RC, Paulus JK, et al. Do observational studies using propensity score methods agree with randomized trials? A systematic comparison of studies on acute coronary syndromes. *European Heart Journal*. 2012;33(15):1893-1901. doi:10.1093/eurheartj/ehs114.

338. Zhang Z, Ni H, Xu X. Do the observational studies using propensity score analysis agree with randomized controlled trials in the area of sepsis? *Journal of Critical Care*. 2014;29(5):886.e9-886.e15. doi:10.1016/j.jcrc.2014.05.023.

339. Lonjon G, Boutron I, Trinquart L, et al. Comparison of Treatment Effect Estimates From Prospective Nonrandomized Studies With Propensity Score Analysis and Randomized Controlled Trials of Surgical Procedures. *Annals of Surgery*. 2014;259(1):18-25. doi:10.1097/SLA.0000000000000256.

340. Feinstein AR, Horwitz RI. Double standards, scientific methods, and epidemiologic research. *New England Journal of Medicine*. 1982;307(26):1611-1617. doi:10.1056/NEJM198212233072604.

341. Jones DS, Podolsky SH. The history and fate of the gold standard. *The Lancet*. 2015;385(9977):1502-1503. doi:10.1016/S0140-6736(15)60742-5.

342. Greenland S. Randomization, statistics, and causal inference. *Epidemiology*. 1990;1(6):421-429.

343. Senn SJ. *Statistical issues in drug development*. 2nd ed. Chichester, England: John Wiley et Sons; 2007.

344. Hróbjartsson A, Boutron I. Blinding in Randomized Clinical Trials: Imposed Impartiality. *Clinical Pharmacology & Therapeutics*. 2011;90(5):732-736. doi:10.1038/clpt.2011.207.

345. Duggan PF. Time to abolish "gold standard". *BMJ*. 1992;304(6841):1568-1569. doi:10.1136/bmj.304.6841.1568-b.

346. Claassen JAHR. The gold standard: not a golden standard. *BMJ*. 2005;330(7500):1121. doi:10.1136/bmj.330.7500.1121.

347. Concato J, Horwitz RI. Randomized trials and evidence in medicine: A commentary on Deaton and Cartwright. *Social Science & Medicine*. 2018;210:32-36. doi:10.1016/j.socscimed.2018.04.010.

348. Deaton A, Cartwright N. Reflections on Randomized Control Trials. *Social Science & Medicine*. 2018;210:86-90. doi:10.1016/j.socscimed.2018.04.046.

349. Gelman A. Benefits and limitations of randomized controlled trials: A commentary on Deaton and Cartwright. *Social Science & Medicine*. 2018;210:48-49. doi:10.1016/j.socscimed.2018.04.034.

350. Horwitz RI, Singer B. Introduction. What works? And for whom? *Social Science & Medicine*. 2018;210:22-25. doi:10.1016/j.socscimed.2018.05.013.

351. Jones A, Steel D. A combined theoretical and empirical approach to evidence quality evaluation: A commentary on Deaton and Cartwright. *Social Science & Medicine*. 2018;210:74-76. doi:10.1016/j.socscimed.2018.04.035.

352. Sampson RJ. After the experimental turn: A commentary on Deaton and Cartwright. *Social Science & Medicine*. 2018;210:67-69. doi:10.1016/j.socscimed.2018.04.013.

353. Stampfer MJ, Colditz GA. Estrogen replacement therapy and coronary heart disease: A quantitative assessment of the epidemiologic evidence. *Preventive Medicine*. 1991;20(1):47-63. doi:10.1016/0091-7435(91)90006-P.

354. Grady D, Rubin SM, Petitti DB, et al. Hormone Therapy To Prevent Disease and Prolong Life in Postmenopausal Women. *Annals of Internal Medicine*. 1992;117(12):1016. doi:10.7326/0003-4819-117-12-1016.

355. Hulley S. Randomized Trial of Estrogen Plus Progestin for Secondary Prevention of Coronary Heart Disease in Postmenopausal Women. *JAMA*. 1998;280(7):605. doi:10.1001/jama.280.7.605.

356. Humphrey LL, Chan BKS, Sox HC. Postmenopausal Hormone Replacement Therapy and the Primary Prevention of Cardiovascular Disease. *Annals of Internal Medicine*. 2002;137(4):273. doi:10.7326/0003-4819-137-4-200208200-00012.

357. Viscoli CM, Brass LM, Kernan WN, Sarrel PM, Suissa S, Horwitz RI. A clinical trial of estrogen-replacement therapy after ischemic stroke. *New England Journal of Medicine*. 2001;345(17):1243-1249. doi:10.1056/NEJMoa010534.

358. Writing Group for the Women's Health Initiative. Risks and Benefits of Estrogen Plus Progestin in Healthy Postmenopausal Women: Principal Results From the Women's Health Initiative Randomized Controlled Trial. *JAMA*. 2002;288(3):321-333. doi:10.1001/jama.288.3.321.

359. Laine C. Postmenopausal Hormone Replacement Therapy: How Could We Have Been So Wrong? *Annals of Internal Medicine*. 2002;137(4):290. doi:10.7326/0003-4819-137-4-200208200-00015.

360. Petitti D. Commentary: hormone replacement therapy and coronary heart disease: four lessons. *International Journal of Epidemiology*. 2004;33(3):461-463. doi:10.1093/ije/dyh192.

361. Prentice RL, Pettinger M, Anderson GL. Statistical issues arising in the Women's Health Initiative. *Biometrics*. 2005;61(4):899-911; discussion 911-41. doi:10.1111/j.0006-341X.2005.454_1.x.

362. Hernán MA, Alonso A, Logan R, et al. Observational Studies Analyzed Like Randomized Experiments. *Epidemiology*. 2008;19(6):766-779. doi:10.1097/EDE.0b013e3181875e61.

363. Vandenbroucke JP. The HRT controversy: observational studies and RCTs fall in line. *The Lancet*. 2009;373(9671):1233-1235. doi:10.1016/S0140-6736(09)60708-X.

364. von Elm E, Egger M. The scandal of poor epidemiological research. *BMJ*. 2004;329(7471):868-869. doi:10.1136/bmj.329.7471.868.

365. Taubes G. Do we really know what makes us healthy? *New York Times magazine*. Updated September 16, 2007:52 et seq.

366. Shen L, Ji H-F. Is antioxidant supplement beneficial? New avenue to explore. *Trends in Food Science & Technology*. 2017;68:51-55. doi:10.1016/j.tifs.2017.08.010.

367. Bjelakovic G, Nikolova D, Gluud LL, Simonetti RG, Gluud C. Mortality in randomized trials of antioxidant supplements for primary and secondary prevention: systematic review and meta-analysis. *JAMA*. 2007;297(8):842-857. doi:10.1001/jama.297.8.842.

368. Enstrom JE, Kanim LE, Klein MA. Vitamin C Intake and Mortality among a Sample of the United States Population. *Epidemiology*. 1992;3(3):194-202. doi:10.1097/00001648-199205000-00003.

369. Rimm EB, Stampfer MJ, Ascherio A, Giovannucci E, Colditz GA, Willett WC. Vitamin E consumption and the risk of coronary heart disease in men. *New England Journal of Medicine*. 1993;328(20):1450-1456. doi:10.1056/NEJM199305203282004.

370. Myung S-K, Ju W, Cho B, et al. Efficacy of vitamin and antioxidant supplements in prevention of cardiovascular disease: systematic review and meta-analysis of randomised controlled trials. *BMJ*. 2013;346:f10.

371. Bjelakovic G, Nikolova D, Simonetti RG, Gluud C. Antioxidant supplements for prevention of gastrointestinal cancers: a systematic review and meta-analysis. *The Lancet*. 2004;364(9441):1219-1228. doi:10.1016/S0140-6736(04)17138-9.

372. Miller ER, Pastor-Barriuso R, Dalal D, Riemersma RA, Appel LJ, Guallar EL. Meta-Analysis: High-Dosage Vitamin E Supplementation May Increase All-Cause Mortality. *ACC Current Journal Review*. 2005;14(5):17. doi:10.1016/j.accreview.2005.04.017.

373. Davey Smith G, Ebrahim S. Epidemiology—is it time to call it a day? *International Journal of Epidemiology*. 2001;30(1):1-11. doi:10.1093/ije/30.1.1.

374. Forman D, Altman D. Vitamins to prevent cancer: supplementary problems. *The Lancet*. 2004;364(9441):1193-1194. doi:10.1016/S0140-6736(04)17153-5.

375. Lawlor DA, Smith GD, Bruckdorfer KR, Kundu D, Ebrahim S. Those confounded vitamins: what can we learn from the differences between observational versus randomised trial evidence? *The Lancet*. 2004;363(9422):1724-1727. doi:10.1016/S0140-6736(04)16260-0.

376. Chou R, Dana T, Blazina I, Daeges M, Jeanne TL. Statins for Prevention of Cardiovascular Disease in Adults: Evidence Report and Systematic Review for the US Preventive Services Task Force. *JAMA*. 2016;316(19):2008-2024. doi:10.1001/jama.2015.15629.

377. Gupta A, Thompson D, Whitehouse A, et al. Adverse events associated with unblinded, but not with blinded, statin therapy in the Anglo-Scandinavian Cardiac Outcomes Trial—Lipid-Lowering Arm (ASCOT-LLA): a randomised double-blind placebo-controlled trial and its non-randomised non-blind extension phase. *The Lancet*. 2017;389(10088):2473-2481. doi:10.1016/S0140-6736(17)31075-9.

378. Goldfine AB. Statins: is it really time to reassess benefits and risks? *New England Journal of Medicine*. 2012;366(19):1752-1755. doi:10.1056/NEJMp1203020.

379. Collins R, Reith C, Emberson J, et al. Interpretation of the evidence for the efficacy and safety of statin therapy. *The Lancet*. 2016;6736(16). doi:10.1016/S0140-6736(16)31357-5.

380. Pedro-Botet J, Rubiés-Prat J. Statin-associated muscle symptoms: beware of the nocebo effect. *The Lancet*. 2017;389(10088):2445-2446. doi:10.1016/S0140-6736(17)31163-7.

381. Bonovas S, Filioussi K, Sitaras NM. Statin use and the risk of prostate cancer: A metaanalysis of 6 randomized clinical trials and 13 observational studies. *International Journal of Cancer*. 2008;123(4):899-904. doi:10.1002/ijc.23550.

382. Emilsson L, García-Albéniz X, Logan RW, Caniglia EC, Kalager M, Hernán MA. Examining Bias in Studies of Statin Treatment and Survival in Patients With Cancer. *JAMA Oncology*. 2017;02115:1-8. doi:10.1001/jamaoncol.2017.2752.

383. Wasserstein RL, Lazar NA. The ASA's Statement on p -Values: Context, Process, and Purpose. *The American Statistician*. 2016;70(2):129-133. doi:10.1080/00031305.2016.1154108.

384. Peng R. The reproducibility crisis in science: A statistical counterattack. *Significance*. 2015;12(3):30-32. doi:10.1111/j.1740-9713.2015.00827.x.

385. Goodman SN, Fanelli D, Ioannidis JPA. What does research reproducibility mean? *Science Translational Medicine*. 2016;8(341):341ps12-341ps12. doi:10.1126/scitranslmed.aaf5027.

386. Lecoutre M-P, Poitevineau J, Lecoutre B. Even statisticians are not immune to misinterpretations of Null Hypothesis Significance Tests. *International Journal of Psychology*. 2003;38(1):37-45. doi:10.1080/00207590244000250.

387. Weinberg CR. Invited Commentary: Can Issues With Reproducibility in Science Be Blamed on Hypothesis Testing? *American Journal of Epidemiology*. 2017;186(6):636-638. doi:10.1093/aje/kwx258.

388. Sterling TD. Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance - Or Vice Versa. *Journal of the American Statistical Association*. 1959;54(285):30-34.

389. Rozeboom WW. The Fallacy of the Null-Hypothesis Significance Test. *Psychological Bulletin*. 1960;57(5):416-428. doi:10.1037/h0042040.

390. Altman DG. Statistics in Medical Journals. *Statistics in Medicine*. 1982;1(1):59-71. doi:10.1002/sim.4780010109.

391. Vandenbroucke JP. How Trustworthy is Epidemiologic Research? *Epidemiology*. 1990;1(1):83-84. doi:10.1097/00001648-199001000-00018.

392. Ioannidis JPA. Why Most Published Research Findings Are False. *PLOS Medicine*. 2005;2(8):e124. doi:10.1371/journal.pmed.0020124.

393. Smith R. Medical Research - Still a Scandal. https://blogs-bmj-com/bmj/2014/01/31/richard-smith-medical-research-still-a-scandal/. Accessed June 13, 2018.

394. Shojania KG, Sampson M, Ansari MT, Ji J, Doucette S, Moher D. How Quickly Do Systematic Reviews Go Out of Date? A Survival Analysis. *Annals of Internal Medicine*. 2007;147(4):224. doi:10.7326/0003-4819-147-4-200708210-00179.

395. Prasad V, Gall V, Cifu A. The Frequency of Medical Reversal. *Archives of Internal Medicine*. 2011;171(18):1675-1676. doi:10.1001/archinternmed.2011.295.

396. de Vries F, Zeegers M, Goossens ME. Pioglitazone and bladder cancer: two studies, same database, two answers. *British Journal of Clinical Pharmacology*. 2013;76(3):484-485. doi:10.1111/bcp.12145.

397. Prasad V, Vandross A, Toomey C, et al. A decade of reversal: an analysis of 146 contradicted medical practices. *Mayo Clinic Proceedings*. 2013;88(8):790-798. doi:10.1016/j.mayocp.2013.05.012.

398. Eckel RH. Eggs and beyond: Is dietary cholesterol no longer important? *American Journal of Clinical Nutrition*. 2015;102(2):235-236. doi:10.3945/ajcn.115.116905.

399. Alamri A, Stevenson MR. A corpus of potentially contradictory research claims from cardiovascular research abstracts. *Journal of Biomedical Semantics*. 2016;7(1):36. doi:10.1186/s13326-016-0083-z.

400. Savović J, Turner RM, Mawdsley D, et al. Association Between Risk-of-Bias Assessments and Results of Randomized Trials in Cochrane Reviews: The ROBES Meta-Epidemiologic Study. *American Journal of Epidemiology*. 2018;187(5):1113-1122. doi:10.1093/aje/kwx344.

401. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical Evidence of Bias: Dimensions of Methodological Quality Associated With Estimates of Treatment Effects in Controlled Trials. *JAMA*. 1995;273(5):408. doi:10.1001/jama.1995.03520290060030.

402. Chan A-W, Altman DG. Identifying outcome reporting bias in randomised trials on PubMed: review of publications and survey of authors. *BMJ*. 2005;330(7494):753-0. doi:10.1136/bmj.38356.424606.8F.

403. Wood L, Egger M, Gluud LL, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. *BMJ*. 2008;336(7644):601-605. doi:10.1136/bmj.39465.451748.AD.

404. Hróbjartsson A, Thomsen ASS, Emanuelsson F, et al. Observer bias in randomised clinical trials with binary outcomes: systematic review of trials with both blinded and non-blinded outcome assessors. *BMJ*. 2012;344(feb27 2):e1119-e1119. doi:10.1136/bmj.e1119.

405. Jordan S, Watkins A, Storey M, et al. Volunteer Bias in Recruitment, Retention, and Blood Sample Donation in a Randomised Controlled Trial Involving Mothers and Their Children at Six Months and Two Years: A Longitudinal Analysis. *PLOS ONE*. 2013;8(7):e67912. doi:10.1371/journal.pone.0067912.

406. Hróbjartsson A, Emanuelsson F, Skou Thomsen AS, Hilden J, Brorson S. Bias due to lack of patient blinding in clinical trials. A systematic review of trials randomizing patients to blind and nonblind sub-studies. *International Journal of Epidemiology*. 2014;43(4):1272-1283. doi:10.1093/ije/dyu115.

407. Moroz V, Wilson JS, Kearns P, Wheatley K. Comparison of anticipated and actual control group outcomes in randomised trials in paediatric oncology provides evidence that historically controlled studies are biased in favour of the novel treatment. *Trials*. 2014;15(1):481. doi:10.1186/1745-6215-15-481.

408. Gewandter JS, McKeown A, McDermott MP, et al. Data Interpretation in Analgesic Clinical Trials With Statistically Nonsignificant Primary Analyses: An ACTTION Systematic Review. *Journal of Pain*. 2015;16(1):3-10. doi:10.1016/j.jpain.2014.10.003.

409. Page MJ, Higgins JPT, Clayton G, Sterne JAC, Hróbjartsson A, Savović J. Empirical Evidence of Study Design Biases in Randomized Trials: Systematic Review of Meta-Epidemiological Studies. *PLOS ONE*. 2016;11(7):e0159267. doi:10.1371/journal.pone.0159267.

410. Allemann SS, Nieuwlaat R, Navarro T, Haynes B, Hersberger KE, Arnet I. Congruence between patient characteristics and interventions may partly explain medication adherence intervention effectiveness: an analysis of 190 randomized controlled trials from a Cochrane systematic review. *Journal of Clinical Epidemiology*. 2017;91:70-79. doi:10.1016/j.jclinepi.2017.07.011.

411. Mayo-Wilson E, Li T, Fusco N, et al. Cherry-picking by trialists and meta-analysts can drive conclusions about intervention efficacy. *Journal of Clinical Epidemiology*. 2017;91:95-110. doi:10.1016/j.jclinepi.2017.07.014.

412. Nguyen T-L, Collins GS, Lamy A, et al. Simple randomization did not protect against bias in smaller trials. *Journal of Clinical Epidemiology*. 2017;84:105-113. doi:10.1016/j.jclinepi.2017.02.010.

413. Bolzern J, Mnyama N, Bosanquet K, Torgerson DJ. A review of cluster randomized trials found statistical evidence of selection bias. *Journal of Clinical Epidemiology*. 2018;99:106-112. doi:10.1016/j.jclinepi.2018.03.010.

414. Carlisle JB. Data fabrication and other reasons for non-random sampling in 5087 randomised, controlled trials in anaesthetic and general medical journals. *Anaesthesia*. 2017;72(8):944-952. doi:10.1111/anae.13938.

415. Loadsman JA, McCulloch TJ. Widening the search for suspect data - is the flood of retractions about to become a tsunami? *Anaesthesia*. 2017;72(8):931-935. doi:10.1111/anae.13962.

416. Steen RG, Casadevall A, Fang FC, Derrick GE. Why Has the Number of Scientific Retractions Increased? *PLOS ONE*. 2013;8(7):e68397. doi:10.1371/journal.pone.0068397.

417. Sawin VI, Robinson KA. Biased and inadequate citation of prior research in reports of cardiovascular trials is a continuing source of waste in research. *Journal of Clinical Epidemiology*. 2015;69:174-178. doi:10.1016/j.jclinepi.2015.03.026.

418. Albarqouni LN, López-López JA, Higgins JPT. Indirect evidence of reporting biases was found in a survey of medical research studies. *Journal of Clinical Epidemiology*. 2017;83:57-64. doi:10.1016/j.jclinepi.2016.11.013.

419. Carmona-Bayonas A, Jimenez-Fonseca P, Fernández-Somoano A, et al. Top ten errors of statistical analysis in observational studies for cancer research. *Clinical and Translational Oncology*. 2017;32(15):817. doi:10.1007/s12094-017-1817-9.

420. Perneger TV, Combescure C. The distribution of P -values in medical research articles suggested selective reporting associated with statistical significance. *Journal of Clinical Epidemiology*. 2017;87(0):70-77. doi:10.1016/j.jclinepi.2017.04.003.

421. Wolkewitz M, Schumacher M. Survival biases lead to flawed conclusions in observational treatment studies of influenza patients. *Journal of Clinical Epidemiology*. 2017;84:121-129. doi:10.1016/j.jclinepi.2017.01.008.

422. Di Girolamo N, Winter A, Meursinge Reynders R. High and unclear risk of bias assessments are predominant in diagnostic accuracy studies included in Cochrane reviews. *Journal of Clinical Epidemiology*. 2018;101:73-78. doi:10.1016/j.jclinepi.2018.05.001.

423. Hemkens LG, Ewald H, Naudet F, et al. Interpretation of epidemiologic studies very often lacked adequate consideration of confounding. *Journal of Clinical Epidemiology*. 2018;93:94-102. doi:10.1016/j.jclinepi.2017.09.013.

424. Lacny S, Wilson T, Clement F, et al. Kaplan-Meier survival analysis overestimates cumulative incidence of health-related events in competing risk settings: a meta-

analysis. *Journal of Clinical Epidemiology*. 2018;93:25-35. doi:10.1016/j.jclinepi.2017.10.006.

425. Losilla J-M, Oliveras I, Marin-Garcia JA, Vives J. Three risk of bias tools lead to opposite conclusions in observational research synthesis. *Journal of Clinical Epidemiology*. 2018;101:61-72. doi:10.1016/j.jclinepi.2018.05.021.

426. Vickers A. Interpreting data from randomized trials: the Scandinavian prostatectomy study illustrates two common errors. *Nature Clinical Practice Urology*. 2005;2(9):404-405. doi:10.1038/ncpuro0294.

427. NHMRC. Outcomes of funding rounds. https://www.nhmrc.gov.au/grants-funding/outcomes-of-funding-rounds. Accessed September 10, 2018.

428. Horrobin DF. The Philosophical Basis of Peer Review and the Suppression of Innovation. *JAMA*. 1990;263(10):1438. doi:10.1001/jama.1990.03440100162024.

429. Erren TC, Shaw DM, Groß JV. How to avoid haste and waste in occupational, environmental and public health research. *Journal of Epidemiology and Community Health*. 2015;69(9):823-825. doi:10.1136/jech-2015-205543.

430. Davidian M, Louis TA. Why Statistics? *Science*. 2012;336(April):12. doi:10.1126/science.1218685.

431. Bruce R, Chauvin A, Trinquart L, Ravaud P, Boutron I. Impact of interventions to improve the quality of peer review of biomedical journals: a systematic review and meta-analysis. *BMC Medicine*. 2016;14(1):85. doi:10.1186/s12916-016-0631-5.

432. Morey RD, Chambers CD, Etchells PJ, et al. The Peer Reviewers' Openness Initiative: incentivizing open research practices through peer review. *Royal Society Open Science*. 2016;3(1):150547. doi:10.1098/rsos.150547.

433. Mathieu S, Boutron I, Moher D, Altman DG, Ravaud P. Comparison of Registered and Published Primary Outcomes in Randomized Controlled Trials. *JAMA*. 2009;302(9):977. doi:10.1001/jama.2009.1242.

434. Loder E, Groves T, Macauley D. Registration of observational studies. *BMJ*. 2010;340:c950. doi:10.1136/bmj.c950.

435. Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ*. 2010;340(mar23 1):c869-c869. doi:10.1136/bmj.c869.

436. Lash TL, Collin LJ, van Dyke ME. The Replication Crisis in Epidemiology: Snowball, Snow Job, or Winter Solstice? *Current Epidemiology Reports*. 2018;5(2):175-183. doi:10.1007/s40471-018-0148-x.

437. MacCoun RJ, Perlmutter S. Blind analysis: Hide results to seek the truth. *Nature*. 2015;526(7572):187-189. doi:10.1038/526187a.

438. Ioannidis JPA. The Proposal to Lower P Value Thresholds to 005. *JAMA*. 2018;319(14):1429-1430. doi:10.1001/jama.2018.1536.

439. Evidence-Based Medicine Working Group. Evidence-Based Medicine. *JAMA*. 1992;268(17):2420. doi:10.1001/jama.1992.03490170092032.

440. Sauerbrei W, Abrahamowicz M, Altman DG, Le Cessie S, Carpenter JR. STRengthening Analytical Thinking for Observational Studies: the STRATOS initiative. *Statistics in Medicine*. 2014;33(30):5413-5432. doi:10.1002/sim.6265.

441. Sarewitz D. The pressure to publish pushes down quality. *Nature*. 2016;533(7602):147. doi:10.1038/533147a.

442. Madigan D, Ryan PB, Schuemie MJ. Does design matter? Systematic evaluation of the impact of analytical choices on effect estimates in observational studies. *Therapeutic Advances in Drug Safety*. 2013;4(2):53-62. doi:10.1177/2042098613477445.

443. Greenland S. Modeling and variable selection in epidemiologic analysis. *American Journal of Public Health*. 1989;79(3):340-349. doi:10.2105/AJPH.79.3.340.

444. Gelman A. How do we choose our default methods. In: Lin X, Genest C, Banks DL, Molenberghs G, Scott DW, Wang J-L, eds. *Past, Present, and Future of Statistical Science*: CRC Press, New York; 2014:293-301.

445. Keren G, Wu G. A Bird's-Eye View of the History of Judgment and Decision Making. In: Keren G, Wu G, eds. *The Wiley Blackwell Handbook of Judgment and Decision Making*. Vol. 82. Chichester, UK: John Wiley & Sons, Ltd; 2015:1-39.

446. Wynder EL, Higgins IT, Harris RE. The wish bias. *Journal of Clinical Epidemiology*. 1990;43(6):619-621. doi:10.1016/0895-4356(90)90167-N.

447. Schulz KF. Randomised trials, human nature, and reporting guidelines. *The Lancet*. 1996;348(9027):596-598. doi:10.1016/S0140-6736(96)01201-9.

448. Tversky A, Kahneman D. Judgment under Uncertainty: Heuristics and Biases. *Science*. 1974;185(4157):1124-1131.

449. Thaler RH. From Cashews to Nudges: The Evolution of Behavioral Economics. *American Economic Review*. 2018;108(6):1265-1287. doi:10.1257/aer.108.6.1265.

450. Kahneman D. A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*. 2003;58(9):697-720. doi:10.1037/0003-066X.58.9.697.

451. Grüne-Yanoff T. Reflections on the 2017 Nobel Memorial Prize Awarded to Richard Thaler. *Erasmus Journal for Philosophy and Economics*. 2017;10(2):61. doi:10.23941/ejpe.v10i2.307.

452. Taleb NN. *Fooled by randomness: The hidden role of chance in life and in the markets.* 2nd ed. New York: Thomson/Texere; 2004.

453. Taleb NN. *The black swan: The impact of the highly improbable.* 2nd ed., Random trade pbk. ed. New York: Random House Trade Paperbacks; 2010.

454. Meehl PE. *Clinical Versus Statistical Prediction*. Minneapolis, MN: University of Minnesota Press; 1954.

455. Politser P. Decision analysis and clinical judgment. A re-evaluation. *Medical Decision Making*. 1981;1(4):361-389. doi:10.1177/0272989X8100100406.

456. Dawes RM, Faust D, Meehl PE. Clinical versus actuarial judgment. *Science*. 1989;243(4899):1668-1674. doi:10.1126/science.2648573.

457. Eddy DM. Probabilistic reasoning in clinical medicine: Problems and opportunities. In: Kahneman D, Slovic P, Tversky A, eds. *Judgment under Uncertainty: Heuristics and Biases*. New York: Cambridge University Press; 1982:249-267.

458. Redelmeier DA, Tversky A. Discrepancy between Medical Decisions for Individual Patients and for Groups. *New England Journal of Medicine*. 1990;322(16):1162-1164.

459. Hershberger PJ, Markert RJ, Part HM, Cohen SM, Finger WW. Understanding and addressing cognitive bias in medical education. *Advances in Health Sciences Education*. 1996;1(3):221-226. doi:10.1007/BF00162919.

460. Croskerry P. The Cognitive Imperative Thinking about How We Think. *Academic Emergency Medicine*. 2000;7(11):1223-1231. doi:10.1111/j.1553-2712.2000.tb00467.x.

461. Bornstein BH, Emler AC. Rationality in medical decision making: a review of the literature on doctors' decision-making biases. *Journal of Evaluation in Clinical Practice*. 2001;7(2):97-107. doi:10.1046/j.1365-2753.2001.00284.x.

462. Redelmeier DA, Ferris LE, Tu JV, Hux JE, Schull MJ. Problems for clinical judgement: introducing cognitive psychology as one more basic science. *Canadian Medical Association Journal*. 2001;164(3):358.

463. Blumenthal-Barby JS, Krieger H. Cognitive Biases and Heuristics in Medical Decision Making. *Medical Decision Making*. 2015;35(4):539-557. doi:10.1177/0272989X14547740.

464. Ludolph R, Schulz PJ. Debiasing Health-Related Judgments and Decision Making: A Systematic Review. *Medical Decision Making*. 2018;38(1):3-13. doi:10.1177/0272989X17716672.

465. Vandenbroucke JP. Medical journals and the shaping of medical knowledge*. *The Lancet*. 1998;352(9145):2001-2006. doi:10.1016/S0140-6736(98)10208-8.

466. Catalogue of Bias Collaboration. Confirmation bias - Catalog of Bias. https://catalogofbias.org/biases/confirmation-bias/. Accessed May 26, 2018.

467. Catalogue of Bias Collaboration. Positive results bias - Catalog of Bias. https://catalogofbias.org/biases/positive-results-bias/. Accessed May 26, 2018.

468. Catalogue of Bias Collaboration. Biases of rhetoric - Catalog of Bias. https://catalogofbias.org/biases/biases-of-rhetoric/. Accessed May 26, 2018.

469. Cain DM. Everyone's a Little Bit Biased (Even Physicians). *JAMA*. 2008;299(24):2893. doi:10.1001/jama.299.24.2893.

470. Seshia SS, Makhinson M, Young GB. 'Cognitive biases plus': covert subverters of healthcare evidence. *Evidence Based Medicine*. 2016;21(2):41-45. doi:10.1136/ebmed-2015-110302.

471. Marriott N. The future of statistical thinking. *Significance*. 2014;11(5):78-80. doi:10.1111/j.1740-9713.2014.00787.x.

472. Spiegelhalter DJ. Trust in numbers. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 2017;180(4):1-16. doi:10.1111/rssa.12302.

473. Rosenblueth A, Wiener N. The Role of Models in Science. *Philosophy of Science*. 1945;12(4):316-321.

474. Wild CJ, Pfannkuch M. Statistical Thinking in Empirical Enquiry. *International Statistical Review*. 1999;67(3):223-248. doi:10.1111/j.1751-5823.1999.tb00442.x.

475. Evans JSBT, Stanovich KE. Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*. 2013;8(3):223-241. doi:10.1177/1745691612460685.

476. Frankish K, Evans JSBT. The duality of mind: An historical perspective. In: Evans JSBT, Frankish K, eds. *In two minds: Dual processes and beyond / edited by Jonathan St. B.T. Evans and Keith Frankish*. Oxford: Oxford University Press; 2009.

477. Gigerenzer G, Gaissmaier W. Heuristic Decision Making. *Annual Review of Psychology*. 2011;62(1):451-482. doi:10.1146/annurev-psych-120709-145346.

478. Evans JSBT. Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*. 2008;59:255-278. doi:10.1146/annurev.psych.59.103006.093629.

479. Kahneman D, Klein G. Conditions for intuitive expertise: A failure to disagree. *American Psychologist*. 2009;64(6):515-526. doi:10.1037/a0016755.

480. Gigerenzer G. *Simply rational: Decision making in the real world*. Oxford, New York: Oxford University Press; 2015.

481. Keren G. A Tale of Two Systems: A Scientific Advance or a Theoretical Stone Soup? Commentary on Evans & Stanovich (2013). *Perspectives on Psychological Science*. 2013;8(3):257-262. doi:10.1177/1745691613483474.

482. Gigerenzer G. Why Heuristics Work. *Perspectives on Psychological Science*. 2008;3(1):20-29. doi:10.1111/j.1745-6916.2008.00058.x.

483. Boyd R, Richerson PJ. *The origin and evolution of cultures*. New York, Oxford: Oxford University Press; 2005.

484. Masicampo EJ, Baumeister RF. Toward a physiology of dual-process reasoning and judgment: lemonade, willpower, and expensive rule-based analysis. *Psychological Science*. 2008;19(3):255-260. doi:10.1111/j.1467-9280.2008.02077.x.

485. Magistretti PJ, Allaman I. A cellular perspective on brain energy metabolism and functional imaging. *Neuron*. 2015;86(4):883-901. doi:10.1016/j.neuron.2015.03.035.

486. Ferrer i Cancho R, Sole RV. Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences of the United States of America*. 2003;100(3):788-791. doi:10.1073/pnas.0335980100.

487. Kool W, McGuire JT, Rosen ZB, Botvinick MM. Decision making and the avoidance of cognitive demand. *Journal of Experimental Psychology: General*. 2010;139(4):665-682. doi:10.1037/a0020198.

488. Reber R, Greifeneder R. Processing Fluency in Education: How Metacognitive Feelings Shape Learning, Belief Formation, and Affect. *Educational Psychologist*. 2017;52(2):84-103. doi:10.1080/00461520.2016.1258173.

489. Westbrook A, Braver TS. Cognitive effort: A neuroeconomic approach. *Cognitive, Affective & Behavioral Neuroscience*. 2015;15(2):395-415. doi:10.3758/s13415-015-0334-y.

490. Zohar D, Tzischinsky O, Epstein R, Lavie P. The Effects of Sleep Loss on Medical Residents' Emotional Reactions to Work Events: a Cognitive-Energy Model. *Sleep*. 2005;28(1):47-54. doi:10.1093/sleep/28.1.47.

491. Natale V, Alzani A, Cicogna P. Cognitive efficiency and circadian typologies: a diurnal study. *Personality and Individual Differences*. 2003;35(5):1089-1105. doi:10.1016/S0191-8869(02)00320-3.

492. Amer T, Campbell KL, Hasher L. Cognitive Control As a Double-Edged Sword. *Trends in Cognitive Sciences*. 2016;20(12):905-915. doi:10.1016/j.tics.2016.10.002.

493. Mitchell PJ, Redman JR. Effects of caffeine, time of day and user history on study-related performance. *Psychopharmacology*. 1992;109(1-2):121-126. doi:10.1007/BF02245489.

494. Wardle MC, Treadway MT, Wit H de. Caffeine increases psychomotor performance on the effort expenditure for rewards task. *Pharmacology, Biochemistry, and Behavior*. 2012;102(4):526-531. doi:10.1016/j.pbb.2012.06.016.

495. Stanovich KE. *Rationality and the reflective mind*. New York: Oxford University Press; 2011.

496. Neys W de, Rossi S, Houdé O. Bats, balls, and substitution sensitivity: cognitive misers are no happy fools. *Psychonomic Bulletin & Review*. 2013;20(2):269-273. doi:10.3758/s13423-013-0384-5.

497. Hertwig R, Herzog SM, Schooler LJ, Reimer T. Fluency heuristic: a model of how the mind exploits a by-product of information retrieval. *Journal of Experimental Psychology. Learning, Memory, and Cognition*. 2008;34(5):1191-1206. doi:10.1037/a0013025.

498. Reber R. Availability. In: Pohl RF, ed. *Cognitive Illusions*. 2nd ed. New York: Routledge; 2017:185-203. http://www.tandfebooks.com/isbn/9781315696935.

499. Rodriguez F, Rhodes RE, Miller KF, Shah P. Examining the influence of anecdotal stories and the interplay of individual differences on reasoning. *Thinking & Reasoning*. 2016;22(3):274-296. doi:10.1080/13546783.2016.1139506.

500. Wilcken H. The thinking that drives low value care. https://www.doctorportal.com.au/mjainsight/2017/17/the-thinking-that-drives-low-value-care/.

501. Haselton MG, Nettle D, Murray DR. The Evolution of Cognitive Bias. In: Buss DM, ed. *Handbook of Evolutionary Psychology*. Second: John Wiley & Sons, Inc; 2016.

502. Bornstein RF, Craver-Lemley C. Mere exposure effect. In: Pohl RF, ed. *Cognitive Illusions*. 2nd ed. New York: Routledge; 2017:256-275. http://www.tandfebooks.com/isbn/9781315696935.

503. Lombrozo T, Vasilyeva N. Causal Explanation. In: Waldmann MR, ed. *The Oxford handbook of causal reasoning*. New York, NY: Oxford University Press; 2017. *Oxford library of psychology*.

504. Le Pelley ME, Griffiths O, Beesley T. Associative Accounts of Causal Cognition. In: Waldmann MR, ed. *The Oxford handbook of causal reasoning*. New York, NY: Oxford University Press; 2017. *Oxford library of psychology*.

505. Johnson SGB, Ahn W-k. Causal Mechanisms. In: Waldmann MR, ed. *The Oxford handbook of causal reasoning*. New York, NY: Oxford University Press; 2017. *Oxford library of psychology*.

506. Hayes BK, Hawkins GE, Newell BR. Consider the alternative: The effects of causal knowledge on representing and using alternative hypotheses in judgments under uncertainty. *Journal of Experimental Psychology. Learning, Memory, and Cognition*. 2016;42(5):723-739. doi:10.1037/xlm0000205.

507. Ajzen I. Intuitive theories of events and the effects of base-rate information on prediction. *Journal of Personality and Social Psychology*. 1977;35(5):303-314. doi:10.1037//0022-3514.35.5.303.

508. Krynski TR, Tenenbaum JB. The role of causality in judgment under uncertainty. *Journal of Experimental Psychology: General*. 2007;136(3):430-450. doi:10.1037/0096-3445.136.3.430.

509. Zakay D, Fleisig D. Motivation and Heuristic Thinking. In: Kreitler S, ed. *Cognition and motivation: Forging an interdisciplinary perspective*. Cambridge: Cambridge University Press; 2013:289-306.

510. Arah OA. 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(1):5. doi:10.1186/1742-7622-5-5.

511. Gelman A. Causality and Statistical Learning. *American Journal of Sociology*. 2010;117(3):955-966. doi:10.1086/662659.

512. Pinker S. *The stuff of thought: Language as a window into human nature*. London: Penguin; 2008.

513. Matute H, Yarritu I, Vadillo MA. Illusions of causality at the heart of pseudoscience. *British Journal of Psychology*. 2011;102(3):392-405. doi:10.1348/000712610X532210.

514. Kaufman JS, Poole C. Looking Back on "Causal Thinking in the Health Sciences". *Annual Review of Public Health*. 2000;21(1):101-119. doi:10.1146/annurev.publhealth.21.1.101.

515. Rips LJ. Causal Thinking. In: Adler JE, Rips LJ, eds. *Reasoning*. Cambridge: Cambridge University Press; 2005:597-631.

516. Reichenbach H, Reichenbach M. *The Direction of Time*. Berkeley, London: University of California Press; 1956 (1991 [printing]).

517. Szabó LE, Rédei M, Hofer-Szabó G. *The Principle of the Common Cause*. Cambridge: Cambridge University Press; 2013.

518. Pearl J, Mackenzie D. *The Book of Why: The New Science of Cause and Effect.* First edition. New York: Basic Books; 2018.

519. Greenland S. Probability versus Popper: An Elaboration of the Insufficiency of Current Popperian Approaches for Epidemiologic Analysis. In: Rothman KJ, ed. *Causal Inference*. Chester Hill, MA: Epidemiology Resources Inc; 1988:95-104.

520. Altman DG, Bland JM. Absence of evidence is not evidence of absence. *BMJ*. 1995;311(7003):485. doi:10.1136/bmj.311.7003.485.

521. Tavel ME. The placebo effect: the good, the bad, and the ugly. *The American Journal of Medicine*. 2014;127(6):484-488. doi:10.1016/j.amjmed.2014.02.002.

522. Sherman R, Hickner J. Academic physicians use placebos in clinical practice and believe in the mind-body connection. *Journal of General Internal Medicine*. 2008;23(1):7-10. doi:10.1007/s11606-007-0332-z.

523. Senn SJ. Francis Galton and regression to the mean. *Significance*. 2011;8(3):124-126. doi:10.1111/j.1740-9713.2011.00509.x.

524. Hróbjartsson A, Gøtzsche PC. Is the Placebo Powerless? An Analysis of Clinical Trials Comparing Placebo with No Treatment. *New England Journal of Medicine*. 2001;344(21):1594-1602. doi:10.1056/NEJM200105243442106.

525. Kienle GS, Kiene H. The powerful placebo effect: Fact or fiction? *Journal of Clinical Epidemiology*. 1997;50(12):1311-1318. doi:10.1016/S0895-4356(97)00203-5.

526. Scudellari M. The science myths that will not die. *Nature*. 2015;528(7582):322-325. doi:10.1038/528322a.

527. Ghezzi P, Jaquet V, Marcucci F, Schmidt HHHW. The oxidative stress theory of disease: levels of evidence and epistemological aspects. *British Journal of Pharmacology*. 2017;174(12):1784-1796. doi:10.1111/bph.13544.

528. Stanovich KE, West RF. On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology*. 2008;94(4):672-695. doi:10.1037/0022-3514.94.4.672.

529. Soll JB, Milkman KL, Payne JW. A User's Guide to Debiasing. In: Keren G, Wu G, eds. *The Wiley Blackwell Handbook of Judgment and Decision Making*. Vol. 61. Chichester, UK: John Wiley & Sons, Ltd; 2015:924-951.

530. Forgas JP, East R. On being happy and gullible: Mood effects on skepticism and the detection of deception. *Journal of Experimental Social Psychology*. 2008;44(5):1362-1367. doi:10.1016/j.jesp.2008.04.010.

531. Pronin E, Gilovich T, Ross L. Objectivity in the Eye of the Beholder: Divergent Perceptions of Bias in Self Versus Others. *Psychological Review*. 2004;111(3):781-799. doi:10.1037/0033-295X.111.3.781.

532. Brown AW, Bohan Brown MM, Allison DB. Belief beyond the evidence: using the proposed effect of breakfast on obesity to show 2 practices that distort scientific evidence. *American Journal of Clinical Nutrition*. 2013;98(5):1298-1308. doi:10.3945/ajcn.113.064410.

533. Kunda Z. The case for motivated reasoning. *Psychological Bulletin*. 1990;108(3):480-498. doi:10.1037/0033-2909.108.3.480.

534. Epley N, Gilovich T. The Mechanics of Motivated Reasoning. *Journal of Economic Perspectives*. 2016;30(3):133-140. doi:10.1257/jep.30.3.133.

535. Cox LAT, Popken DA. Overcoming confirmation bias in causal attribution: a case study of antibiotic resistance risks. *Risk Analysis*. 2008;28(5):1155-1172. doi:10.1111/j.1539-6924.2008.01122.x.

536. Mercier H. Confirmation bias – myside bias. In: Pohl RF, ed. *Cognitive Illusions*. 2nd ed. New York: Routledge; 2017:99-114. http://www.tandfebooks.com/isbn/9781315696935.

537. Marmot M. Facts, opinions and affaires du coeur. *American Journal of Epidemiology*. 1976;103(6):519-526. doi:10.1093/oxfordjournals.aje.a112254.

538. Mercier H, Sperber D. Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*. 2011;34(02):57-74. doi:10.1017/S0140525X10000968.

539. Moore DA, Healy PJ. The trouble with overconfidence. *Psychological Review*. 2008;115(2):502-517. doi:10.1037/0033-295X.115.2.502.

540. Mannes A, Moore D. I know I'm right! A behavioural view of overconfidence. *Significance*. 2013;10(4):10-14. doi:10.1111/j.1740-9713.2013.00674.x.

541. Buehler R, Griffin D, Ross M. Inside the Planning Fallacy: The Causes and Consequences of Optimistic Time Predictions. In: Gilovich T, Griffin D, Kahneman D, eds. *Heuristics and biases: The psychology of intuitive judgement*. Cambridge: Cambridge University Press; 2002:250-270.

542. Hoffrage U. Overconfidence. In: Pohl RF, ed. *Cognitive Illusions*. 2nd ed. New York: Routledge; 2017:291-314. http://www.tandfebooks.com/isbn/9781315696935.

543. Kruger J, Dunning D. Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments. *Journal of Personality and Social Psychology*. 1999;77(6):1121-1134.

544. Feld J, Sauermann J, Grip A de. Estimating the relationship between skill and overconfidence. *Journal of Behavioral and Experimental Economics*. 2017;68:18-24. doi:10.1016/j.socec.2017.03.002.

545. Ehrlinger J, Johnson K, Banner M, Dunning D, Kruger J. Why the Unskilled Are Unaware: Further Explorations of (Absent) Self-Insight Among the Incompetent. *Organizational Behavior and Human Decision Processes*. 2008;105(1):98-121. doi:10.1016/j.obhdp.2007.05.002.

546. Wikipedia. Ambiguity aversion. https://en.wikipedia.org/wiki/Ambiguity_aversion. Accessed June 7, 2018.

547. Trautmann ST, Vieider FM, Wakker PP. Causes of ambiguity aversion: Known versus unknown preferences. *Journal of Risk and Uncertainty*. 2008;36(3):225-243. doi:10.1007/s11166-008-9038-9.

548. Frederick S. Automated Choice Heuristics. In: Gilovich T, Griffin D, Kahneman D, eds. *Heuristics and biases: The psychology of intuitive judgement*. Cambridge: Cambridge University Press; 2002:548-558.

549. Blumenthal-Barby JS. Biases and Heuristics in Decision Making and Their Impact on Autonomy. *The American Journal of Bioethics*. 2016;16(5):5-15. doi:10.1080/15265161.2016.1159750.

550. Probst CA, Shaffer VA, Chan YR. The effect of defaults in an electronic health record on laboratory test ordering practices for pediatric patients. *Health Psychology*. 2013;32(9):995-1002. doi:10.1037/a0032925.

551. Cole SR, Chu H, Greenland S. Maximum Likelihood, Profile Likelihood, and Penalized Likelihood: A Primer. *American Journal of Epidemiology*. 2014;179(2):252-260. doi:10.1093/aje/kwt245.

552. Newgard CD, Lewis RJ. Missing Data: How to Best Account for What Is Not Known. *JAMA*. 2015;314(9):940. doi:10.1001/jama.2015.10516.

553. Kahneman D, Knetsch JL, Thaler RH. Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias. *Journal of Economic Perspectives*. 1991;5(1):193-206. doi:10.1257/jep.5.1.193.

554. Grant AM, Sandberg S. *Originals: How non-conformists change the world*. London: Virgin Digital; 2016.

555. Forgas JP, Laham SM. Halo effects. In: Pohl RF, ed. *Cognitive Illusions*. 2nd ed. New York: Routledge; 2017:276-290. http://www.tandfebooks.com/isbn/9781315696935.

556. Halford GS, Baker R, McCredden JE, Bain JD. How many variables can humans process? *Psychological Science*. 2005;16(1):70-76. doi:10.1111/j.0956-7976.2005.00782.x.

557. VanderWeele TJ, Robins JM. Signed directed acyclic graphs for causal inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2010;72(1):111-127. doi:10.1111/j.1467-9868.2009.00728.x.

558. Berry D, Wathen JK, Newell M. Bayesian model averaging in meta-analysis: vitamin E supplementation and mortality. *Clinical Trials*. 2009;6(1):28-41. doi:10.1177/1740774508101279.

559. Basu A, Manca A. Regression Estimators for Generic Health-Related Quality of Life and Quality-Adjusted Life Years. *Medical Decision Making*. 2012;32(1):56-69. doi:10.1177/0272989X11416988.

560. Schomaker M, Heumann C. Model selection and model averaging after multiple imputation. *Computational Statistics & Data Analysis*. 2014;71:758-770. doi:10.1016/j.csda.2013.02.017.

561. Silberzahn R, Uhlmann EL. Many Hands Make Tight Work. *Nature*. 2015;526(7572):189-191. doi:10.1093/nq/s6-VIII.201.347-e.

562. Boulesteix A-L, Wilson R, Hapfelmeier A. Towards evidence-based computational statistics: lessons from clinical research on the role and design of real-data benchmark studies. *BMC Medical Research Methodology*. 2017;17(1):138. doi:10.1186/s12874-017-0417-2.

563. Gharibzadeh S, Mansournia MA, Foroushani A, et al. Comparing Different Propensity Score Estimation Methods for Estimating the Marginal Causal Effect through Standardization to Propensity Scores. *Communications in Statistics - Simulation and Computation*. 2017;0918(July):0. doi:10.1080/03610918.2017.1300267.

564. Silberzahn R, Uhlmann EL, Martin DP, et al. Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Advances in Methods and Practices in Psychological Science*. 2018;1(3):337-356. doi:10.1177/2515245917747646.

565. Dembe AE, Partridge JS, Geist LC. Statistical software applications used in health services research: analysis of published studies in the U.S. *BMC Health Services Research*. 2011;11(1):252. doi:10.1186/1472-6963-11-252.

566. Wicherts JM, Veldkamp CLS, Augusteijn HEM, Bakker M, van Aert RCM, van Assen MALM. Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. *Frontiers in Psychology*. 2016;7:1832. doi:10.3389/fpsyg.2016.01832.

567. Wang J, Johnson DE. An Examination of Discrepancies in Multiple Imputation Procedures Between SAS and SPSS. *The American Statistician*. 2018;86:1-9. doi:10.1080/00031305.2018.1437078.

568. Morris TP, Kahan BC, White IR. Choosing sensitivity analyses for randomised trials: principles. *BMC Medical Research Methodology*. 2014;14(1):11. doi:10.1186/1471-2288-14-11.

569. Drake C. Effects of misspecification on the propensity score on estimatiors of treatment effects. *Biometrics*. 1993;49(4):1231-1236.

570. Rubin DB. Estimating Causal Effects from Large Data Sets Using Propensity Scores. *Annals of Internal Medicine*. 1997;127(8_Part_2):757. doi:10.7326/0003-4819-127-8_Part_2-199710151-00064.

571. Cook NR, Cole SR, Hennekens CH. Use of a Marginal Structural Model to Determine the Effect of Aspirin on Cardiovascular Mortality in the Physicians' Health Study. *American Journal of Epidemiology*. 2002;155(11):1045-1053. doi:10.1093/aje/155.11.1045.

572. Cepeda MS. Comparison of Logistic Regression versus Propensity Score When the Number of Events Is Low and There Are Multiple Confounders. *American Journal of Epidemiology*. 2003;158(3):280-287. doi:10.1093/aje/kwg115.

573. Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermeulen MJ. Analysis of Observational Studies in the Presence of Treatment Selection Bias: Effects of Invasive CardiacManagement on AMI Survival Using Propensity Score and Instrumental Variable Methods. *JAMA*. 2007;297(3):278. doi:10.1001/jama.297.3.278.

574. Ali MS, Groenwold RHH, Klungel OH. Propensity score methods and unobserved covariate imbalance: Comments on "squeezing the balloon". *Health Services Research*. 2014;49(3):1074-1082. doi:10.1111/1475-6773.12152.

575. Davies NM, Thomas KH, Taylor AE, et al. How to compare instrumental variable and conventional regression analyses using negative controls and bias plots. *International Journal of Epidemiology*. 2017;46(6):2067-2077. doi:10.1093/ije/dyx014.

576. Greenland S. Basic methods for sensitivity analysis of biases. *International Journal of Epidemiology*. 1996;25(6):1107-1116. doi:10.1093/ije/25.6.1107.

577. Liu W, Kuramoto SJ, Stuart EA. An Introduction to Sensitivity Analysis for Unobserved Confounding in Nonexperimental Prevention Research. *Prevention Science*. 2013;14(6):570-580. doi:10.1007/s11121-012-0339-5.

578. Lash TL, Fox MP, Fink AK. *Applying Quantitative Bias Analysis to Epidemiologic Data*. New York, NY: Springer New York; 2009.

579. Arah OA. Bias Analysis for Uncontrolled Confounding in the Health Sciences. *Annual Review of Public Health*. 2017;38(1):annurev-publhealth-032315-021644. doi:10.1146/annurev-publhealth-032315-021644.

580. VanderWeele TJ, Ding P. Sensitivity Analysis in Observational Research: Introducing the E-Value. *Annals of Internal Medicine*. 2017;167(4):268-274. doi:10.7326/M16-2607.

581. Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*. 2015;34(28):3661-3679. doi:10.1002/sim.6607.

582. Stürmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *Journal of Clinical Epidemiology*. 2006;59(5):437-447. doi:10.1016/j.jclinepi.2005.07.004.

583. Gayat E, Pirracchio R, Resche-Rigon M, Mebazaa A, Mary J-Y, Porcher R. Propensity scores in intensive care and anaesthesiology literature: a systematic review. *Intensive Care Medicine*. 2010;36(12):1993-2003. doi:10.1007/s00134-010-1991-5.

584. Chen Y, Briesacher BA. Use of instrumental variable in prescription drug research with observational data: a systematic review. *Journal of Clinical Epidemiology*. 2011;64(6):687-700. doi:10.1016/j.jclinepi.2010.09.006.

585. Nietert PJ, Wahlquist AE, Herbert TL. Characteristics of recent biostatistical methods adopted by researchers publishing in general/internal medicine journals. *Statistics in Medicine*. 2013;32(1):1-10. doi:10.1002/sim.5311.

586. Yang S, Eaton CB, Lu J, Lapane KL. Application of marginal structural models in pharmacoepidemiologic studies: a systematic review. *Pharmacoepidemiology and Drug Safety*. 2014;23(6):560-571. doi:10.1002/pds.3569.

587. Laborde-Castérot H, Agrinier N, Thilly N. Performing both propensity score and instrumental variable analyses in observational studies often leads to discrepant results: a systematic review. *Journal of Clinical Epidemiology*. 2015;68(10):1232-1240. doi:10.1016/j.jclinepi.2015.04.003.

588. Meaney C, Moineddin R, Voruganti T, O'Brien MA, Krueger P, Sullivan F. Text mining describes the use of statistical and epidemiological methods in published medical research. *Journal of Clinical Epidemiology*. 2016;74:124-132. doi:10.1016/j.jclinepi.2015.10.020.

589. *EndNote X7*. New York, NY, USA: Thomson Reuters.

590. *Microsoft SQL Server*. Redmond WA, USA: Microsoft Corporation.

591. *Microsoft Visual Studio*. Redmond WA, USA: Microsoft Corporation.

592. Mythicsoft. *FileLocator Pro*. Cambridge, UK: Mythicsoft; 2017. https://www.mythicsoft.com/filelocatorpro.

593. *Stata/IC 15*. College Station, TX, USA: StataCorp.

594. Glynn RJ, Schneeweiss S, Stürmer T. Indications for propensity scores and review of their use in pharmacoepidemiology. *Basic & Clinical Pharmacology & Toxicology*. 2006;98(3):253-259. doi:10.1111/j.1742-7843.2006.pto_293.x.

595. Penning de Vries BBL, Groenwold RHH. Cautionary note: propensity score matching does not account for bias due to censoring. *Nephrology Dialysis Transplantation*. 2018;33(6):914-916. doi:10.1093/ndt/gfx198.

596. Patel A, Billot L. Reality and Truth. *Circulation*. 2017;136(3):260-262. doi:10.1161/CIRCULATIONAHA.117.029233.

597. R Core Team. *R*. Vienna, Austria: R Foundation for Statistical Computing. https://www.r-project.org.

598. *SAS*. Cary, NC, USA: SAS Institute Inc.

599. *IBM SPSS*. Armonk, NY, USA: IBM Corp.

600. *Stata*. College Station, TX, USA: StataCorp.

601. *JMP*. Cary, NC, USA: SAS Institute Inc.

602. *Microsoft Excel.* Version: 2015. Redmond WA, USA: Microsoft Corporation.

603. *GraphPad Prism*. San Diego, CA, USA: GraphPad Software Inc.

604. Tukey JW. We Need Both Exploratory and Confirmatory. *The American Statistician*. 1980;34(1):23-25. doi:10.1080/00031305.1980.10482706.

605. Harel O, Mitchell EM, Perkins NJ, et al. Multiple Imputation for Incomplete Data in Epidemiologic Studies. *American Journal of Epidemiology*. 2018;187(3):576-584. doi:10.1093/aje/kwx349.

606. Bartlett JW, Harel O, Carpenter JR. Asymptotically Unbiased Estimation of Exposure Odds Ratios in Complete Records Logistic Regression. *American Journal of Epidemiology*. 2015;182(8):730-736. doi:10.1093/aje/kwv114.

607. Eekhout I, Boer RM de, Twisk JWR, de Vet, Henrica C. W., Heymans MW. Missing Data: A Systematic Review of How They Are Reported and Handled. *Epidemiology*. 2012;23(5):729-732. doi:10.1097/EDE.0b013e3182576cdb.

608. Vandenbroucke JP. When are observational studies as credible as randomised trials? *The Lancet*. 2004;363(9422):1728-1731. doi:10.1016/S0140-6736(04)16261-2.

609. Bosco JLF, Silliman Ra, Thwin SS, et al. A most stubborn bias: no adjustment method fully resolves confounding by indication in observational studies. *Journal of Clinical Epidemiology*. 2010;63(1):64-74. doi:10.1016/j.jclinepi.2009.03.001.

610. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41-55. doi:10.1093/biomet/70.1.41.

611. Hade EM, Lu B. Bias associated with using the estimated propensity score as a regression covariate. *Statistics in Medicine*. 2014;33(1):74-87. doi:10.1002/sim.5884.

612. Ali MS, Groenwold RHH, Belitser SV, et al. Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review. *Journal of Clinical Epidemiology*. 2015;68(2):122-131. doi:10.1016/j.jclinepi.2014.08.011.

613. Zakrison TL, Austin PC, McCredie VA. A systematic review of propensity score methods in the acute care surgery literature: avoiding the pitfalls and proposing a set of reporting guidelines. *European Journal of Trauma and Emergency Surgery*. 2018;44(3):385-395. doi:10.1007/s00068-017-0786-6.

614. Lonjon G, Porcher R, Ergina P, Fouet M, Boutron I. Potential Pitfalls of Reporting and Bias in Observational Studies With Propensity Score Analysis Assessing a Surgical Procedure: A Methodological Systematic Review. *Annals of Surgery*. 2017;265(5):901-909. doi:10.1097/SLA.0000000000001797.

615. Austin PC, Grootendorst P, Normand S-LT, Anderson GM. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a Monte Carlo study. *Statistics in Medicine*. 2007;26(4):754-768. doi:10.1002/sim.2618.

616. Greenland S. Causal Analysis in the Health Sciences. *Journal of the American Statistical Association*. 2000;95(449):286-289. doi:10.1080/01621459.2000.10473924.

617. Cadarette SM, Ban JK, Consiglio GP, et al. Diffusion of Innovations model helps interpret the comparative uptake of two methodological innovations: co-authorship network analysis and recommendations for the integration of novel methods in practice. *Journal of Clinical Epidemiology*. 2017;84:150-160. doi:10.1016/j.jclinepi.2016.12.006.

618. Pullenayegum EM, Platt RW, Barwick M, Feldman BM, Offringa M, Thabane L. Knowledge translation in biostatistics: a survey of current practices, preferences, and barriers to the dissemination and uptake of new statistical methods. *Statistics in Medicine*. 2016;35(6):805-818. doi:10.1002/sim.6633.

619. Rogers EM. *Diffusion of innovations.* 5th ed. New York: Free Press; 2003.

620. Kramer MS, Lane DA. Causal propositions in clinical research and practice. *Journal of Clinical Epidemiology*. 1992;45(6):639-649. doi:10.1016/0895-4356(92)90136-B.

621. Greenland S, Poole C. Problems in common interpretations of statistics in scientific articles, expert reports, and testimony. *Jurimetrics*. 2011;51:113-129.

622. Kezios KL, Hayes-Larson E. A Clarification on Causal Questions: We Ask Them More Often Than We Realize. *American Journal of Public Health*. 2018;108(8):e4. doi:10.2105/AJPH.2018.304547.

623. Boutron I, Altman DG, Hopewell S, Vera-Badillo F, Tannock I, Ravaud P. Impact of spin in the abstracts of articles reporting results of randomized controlled trials in the field of cancer: The SPIIN randomized controlled trial. *Journal of Clinical Oncology*. 2014;32(36):4120-4126. doi:10.1200/JCO.2014.56.7503.

624. Shinohara K, Aoki T, So R, et al. Influence of overstated abstract conclusions on clinicians: a web-based randomised controlled trial. *BMJ Open*. 2017;7(12):e018355. doi:10.1136/bmjopen-2017-018355.

625. Pulford BD, Colman AM, Buabang EK, Krockow EM. The persuasive power of knowledge: Testing the confidence heuristic. *Journal of Experimental Psychology. General*. 2018;147(10):1431-1444. doi:10.1037/xge0000471.

626. Svenson O. Motivation, Decision Theory, and Human Decision Making. In: Kreitler S, ed. *Cognition and motivation: Forging an interdisciplinary perspective*. Cambridge: Cambridge University Press; 2013:307-320.

627. Anderson C, Hildreth JAD, Howland L. Is the desire for status a fundamental human motive? A review of the empirical literature. *Psychological Bulletin*. 2015;141(3):574-601. doi:10.1037/a0038781.

628. Bendersky C, Pai J. Status Dynamics. *Annual Review of Organizational Psychology and Organizational Behavior*. 2018;5(1):183-199. doi:10.1146/annurev-orgpsych-032117-104602.

629. Blader SL, Yu S. Are Status and Respect Different or Two Sides of the Same Coin? *Academy of Management Annals*. 2017;11(2):800-824. doi:10.5465/annals.2015.0150.

630. Reiss S. Multifaceted Nature of Intrinsic Motivation: The Theory of 16 Basic Desires. *Review of General Psychology*. 2004;8(3):179-193. doi:10.1037/1089-2680.8.3.179.

631. Neel R, Kenrick DT, White AE, Neuberg SL. Individual differences in fundamental social motives. *Journal of Personality and Social Psychology*. 2016;110(6):887-907. doi:10.1037/pspp0000068.

632. Schaller M, Kenrick DT, Neel R, Neuberg SL. Evolution and human motivation: A fundamental motives framework. *Social and Personality Psychology Compass*. 2017;11(6):e12319. doi:10.1111/spc3.12319.

633. Pettit NC, Yong K, Spataro SE. Holding your place: Reactions to the prospect of status gains and losses. *Journal of Experimental Social Psychology*. 2010;46(2):396-401. doi:10.1016/j.jesp.2009.12.007.

634. Maslow AH. A theory of human motivation. *Psychological Review*. 1943;50(4):370-396.

635. Krems JA, Kenrick DT, Neel R. Individual Perceptions of Self-Actualization: What Functional Motives Are Linked to Fulfilling One's Full Potential? *Personality & Social Psychology Bulletin*. 2017;43(9):1337-1352. doi:10.1177/0146167217713191.

636. Crocker J, Canevello A, Brown AA. Social Motivation: Costs and Benefits of Selfishness and Otherishness. *Annual Review of Psychology*. 2017;68:299-325. doi:10.1146/annurev-psych-010416-044145.

637. McNair B. PR must die: spin, anti-spin and political public relations in the UK, 1997–2004. *Journalism Studies*. 2004;5(3):325-338. doi:10.1080/1461670042000246089.

638. Arunachalam L, Hunter IA, Killeen S. Reporting of Randomized Controlled Trials With Statistically Nonsignificant Primary Outcomes Published in High-impact Surgical Journals. *Annals of Surgery*. 2017;265(6):1141-1145. doi:10.1097/SLA.0000000000001795.

639. Boutron I, Dutton S, Ravaud P, Altman DG. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *JAMA*. 2010;303(20):2058-2064. doi:10.1001/jama.2010.651.

640. Boutron I, Ravaud P. Misrepresentation and distortion of research in biomedical literature. *Proceedings of the National Academy of Sciences of the United States of America*. 2018;115(11):2613-2619. doi:10.1073/pnas.1710755115.

641. Brenner RJ. On the more insidious manifestations of bias in scientific reporting. *Journal of the American College of Radiology*. 2010;7(7):490-494. doi:10.1016/j.jacr.2010.02.007.

642. Chan A-W, Ioannidis JPA. Bias, Spin, and Misreporting: Time for Full Access to Trial Protocols and Results. *PLOS Medicine*. 2008;5(11):e230. doi:10.1371/journal.pmed.0050230.

643. Chiu K, Grundy Q, Bero L. 'Spin' in published biomedical literature: A methodological systematic review. *PLOS Biology*. 2017;15(9):e2002173. doi:10.1371/journal.pbio.2002173.

644. Fletcher RH, Black B. "Spin" in Scientific Writing: Scientific Mischief and Legal Jeopardy. *Medicine and Law*. 2007;26:511-525.

645. Glasziou P, Altman DG, Bossuyt PMM, et al. Reducing waste from incomplete or unusable reports of biomedical research. *The Lancet*. 2014;383(9913):267-276. doi:10.1016/S0140-6736(13)62228-X.

646. Haneef R, Yavchitz A, Ravaud P, et al. Interpretation of health news items reported with or without spin: protocol for a prospective meta-analysis of 16 randomised controlled trials. *BMJ Open*. 2017;7(11):e017425. doi:10.1136/bmjopen-2017-017425.

647. Hoffman SJ, Justicz V. Automatically quantifying the scientific quality and sensationalism of news records mentioning pandemics: validating a maximum entropy machine-learning model. *Journal of Clinical Epidemiology*. 2016;75:47-55. doi:10.1016/j.jclinepi.2015.12.010.

648. Horton R. The rhetoric of research. *BMJ*. 1995;310(6985):985-987. doi:10.1136/bmj.310.6985.985.

649. Knottnerus JA, Tugwell P. The way in which effects are analyzed and communicated can make a difference for decision making. *Journal of Clinical Epidemiology*. 2016;72:1-3. doi:10.1016/j.jclinepi.2016.02.005.

650. Lazarus C, Haneef R, Ravaud P, Boutron I. Classification and prevalence of spin in abstracts of non-randomized studies evaluating an intervention. *BMC Medical Research Methodology*. 2015;15(1):85. doi:10.1186/s12874-015-0079-x.

651. Lazarus C, Haneef R, Ravaud P, Hopewell S, Altman DG, Boutron I. Peer reviewers identified spin in manuscripts of nonrandomized studies assessing therapeutic interventions, but their impact on spin in abstract conclusions was limited. *Journal of Clinical Epidemiology*. 2016;77(1):44-51. doi:10.1016/j.jclinepi.2016.04.012.

652. Mathieu S, Giraudeau B, Soubrier M, Ravaud P. Misleading abstract conclusions in randomized controlled trials in rheumatology: Comparison of the abstract conclusions and the results section. *Joint Bone Spine*. 2012;79(3):262-267. doi:10.1016/j.jbspin.2011.05.008.

653. McGrath TA, McInnes MDF, van Es N, Leeflang MMG, Korevaar DA, Bossuyt PMM. Overinterpretation of Research Findings: Evidence of "Spin" in Systematic Reviews of Diagnostic Accuracy Studies. *Clinical Chemistry*. 2017;63(8):1353-1362. doi:10.1373/clinchem.2017.271544.

654. Ochodo EA, Haan MC de, Reitsma JB, Hooft L, Bossuyt PMM, Leeflang MMG. Overinterpretation and Misreporting of Diagnostic Accuracy Studies: Evidence of "Spin". *Radiology*. 2013;267(2):581-588. doi:10.1148/radiol.12120527.

655. Patel SV, Chadi SA, Choi J, Colquhoun PHD. The Use of "Spin" in Laparoscopic Lower GI Surgical Trials with Nonsignificant Results. *Diseases of the Colon & Rectum*. 2013;56(12):1388-1394. doi:10.1097/01.dcr.0000436466.50341.c5.

656. Riggs TW. Spin. *Obstetrics & Gynecology*. 2017;129(2):237-238. doi:10.1097/AOG.0000000000001869.

657. Saquib N, Saquib J, Ioannidis JPA. Practices and impact of primary outcome adjustment in randomized controlled trials: meta-epidemiologic study. *BMJ*. 2013;347(jul12 2):f4313-f4313. doi:10.1136/bmj.f4313.

658. Sumner P, Vivian-Griffiths S, Boivin J, et al. The association between exaggeration in health related science news and academic press releases: retrospective observational study. *BMJ*. 2014;349(dec09 7):g7015-g7015. doi:10.1136/bmj.g7015.

659. Tugwell P, Knottnerus JA. Can we measure 'Sensationalisim' and 'Spin'? *Journal of Clinical Epidemiology*. 2016;75:A6-A8. doi:10.1016/j.jclinepi.2016.06.001.

660. Turrentine M. It's All How You "Spin" It: Interpretive Bias in Research Findings in the Obstetrics and Gynecology Literature. *Obstetrics & Gynecology*. 2017;129(2):239-242. doi:10.1097/AOG.0000000000001818.

661. Weingart P. Is There a Hype Problem in Science? If So, How Is It Addressed? In: Jamieson KH, Kahan DM, Scheufele D, eds. *The Oxford handbook on the science of science communication*. Vol. 1. New York NY United States of America: Oxford University Press; 2017:111-118. *Oxford library of psychology*.

662. Yavchitz A, Ravaud P, Hopewell S, Baron G, Boutron I. Impact of adding a limitations section to abstracts of systematic reviews on readers' interpretation: a randomized controlled trial. *BMC Medical Research Methodology*. 2014;14:123. doi:10.1186/1471-2288-14-123.

663. Yavchitz A, Ravaud P, Altman DG, et al. A new classification of spin in systematic reviews and meta-analyses was developed and ranked according to the severity. *Journal of Clinical Epidemiology*. 2016;75:56-65. doi:10.1016/j.jclinepi.2016.01.020.

664. Gøtzsche PC. Readers as research detectives. *Trials*. 2009;10:2. doi:10.1186/1745-6215-10-2.

665. Barry HC, Ebell MH, Shaughnessy AF, Slawson DC, Neitzke F. Family Physicians' Use of Medical Abstracts To Guide Decision Making: Style or Substance? *The Journal of the American Board of Family Practice*. 2001;14(6):437-442.

666. Smith R. Doctors are not scientists. *BMJ*. 2004;328(7454):0-h-0. doi:10.1136/bmj.328.7454.0-h.

667. The Editors. Addressing the Limitations of Structured Abstracts. *Annals of Internal Medicine*. 2004;140(6):480. doi:10.7326/0003-4819-140-6-200403160-00015.

668. Marcelo A, Gavino A, Isip-Tan IT, et al. A comparison of the accuracy of clinical decisions based on full-text articles and on journal abstracts alone: a study among residents in a tertiary care hospital. *Evidence Based Medicine*. 2013;18(2):48-53. doi:10.1136/eb-2012-100537.

669. Saint S, Christakis DA, Saha S, et al. Journal reading habits of internists. *Journal of General Internal Medicine*. 2000;15(12):881-884. doi:10.1046/j.1525-1497.2000.00202.x.

670. Burke DT, Judelson AL, Schneider JC, DeVito MC, Latta D. Reading Habits of Practicing Physiatrists. *American Journal of Physical Medicine & Rehabilitation*. 2002;81(10):779-787. doi:10.1097/00002060-200210000-00011.

671. Macleod MR, Michie S, Roberts I, et al. Biomedical research: increasing value, reducing waste. *The Lancet*. 2014;383(9912):101-104. doi:10.1016/S0140-6736(13)62329-6.

672. Storz-Pfennig P. Potentially unnecessary and wasteful clinical trial research detected in cumulative meta-epidemiological and trial sequential analysis. *Journal of Clinical Epidemiology*. 2017;82:61-70. doi:10.1016/j.jclinepi.2016.11.003.

673. Gore SM, Jones G, Thompson SG. The Lancet’s statistical review process: areas for improvement by authors. *The Lancet*. 1992;340(8811):100-102. doi:10.1016/0140-6736(92)90409-V.

674. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Journal of Clinical Epidemiology*. 2008;61:344-349. doi:10.1016/j.jclinepi.2007.11.008.

675. Cals JWL, Kotz D. Effective writing and publishing scientific papers, part VI: discussion. *Journal of Clinical Epidemiology*. 2013;66(10):1064. doi:10.1016/j.jclinepi.2013.04.017.

676. Carlin JB. Comment: Is Reform Possible Without a Paradigm Shift? *The American Statistician*. 2016;70(suppl). doi:10.1080/00031305.2016.1154108.

677. Mainland D. The significance of "nonsignificance". *Clinical Pharmacology & Therapeutics*. 1963;4(5):580-586. doi:10.1002/cpt196345580.

678. Hernán MA. The C-Word: Scientific Euphemisms Do Not Improve Causal Inference From Observational Data. *American Journal of Public Health*. 2018;108(5):616-619. doi:10.2105/AJPH.2018.304337.

679. Cofield SS, Corona RV, Allison DB. Use of causal language in observational studies of obesity and nutrition. *Obesity Facts*. 2010;3(6):353-356. doi:10.1159/000322940.

680. Höfler M, Venz J, Trautmann S, Miller R. Writing a discussion section: how to integrate substantive and statistical expertise. *BMC Medical Research Methodology*. 2018;18(1):34. doi:10.1186/s12874-018-0490-1.

681. Jones HE, Schooling CM. Let's Require the "T-Word". *American Journal of Public Health*. 2018;108(5):624. doi:10.2105/AJPH.2018.304365.

682. Lipton R, Ødegaard T. Causal thinking and causal language in epidemiology: it's in the details. *Epidemiologic Perspectives & Innovations*. 2005;2:8. doi:10.1186/1742-5573-2-8.

683. Petitti DB. Associations Are Not Effects. *American Journal of Epidemiology*. 1991;133(2):101-102. doi:10.1093/oxfordjournals.aje.a115848.

684. Ahern J. Start With the "C-Word," Follow the Roadmap for Causal Inference. *American Journal of Public Health*. 2018;108(5):621. doi:10.2105/AJPH.2018.304358.

685. Begg MD, March D. Cause and Association: Missing the Forest for the Trees. *American Journal of Public Health*. 2018;108(5):620. doi:10.2105/AJPH.2018.304366.

686. Broadbent A. *Philosophy of Epidemiology*. London: Palgrave Macmillan UK; 2013.

687. Chiolero A. Data Are Not Enough-Hurray For Causality! *American Journal of Public Health*. 2018;108(5):622. doi:10.2105/AJPH.2018.304379.

688. Galea S, Vaughan RD. Moving Beyond the Cause Constraint: A Public Health of Consequence, May 2018. *American Journal of Public Health*. 2018;108(5):602-603. doi:10.2105/AJPH.2018.304390.

689. Green MJ. Calculating Versus Estimating Causal Effects. *American Journal of Public Health*. 2018;108(8):e4-e5. doi:10.2105/AJPH.2018.304546.

690. Imbens GW, Rubin DB. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge: Cambridge University Press; 2015.

691. Kincaid H. Causal modelling, mechanism, and probability in epidemiology. In: Illari PM, Russo F, Williamson J, eds. *Causality in the sciences*. Oxford: Oxford University Press; 2011:70-90.

692. Savitz DA. Re: "Associations Are Not Effects". *American Journal of Epidemiology*. 1991;134(4):442-443. doi:10.1093/oxfordjournals.aje.a116110.

693. Tam CC. Causal thinking and causal language in epidemiology: a cause by any other name is still a cause: response to Lipton and Ødegaard. *Epidemiologic Perspectives & Innovations*. 2006;3:7. doi:10.1186/1742-5573-3-7.

694. Simon HA. Spurious Correlation: A Causal Interpretation. *Journal of the American Statistical Association*. 1954;49(267):467-479. doi:10.1080/01621459.1954.10483515.

695. Biber D, Gray B. Nominalizing the verb phrase in academic science writing. In: Aarts B, Close J, Leech G, Wallis S, eds. *The Verb Phrase in English*. Cambridge: Cambridge University Press; 2010:99-132.

696. Clemen G. The Concept of Hedging: Origins, Approaches and Definitions. In: Markkanen R, Schröder H, eds. *Hedging and Discourse: Approaches to the Analysis of a Pragmatic Phenomenon in Academic Texts*. Vol. 24. Berlin, Boston: De Gruyter; 1997:235-248.

697. Crystal D. *A dictionary of linguistics and phonetics*. 6th ed. Malden MA, Oxford: Blackwell Pub; 2008.

698. Eastwood J. *Oxford guide to English grammar*. Oxford: Oxford University Press; 1994.

699. Grabe W, Kaplan RB. On the Writing of Science and the Science of Writing: Hedging in Science Text and Elsewhere. In: Markkanen R, Schröder H, eds. *Hedging and Discourse: Approaches to the Analysis of a Pragmatic Phenomenon in Academic Texts*. Vol. 24. Reprint 2010. Berlin, Boston: De Gruyter; 1997:151. *Research in Text Theory*; 24.

700. Haase C. The Syntax of Cause and Effect. https://www.tu-chemnitz.de/phil/english/sections/linguist/independent/kursmaterialien/caus/. Accessed December 19, 2016.

701. Huschová P. Exploring modal verbs conveying possibility in academic discourse. *Discourse and Interaction*. 2015;8(2):35-47. doi:10.5817/DI2015-2-35.

702. Skelton J. How to tell the truth in The British Medical Journal: patterns of judgement in the 19th and 20th centuries. In: Markkanen R, Schröder H, eds. *Hedging and Discourse: Approaches to the Analysis of a Pragmatic Phenomenon in Academic Texts*. Berlin, Boston: De Gruyter; 1997:42-64.

703. Trask RL. *A Dictionary of Grammatical Terms in Linguistics*. New York: Routledge; 1992.

704. Wiley J, Myers JL. Availability and Accessibility of Information and Causal Inferences From Scientific Text. *Discourse Processes*. 2003;36(2):109-129. doi:10.1207/S15326950DP3602_2.

705. Xuelan F, Kennedy G. Expressing Causation in Written English. *RELC Journal*. 1992;23(1):62-80. doi:10.1177/003368829202300105.

706. Asghar N. *Automatic Extraction of Causal Relations from Natural Language Texts: A Comprehensive Survey*; 2016.

707. Blanco E, Castell N, Di Moldovan. Causal Relation Extraction. *Proceedings of the Sixth International Conference on Language Resources and Evaluation*. 2008:310-313.

708. Chukharev-Hudilainen E, Saricaoglu A. Causal discourse analyzer: improving automated feedback on academic ESL writing. *Computer Assisted Language Learning*. 2016;29(3):494-516. doi:10.1080/09588221.2014.991795.

709. Dunietz J, Levin L, Carbonell J. Annotating Causal Language Using Corpus Lexicography of Constructions. *Proceedings of LAW IX - The 9th Linguistic Annotation Workshop*. 2015:188-196.

710. Girju R, Moldovan DI. Text Mining for Causal Relations. *Proceedings of the 15th International Florida Artificial Intelligence Research Society Conference*. 2002;(October 2002):360-364.

711. Kilicoglu H. Inferring Implicit Causal Relationships in Biomedical Literature. *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*. 2016;(Cid):46-55.

712. Mihăilă C, Ananiadou S. Semi-supervised learning of causal relations in biomedical scientific discourse. *BioMedical Engineering OnLine*. 2014;13 Suppl 2:S1. doi:10.1186/1475-925X-13-S2-S1.

713. Mihăilă C, Ohta T, Pyysalo S, Ananiadou S. BioCause: Annotating and analysing causality in the biomedical domain. *BMC Bioinformatics*. 2013;14:2. doi:10.1186/1471-2105-14-2.

714. Tirunagari S. *Data Mining of Causal Relations from Text: Analysing Maritime Accident Investigation Reports*; 2015.

715. Li LC, Moja L, Romero A, Sayre EC, Grimshaw JM. Nonrandomized quality improvement intervention trials might overstate the strength of causal inference of their findings. *Journal of Clinical Epidemiology*. 2009;62(9):959-966. doi:10.1016/j.jclinepi.2008.10.008.

716. Ferstl EC, Garnham A, Manouilidou C. Implicit causality bias in English: a corpus of 300 verbs. *Behavior Research Methods*. 2011;43(1):124-135. doi:10.3758/s13428-010-0023-2.

717. Wolff P, Song G. Models of causation and the semantics of causal verbs. *Cognitive Psychology*. 2003;47(3):276-332. doi:10.1016/S0010-0285(03)00036-7.

718. Wikipedia. English grammar. https://en.wikipedia.org/wiki/English_grammar. Accessed November 10, 2016.

719. Wikipedia. Linguistic modality. https://en.wikipedia.org/wiki/Linguistic_modality. Accessed October 25, 2018.

720. Cambridge Dictionary. English Grammar Today on Cambridge Dictionary. http://dictionary.cambridge.org/grammar/british-grammar/. Accessed December 1, 2016.

721. Collins English Dictionary. https://www.collinsdictionary.com/dictionary/english. Accessed March 12, 2018.

722. Dictionary.com. http://www.dictionary.com/. Accessed November 19, 2016.

723. Farlex. TheFreeDictionary: Parts of Speech. http://www.thefreedictionary.com/Parts-of-Speech.htm. Accessed December 15, 2016.

724. Oxford Dictionaries. Oxford Dictionaries Grammar A-Z. https://en.oxforddictionaries.com/grammar/grammar-a-z. Accessed November 19, 2016.

725. Browne C, Culligan B, Phillips J. New Academic Word List. http://www.newacademicwordlist.org. Accessed December 19, 2016.

726. Jackson E, O'Carroll P, Ardington A. The University of Sydney Write Site. http://writesite.elearn.usyd.edu.au/glossary/glossary.htm. Accessed December 21, 2016.

727. Oxford Dictionaries. Definition of *be*. https://en.oxforddictionaries.com/definition/be. Accessed October 25, 2018.

728. Wikipedia. Copula (linguistics). https://en.wikipedia.org/wiki/Copula_(linguistics). Accessed October 24, 2018.

729. Hyland K, Tse P. Hooking the reader: a corpus study of evaluative that in abstracts. *English for Specific Purposes*. 2005;24(2):123-139. doi:10.1016/j.esp.2004.02.002.

730. Wikipedia. Predicate (grammar). https://en.wikipedia.org/wiki/Predicate_(grammar). Accessed June 28, 2018.

731. Hobbs JR. Toward a useful concept of causality for lexical semantics. *Journal of Semantics*. 2005;22(2):181-209. doi:10.1093/jos/ffh024.

732. Do QX, Chan YS, Roth D. Minimally supervised event causality identification. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing; 2011:294-303.

733. Ariel M. Research paradigms in pragmatics. In: Allan K, Jaszczolt KM, eds. *The Cambridge handbook of pragmatics*. Cambridge, UK: Cambridge University Press; 2012:23-46. *Cambridge handbooks in language and linguistics*.

734. Keysar B, Barr DJ. Self-Anchoring in Conversation: Why Language Users Do Not Do What They "Should". In: Gilovich T, Griffin D, Kahneman D, eds. *Heuristics and biases: The psychology of intuitive judgement*. Cambridge: Cambridge University Press; 2002:150-166.

735. Goodman ND, Frank MC. Pragmatic Language Interpretation as Probabilistic Inference. *Trends in Cognitive Sciences*. 2016;20(11):818-829. doi:10.1016/j.tics.2016.08.005.

736. Norris SP, Phillips LM. Interpreting pragmatic meaning when reading popular reports of science. *Journal of Research in Science Teaching*. 1994;31(9):947-967. doi:10.1002/tea.3660310909.

737. Everitt B, Palmer CR, eds. *Encyclopaedic companion to medical statistics.* 2nd ed. Oxford: Wiley; 2011.

738. Qiu J. Publish or perish in China. *Nature*. 2010;463(7278):142-143. doi:10.1038/463142a.

739. Fanelli D, Costas R, Ioannidis JPA. Meta-assessment of bias in science. *Proceedings of the National Academy of Sciences*. 2017;114(14):3714-3719. doi:10.1073/pnas.1618569114.

740. Bastian H. 'They would say that, wouldn’t they?' A reader’s guide toauthor and sponsor biases in clinical research. *Journal of the Royal Society of Medicine*. 2006;99(12):611-614. doi:10.1258/jrsm.99.12.611.

741. Stead WW. The Complex and Multifaceted Aspects of Conflicts of Interest. *JAMA*. 2017;317(17):1765. doi:10.1001/jama.2017.3435.

742. Altman DG, Goodman SN, Schroter S. How Statistical Expertise Is Used in Medical Research. *JAMA*. 2002;287(21):2817. doi:10.1001/jama.287.21.2817.

743. Goldsmith CH, Jin Y, He F, Thabane L. Statistician credit for collaboration requires extending the number of cited authors in research publications. *Journal of Clinical Epidemiology*. 2018;101:130-131. doi:10.1016/j.jclinepi.2018.04.022.

744. Ericsson KA, Pool R. *Peak: Secrets from the new science of expertise*. Boston: Houghton Mifflin Harcourt; 2016.

745. Hernán MA, Hernández-Díaz S, Robins JM. Randomized Trials Analyzed as Observational Studies. *Annals of Internal Medicine*. 2013;159(8):560-563. doi:10.7326/0003-4819-159-8-201310150-00709.

746. Panagiotou OA, Ioannidis JPA. Primary study authors of significant studies are more likely to believe that a strong association exists in a heterogeneous meta-analysis compared with methodologists. *Journal of Clinical Epidemiology*. 2012;65(7):740-747. doi:10.1016/j.jclinepi.2012.01.008.

747. Arora S, Peters AL, Burner E, Lam CN, Menchine M. Trial to Examine Text Message–Based mHealth in Emergency Department Patients With Diabetes (TExT-MED): A Randomized Controlled Trial. *Annals of Emergency Medicine*. 2014;63(6):745-754.e6. doi:10.1016/j.annemergmed.2013.10.012.

748. Bosi E, Scavini M, Ceriello A, et al. Intensive Structured Self-Monitoring of Blood Glucose and Glycemic Control in Noninsulin-Treated Type 2 Diabetes: The PRISMA randomized trial. *Diabetes Care*. 2013;36(10):2887-2894. doi:10.2337/dc13-0092.

749. Cho J-H, Lee H-C, Lim D-J, Kwon H-S, Yoon K-H. Mobile communication using a mobile phone with a glucometer for glucose control in Type 2 patients with diabetes: as effective as an Internet-based glucose monitoring system. *Journal of Telemedicine and Telecare*. 2009;15(2):77-82. doi:10.1258/jtt.2008.080412.

750. Eakin EG, Winkler EA, Dunstan DW, et al. Living Well With Diabetes: 24-Month Outcomes From a Randomized Trial of Telephone-Delivered Weight Loss and Physical Activity Intervention to Improve Glycemic Control. *Diabetes Care*. 2014;37(8):2177-2185. doi:10.2337/dc13-2427.

751. Faridi Z, Liberti L, Shuval K, Northrup V, Ali A, Katz DL. Evaluating the impact of mobile telephone technology on type 2 diabetic patients' self-management: the NICHE pilot study. *Journal of Evaluation in Clinical Practice*. 2008;14(3):465-469. doi:10.1111/j.1365-2753.2007.00881.x.

752. Hoffmann-Petersen N, Lauritzen T, Bech JN, Pedersen EB. Short-term telemedical home blood pressure monitoring does not improve blood pressure in uncomplicated hypertensive patients. *Journal of Human Hypertension*. 2017;31(2):93-98. doi:10.1038/jhh.2016.43.

753. Jeong JY, Jeon J-H, Bae K-H, et al. Smart Care Based on Telemonitoring and Telemedicine for Type 2 Diabetes Care: Multi-Center Randomized Controlled Trial. *Telemedicine Journal and E-health*. 2018;24(8):604-613. doi:10.1089/tmj.2017.0203.

754. Karhula T, Vuorinen A-L, Rääpysjärvi K, et al. Telemonitoring and Mobile Phone-Based Health Coaching Among Finnish Diabetic and Heart Disease Patients: Randomized Controlled Trial. *Journal of Medical Internet Research*. 2015;17(6):e153. doi:10.2196/jmir.4059.

755. Kumar VS, Wentzell KJ, Mikkelsen T, Pentland A, Laffel LM. The DAILY (Daily Automated Intensive Log for Youth) Trial: A Wireless, Portable System to Improve Adherence and Glycemic Control in Youth with Diabetes. *Diabetes Technology & Therapeutics*. 2004;6(4):445-453. doi:10.1089/1520915041705893.

756. Lim S, Kang SM, Shin H, et al. Improved Glycemic Control Without Hypoglycemia in Elderly Diabetic Patients Using the Ubiquitous Healthcare Service, a New Medical Information System. *Diabetes Care*. 2011;34(2):308-313. doi:10.2337/dc10-1447.

757. Madsen LB, Kirkegaard P, Pedersen EB. Blood pressure control during telemonitoring of home blood pressure. A randomized controlled trial during 6 months. *Blood Pressure*. 2008;17(2):78-86. doi:10.1080/08037050801915468.

758. McKinstry B, Hanley J, Wild SH, et al. Telemonitoring based service redesign for the management of uncontrolled hypertension: multicentre randomised controlled trial. *BMJ*. 2013;346(may24 4):f3030-f3030. doi:10.1136/bmj.f3030.

759. Neumann CL, Menne J, Rieken EM, et al. Blood pressure telemonitoring is useful to achieve blood pressure control in inadequately treated patients with arterial hypertension. *Journal of Human Hypertension*. 2011;25(12):732-738. doi:10.1038/jhh.2010.119.

760. Parati G, Omboni S, Albini F, et al. Home blood pressure telemonitoring improves hypertension control in general practice. The TeleBPCare study. *Journal of Hypertension*. 2009;27(1):198-203. doi:10.1097/HJH.0b013e3283163caf.

761. Quinn CC, Shardell MD, Terrin ML, Barr EA, Ballew SH, Gruber-Baldini AL. Cluster-Randomized Trial of a Mobile Phone Personalized Behavioral Intervention for Blood Glucose Control. *Diabetes Care*. 2011;34(9):1934-1942. doi:10.2337/dc11-0366.

762. Stone RA, Rao RH, Sevick MA, et al. Active Care Management Supported by Home Telemonitoring in Veterans With Type 2 Diabetes: The DiaTel randomized controlled trial. *Diabetes Care*. 2010;33(3):478-484. doi:10.2337/dc09-1012.

763. Wakefield BJ, Koopman RJ, Keplinger LE, et al. Effect of Home Telemonitoring on Glycemic and Blood Pressure Control in Primary Care Clinic Patients with Diabetes. *Telemedicine and e-Health*. 2014;20(3):199-205. doi:10.1089/tmj.2013.0151.

764. Warren R, Carlisle K, Mihala G, Scuffham PA. Effects of telemonitoring on glycaemic control and healthcare costs in type 2 diabetes: A randomised controlled trial. *Journal of Telemedicine and Telecare*. 2018;24(9):586-595. doi:10.1177/1357633X17723943.

765. Wild SH, Hanley J, Lewis SC, et al. Supported Telemonitoring and Glycemic Control in People with Type 2 Diabetes: The Telescot Diabetes Pragmatic Multicenter Randomized

Controlled Trial. *PLOS Medicine*. 2016;13(7):e1002098. doi:10.1371/journal.pmed.1002098.

766. Yoo HJ, Park MS, Kim TN, et al. A Ubiquitous Chronic Disease Care system using cellular phones and the internet. *Diabetic Medicine*. 2009;26(6):628-635. doi:10.1111/j.1464-5491.2009.02732.x.

767. Aikens JE, Zivin K, Trivedi R, Piette JD. Diabetes self-management support using mHealth and enhanced informal caregiving. *Journal of Diabetes and its Complications*. 2014;28(2):171-176. doi:10.1016/j.jdiacomp.2013.11.008.

768. Bernocchi P, Scalvini S, Bertacchini F, Rivadossi F, Muiesan M. Home based telemedicine intervention for patients with uncontrolled hypertension: - a real life - non-randomized study. *BMC Medical Informatics and Decision Making*. 2014;14(1):52. doi:10.1186/1472-6947-14-52.

769. Carral F, Ayala MDC, Fernández JJ, et al. Web-Based Telemedicine System Is Useful for Monitoring Glucose Control in Pregnant Women with Diabetes. *Diabetes Technology & Therapeutics*. 2015;17(5):349-354. doi:10.1089/dia.2014.0223.

770. McFarland M, Davis K, Wallace J, et al. Use of home telehealth monitoring with active medication therapy management by clinical pharmacists in veterans with poorly controlled type 2 diabetes mellitus. *Pharmacotherapy*. 2012;32(5):420-426. doi:10.1002/j.1875-9114.2011.01038.x.

771. Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutic trials. *Journal of Chronic Diseases*. 1967;20(8):637-648. doi:10.1016/0021-9681(67)90041-0.

772. Zuidgeest MGP, Goetz I, Groenwold RHH, Irving E, van Thiel GJMW, Grobbee DE. Series: Pragmatic trials and real world evidence: Paper 1. Introduction. *Journal of Clinical Epidemiology*. 2017;88:7-13. doi:10.1016/j.jclinepi.2016.12.023.

773. Meinecke A-K, Welsing PMJ, Kafatos G, et al. Series: Pragmatic trials and real world evidence: Paper 8. Data collection and management. *Journal of Clinical Epidemiology*. 2017;91:13-22. doi:10.1016/j.jclinepi.2017.07.003.

774. Welsing PMJ, Oude Rengerink K, Collier S, et al. Series: Pragmatic trials and real world evidence: Paper 6. Outcome measures in the real world. *Journal of Clinical Epidemiology*. 2017;90:99-107. doi:10.1016/j.jclinepi.2016.12.022.

775. Ford I, Norrie J, Drazen JM, et al. Pragmatic Trials. *New England Journal of Medicine*. 2016;375(5):454-463. doi:10.1056/NEJMra1510059.

776. Diaz-Ordaz K, Kenward MG, Cohen A, Coleman CL, Eldridge S. Are missing data adequately handled in cluster randomised trials? A systematic review and guidelines. *Clinical Trials*. 2014;11(5):590-600. doi:10.1177/1740774514537136.

777. Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. *BMC Medical Research Methodology*. 2017;17(1):162. doi:10.1186/s12874-017-0442-1.

778. Bell ML, Fiero M, Horton NJ, Hsu C-H. Handling missing data in RCTs; a review of the top medical journals. *BMC Medical Research Methodology*. 2014;14(1):118. doi:10.1186/1471-2288-14-118.

779. Karahalios A, Baglietto L, Carlin JB, English DR, Simpson JA. A review of the reporting and handling of missing data in cohort studies with repeated assessment of exposure measures. *BMC Medical Research Methodology*. 2012;12. doi:10.1186/1471-2288-12-96.

780. Malla L, Perera-Salazar R, McFadden E, Ogero M, Stepniewska K, English M. Handling missing data in propensity score estimation in comparative effectiveness evaluations: a systematic review. *Journal of Comparative Effectiveness Research*. 2018;7(3):271-279. doi:10.2217/cer-2017-0071.

781. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009;338(jun29 1):b2393-b2393. doi:10.1136/bmj.b2393.

782. Sullivan TR, Yelland LN, Lee KJ, Ryan P, Salter AB. Treatment of missing data in follow-up studies of randomised controlled trials: A systematic review of the literature. *Clinical Trials*. 2017;14(4):387-395. doi:10.1177/1740774517703319.

783. Zhang Y, Alyass A, Vanniyasingam T, et al. A systematic survey of the methods literature on the reporting quality and optimal methods of handling participants with missing outcome data for continuous outcomes in randomized controlled trials. *Journal of Clinical Epidemiology*. 2017;88:67-80. doi:10.1016/j.jclinepi.2017.05.016.

784. Bell ML, Fairclough DL. Practical and statistical issues in missing data for longitudinal patient-reported outcomes. *Statistical Methods in Medical Research*. 2014;23(5):440-459. doi:10.1177/0962280213476378.

785. Fitzmaurice GM, Laird NM, Ware JH. *Applied longitudinal analysis.* 2nd ed. Oxford: Wiley; 2011.

786. Rubin DB. Inference and missing data. *Biometrika*. 1976;63(3):581-592. doi:10.1093/biomet/63.3.581.

787. Little RJA, Rubin DB. *Statistical analysis with missing data.* 2nd ed. Hoboken, N.J., Chichester: Wiley; 2002.

788. Seaman S, Galati J, Jackson D, Carlin JB. What Is Meant by "Missing at Random"? *Statistical Science*. 2013;28(2):257-268. doi:10.1214/13-STS415.

789. Mealli F, Rubin DB. Clarifying missing at random and related definitions, and implications when coupled with exchangeability: Table 1. *Biometrika*. 2015;102(4):995-1000. doi:10.1093/biomet/asv035.

790. Mealli F, Rubin DB. 'Clarifying missing at random and related definitions, and implications when coupled with exchangeability'. *Biometrika*. 2016;103(2):491. doi:10.1093/biomet/asw017.

791. Doretti M, Geneletti S, Stanghellini E. Missing Data: A Unified Taxonomy Guided by Conditional Independence. *International Statistical Review*. 2018;86(2):189-204. doi:10.1111/insr.12242.

792. Perkins NJ, Cole SR, Harel O, et al. Principled Approaches to Missing Data in Epidemiologic Studies. *American Journal of Epidemiology*. 2018;187(3):568-575. doi:10.1093/aje/kwx348.

793. National Research Council. *The Prevention and Treatment of Missing Data in Clinical Trials*. Washington, D.C.: National Academies Press; 2010.

794. White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine*. 2010;29(28):2920-2931. doi:10.1002/sim.3944.

795. Lewin A, Brondeel R, Benmarhnia T, Thomas F, Chaix B. Attrition Bias Related to Missing Outcome Data: A Longitudinal Simulation Study. *Epidemiology*. 2018;29(1):87-95. doi:10.1097/EDE.0000000000000755.

796. Thoemmes FJ, Mohan K. Graphical Representation of Missing Data Problems. *Structural Equation Modeling*. 2015;22(4):631-642. doi:10.1080/10705511.2014.937378.

797. Vach W, Blettner M. Biased Estimation of the Odds Ratio in Case-Control Studies due to the Use of Ad Hoc Methods of Correcting for Missing Values for Confounding Variables. *American Journal of Epidemiology*. 1991;134(8):895-907. doi:10.1093/oxfordjournals.aje.a116164.

798. Little RJ, Rubin DB. Causal Effects in Clinical and Epidemiological Studies Via Potential Outcomes: Concepts and Analytical Approaches. *Annual Review of Public Health*. 2000;21(1):121-145. doi:10.1146/annurev.publhealth.21.1.121.

799. Wood AM, White IR, Hillsdon M, Carpenter JR. Comparison of imputation and modelling methods in the analysis of a physical activity trial with missing outcomes. *International Journal of Epidemiology*. 2004;34(1):89-99. doi:10.1093/ije/dyh297.

800. Horton NJ, Kleinman KP. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*. 2007;61(1):79-90. doi:10.1198/000313007X172556.

801. Kenward MG, Carpenter JR. Multiple imputation: current perspectives. *Statistical Methods in Medical Research*. 2007;16(3):199-218. doi:10.1177/0962280206075304.

802. Janssen KJM, Donders ART, Harrell FE, et al. Missing covariate data in medical research: to impute is better than to ignore. *Journal of Clinical Epidemiology*. 2010;63(7):721-727. doi:10.1016/j.jclinepi.2009.12.008.

803. Dong Y, Peng C-YJ. Principled missing data methods for researchers. *SpringerPlus*. 2013;2(1). doi:10.1186/2193-1801-2-222.

804. Hussain JA, Bland M, Langan D, Johnson MJ, Currow DC, White IR. Quality of missing data reporting and handling in palliative care trials demonstrates that further development of the CONSORT statement is required: a systematic review. *Journal of Clinical Epidemiology*. 2017;88:81-91. doi:10.1016/j.jclinepi.2017.05.009.

805. Lang KM, Little TD. Principled Missing Data Treatments. *Prevention Science*. 2018;19(3):284-294. doi:10.1007/s11121-016-0644-5.

806. Nguyen CD, Strazdins L, Nicholson JM, Cooklin AR. Impact of missing data strategies in studies of parental employment and health: Missing items, missing waves, and missing mothers. *Social Science & Medicine*. 2018;209:160-168. doi:10.1016/j.socscimed.2018.03.009.

807. Lee KJ, Simpson JA. Introduction to multiple imputation for dealing with missing data. *Respirology*. 2014;19(2):162-167. doi:10.1111/resp.12226.

808. Hayati Rezvan P, Lee KJ, Simpson JA. The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. *BMC Medical Research Methodology*. 2015;15(1):30. doi:10.1186/s12874-015-0022-1.

809. Wood AM, White IR, Thompson SG. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clinical Trials*. 2004;1(4):368-376. doi:10.1191/1740774504cn032oa.

810. Harrell FE. *Regression Modeling Strategies*. Cham: Springer International Publishing; 2015.

811. Groenwold RHH, White IR, Donders ART, Carpenter JR, Altman DG, Moons KGM. Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. *CMAJ*. 2012;184(11):1265-1269. doi:10.1503/cmaj.110977.

812. White IR, Thompson SG. Adjusting for partially missing baseline measurements in randomized trials. *Statistics in Medicine*. 2005;24(7):993-1007. doi:10.1002/sim.1981.

813. Greenland S, Finkle WD. A Critical Look at Methods for Handling Missing Covariates in Epidemiologic Regression Analyses. *American Journal of Epidemiology*. 1995;142(12):1255-1264. doi:10.1093/oxfordjournals.aje.a117592.

814. Jones MP. Indicator and Stratification Methods for Missing Explanatory Variables in Multiple Linear Regression. *Journal of the American Statistical Association*. 1996;91(433):222. doi:10.2307/2291399.

815. Knol MJ, Janssen KJM, Donders ART, et al. Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example. *Journal of Clinical Epidemiology*. 2010;63(7):728-736. doi:10.1016/j.jclinepi.2009.08.028.

816. Lachin JM. Fallacies of last observation carried forward analyses. *Clinical Trials*. 2016;13(2):161-168. doi:10.1177/1740774515602688.

817. Cambridge English Dictionary. https://dictionary.cambridge.org/. Accessed December 1, 2018.

818. Rubin DB. *Multiple imputation for nonresponse in surveys*. New York, Chichester: Wiley; 1987.

819. Schafer JL, Olsen MK. Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective. *Multivariate Behavioral Research*. 1998;33(4):545-571. doi:10.1207/s15327906mbr3304_5.

820. Azur MJ, Stuart EA, Frangakis CE, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research*. 2011;20(1):40-49. doi:10.1002/mpr.329.

821. Liublinska V, Rubin DB. Re: "dealing with missing outcome data in randomized trials and observational studies". *American Journal of Epidemiology*. 2012;176(4):357-8; author reply 358-9. doi:10.1093/aje/kws215.

822. Nguyen CD, Carlin JB, Lee KJ. Model checking in multiple imputation: an overview and case study. *Emerging Themes in Epidemiology*. 2017;14(1):8. doi:10.1186/s12982-017-0062-6.

823. Murray JS. Multiple Imputation: A Review of Practical and Theoretical Findings. *Statistical Science*. 2018;33(2):142-159. doi:10.1214/18-STS644.

824. Sidi Y, Harel O. The treatment of incomplete data: Reporting, analysis, reproducibility, and replicability. *Social Science & Medicine*. 2018;209:169-173. doi:10.1016/j.socscimed.2018.05.037.

825. Little RJA, Rubin DB. Missing data in large data sets. In: Wright T, ed. *Statistical methods and the improvement of data quality: The proceedings of the Small Conference on the Improvement of the Quality of Data Collected by Data Collection Systems, November 11-12, 1982, Oak Ridge, Tennessee / edited by Tommy Wright*. Orlando, Fla.: Academic Press; 1983:215-243.

826. Doidge JC. Responsiveness-informed multiple imputation and inverse probability-weighting in cohort studies with missing data that are non-monotone or not missing at

random. *Statistical Methods in Medical Research*. 2018;27(2):352-363. doi:10.1177/0962280216628902.

827. Kaufman JS, Hernán MA. Epidemiologic Methods Are Useless: They Can Only Give You Answers. *Epidemiology*. 2012;23(6):785-786. doi:10.1097/EDE.0b013e31826c30e6.

828. Hernán MA, VanderWeele TJ. Compound Treatments and Transportability of Causal Inference. *Epidemiology*. 2011;22(3):368-377. doi:10.1097/EDE.0b013e3182109296.

829. Hernán MA, Robins JM. Per-Protocol Analyses of Pragmatic Trials. *New England Journal of Medicine*. 2017;377(14):1391-1398. doi:10.1056/NEJMsm1605385.

830. Hamar GB, Rula EY, Wells A, Coberley C, Pope JE, Larkin S. Impact of a Chronic Disease Management Program on Hospital Admissions and Readmissions in an Australian Population with Heart Disease or Diabetes. *Population Health Management*. 2012;16(2):121031074243002. doi:10.1089/pop.2012.0027.

831. Chon S, Lee YJ, Fraterrigo G, et al. Evaluation of Glycemic Variability in Well-Controlled Type 2 Diabetes Mellitus. *Diabetes Technology & Therapeutics*. 2013;15(6):455-460. doi:10.1089/dia.2012.0315.

832. National Heart Foundation of Australia. Measuring your blood pressure at home. https://www.heartfoundation.org.au/images/uploads/publications/Measuring-your-blood-pressure-at-home.PDF. Accessed November 27, 2017.

833. Gravel J, Opatrny L, Shapiro S. The intention-to-treat approach in randomized controlled trials: are authors saying what they do and doing what they say? *Clinical Trials*. 2007;4(4):350-356. doi:10.1177/1740774507081223.

834. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*. 2013;22(3):278-295. doi:10.1177/0962280210395740.

835. Rubin DB. Multiple Imputations in Sample Surveys-A Phenomenological Bayesian Approach to Nonresponse. *Proceedings of the Survey Research Methods Section, American Statistical Association*. 1978:20-34.

836. Rubin DB. Multiple Imputation After 18+ Years. *Journal of the American Statistical Association*. 1996;91(434):473. doi:10.2307/2291635.

837. Rubin DB, Schenker N. Multiple Imputation for Interval Estimation From Simple Random Samples With Ignorable Nonresponse. *Journal of the American Statistical Association*. 1986;81(394):366. doi:10.2307/2289225.

838. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*. 2011;30(4):377-399. doi:10.1002/sim.4067.

839. Sullivan TR, Lee KJ, Ryan P, Salter AB. Multiple imputation for handling missing outcome data when estimating the relative risk. *BMC Medical Research Methodology*. 2017;17(1):134. doi:10.1186/s12874-017-0414-5.

840. Abayomi K, Gelman A, Levy M. Diagnostics for multivariate imputations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. 2008;57(3):273-291. doi:10.1111/j.1467-9876.2007.00613.x.

841. Moons KGM, Donders RART, Stijnen T, Harrell FE. Using the outcome for imputation of missing predictor values was preferred. *Journal of Clinical Epidemiology*. 2006;59(10):1092-1101. doi:10.1016/j.jclinepi.2006.01.009.

842. Rodwell L, Lee KJ, Romaniuk H, Carlin JB. Comparison of methods for imputing limited-range variables: a simulation study. *BMC Medical Research Methodology*. 2014;14. doi:10.1186/1471-2288-14-57.

843. Sacks DB. A1C Versus Glucose Testing: A Comparison. *Diabetes Care*. 2011;34(2):518-523. doi:10.2337/dc10-1546.

844. Agoritsas T, Merglen A, Shah ND, O'Donnell M, Guyatt GH. Adjusted Analyses in Studies Addressing Therapy and Harm: Users' Guides to the Medical Literature. *JAMA*. 2017;317(7):748-759. doi:10.1001/jama.2016.20029.

845. Greenland S. A serious misinterpretation of a consistent inverse association of statin use with glioma across 3 case-control studies. *European Journal of Epidemiology*. 2017;32(1):87-88. doi:10.1007/s10654-016-0205-z.

846. Charnes A, Cooper WW, Mellon B. A Model for Programming and Sensitivity Analysis in an Integrated Oil Company. *Econometrica*. 1954;22(2):193. doi:10.2307/1907542.

847. Maffei RB. Mathematical Models, Values of Parameters, and the Sensitivity Analysis of Management-Decision Rules. *Journal of Marketing*. 1957;21(4):419. doi:10.2307/1247265.

848. Ding P, VanderWeele TJ. Sensitivity Analysis Without Assumptions. *Epidemiology*. 2016;27(3):368-377. doi:10.1097/EDE.0000000000000457.

849. Hu FB, Goldberg J, Hedeker D, Flay BR, Pentz MA. Comparison of Population-Averaged and Subject-Specific Approaches for Analyzing Repeated Binary Outcomes. *American Journal of Epidemiology*. 1998;147(7):694-703. doi:10.1093/oxfordjournals.aje.a009511.

850. Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. *International Journal of Epidemiology*. 2014;43(6):1969-1985. doi:10.1093/ije/dyu149.

851. Lash TL, Ahern TP. Bias Analysis to Guide New Data Collection. *International Journal of Biostatistics*. 2012;8(2):1-23. doi:10.2202/1557-4679.1345.

852. Groenwold RHH, Sterne JAC, Lawlor DA, Moons KGM, Hoes AW, Tilling K. Sensitivity analysis for the effects of multiple unmeasured confounders. *Annals of Epidemiology*. 2016;26(9):605-611. doi:10.1016/j.annepidem.2016.07.009.

853. Vadillo MA, Matute H, Blanco F. Fighting the Illusion of Control: How to Make Use of Cue Competition and Alternative Explanations. *Universitas Psychologica*. 2013;12(1):261-270.

854. MacFarlane D, Hurlstone MJ, Ecker UKH. Reducing demand for ineffective health remedies: overcoming the illusion of causality. *Psychology & Health*. 2018;33(12):1472-1489. doi:10.1080/08870446.2018.1508685.

855. Thomas L, Peterson ED. The Value of Statistical Analysis Plans in Observational Research. *JAMA*. 2012;308(8):773. doi:10.1001/jama.2012.9502.

856. Oestermeier U, Hesse FW. Verbal and visual causal arguments. *Cognition*. 2000;75(1):65-104. doi:10.1016/S0010-0277(00)00060-3.

857. Rosenbaum PR. The Consquences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment. *Journal of the Royal Statistical Society. Series A (General)*. 1984;147(5):656. doi:10.2307/2981697.

858. Berger VW. Valid Adjustment of Randomized Comparisons for Binary Covariates. *Biometrical Journal*. 2004;46(5):589-594. doi:10.1002/bimj.200410055.

859. Keiding N, Clayton D. Standardization and Control for Confounding in Observational Studies: A Historical Perspective. *Statistical Science*. 2014;29(4):529-558. doi:10.1214/13-STS453.

860. Robins JM. A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect. *Mathematical Modelling*. 1986;7(9-12):1393-1512. doi:10.1016/0270-0255(86)90088-6.

861. Taubman SL, Robins JM, Mittleman MA, Hernán MA. Intervening on risk factors for coronary heart disease: An application of the parametric g-formula. *International Journal of Epidemiology*. 2009;38(6):1599-1611. doi:10.1093/ije/dyp192.

862. Pearl J. Rejoinder to Discussions of Causal Diagrams for Empirical Research. *Biometrika*. 1995;82(4):702. doi:10.2307/2337339.

863. Robins JM. The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In: Sechrest L, Freeman H, Mulley A, eds. *Health Services Research Methodology: A Focus on AIDS*: NCHRS, U.S. Public Health Service; 1989:113-159.

864. Clare PJ, Dobbins TA, Mattick RP. Causal models adjusting for time-varying confounding-a systematic review of the literature. *International Journal of Epidemiology*. 2018;48(1):254-265. doi:10.1093/ije/dyy218.

865. Robins JM, Hernán MA. Estimation of the causal effects of time-varying exposures. In: Fitzmaurice GM, Davidian M, Verbeke G, Molenberghs G, eds. *Longitudinal Data Analysis*. Boca Raton: CRC Press; 2009. *Chapman & Hall/CRC Handbooks of Modern Statistical Methods*.

866. Naimi AI, Cole SR, Kennedy EH. An Introduction to G Methods. *International Journal of Epidemiology*. 2017;46(2):756-762. doi:10.1093/ije/dyw323.

867. Hernán MA, Brumback BA, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*. 2000;11(5):561-570.

868. Joffe MM. Structural Nested Models, G-Estimation, and the Healthy Worker Effect. *Epidemiology*. 2012;23(2):220-222. doi:10.1097/EDE.0b013e318245f798.

869. Hernán MA, Brumback BA, Robins JM. Estimating the causal effect of zidovudine on CD4 count with a marginal structural model for repeated measures. *Statistics in Medicine*. 2002;21(12):1689-1709. doi:10.1002/sim.1144.

870. Westreich DJ, Greenland S. The Table 2 Fallacy: Presenting and Interpreting Confounder and Modifier Coefficients. *American Journal of Epidemiology*. 2013;177(4):292-298. doi:10.1093/aje/kws412.

871. Daniel RM, Stavola BL de, Cousens SN. gformula : Estimating causal effects in the presence of time-dependent confounding or mediation. *The Stata Journal*. 2011;11(4):479-517.

872. Zeger SL, Irizarry RA, Peng RD. *On Time Series Analysis of Public Health and Biomedical Data*; 2004. Dept. of Biostatistics Working Papers. http://www.bepress.com/jhubiostat/paper54.

873. Box GEP, Jenkins GM, Reinsel GC, Ljung GM. *Time series analysis: Forecasting and control.* Fifth edition. Hoboken, New Jersey: Wiley; 2016.

874. Ettehad D, Emdin CA, Kiran A, et al. Blood pressure lowering for prevention of cardiovascular disease and death: a systematic review and meta-analysis. *The Lancet*. 2016;387(10022):957-967. doi:10.1016/S0140-6736(15)01225-8.

875. Pezzin LE, Feldman PH, Mongoven JM, McDonald MV, Gerber LM, Peng TR. Improving blood pressure control: results of home-based post-acute care interventions. *Journal of General Internal Medicine*. 2011;26(3):280-286. doi:10.1007/s11606-010-1525-4.

876. Myers MG, Stergiou GS. Reporting bias: Achilles' heel of home blood pressure monitoring. *Journal of the American Society of Hypertension*. 2014;8(5):350-357. doi:10.1016/j.jash.2014.02.001.

877. Clark CE, Horvath IA, Taylor RS, Campbell JL. Doctors record higher blood pressures than nurses: systematic review and meta-analysis. *British Journal of General Practice*. 2014;64(621):e223-32. doi:10.3399/bjgp14X677851.

878. Bengtsson U, Kjellgren K, Hallberg I, Lindwall M, Taft C. Improved Blood Pressure Control Using an Interactive Mobile Phone Support System. *The Journal of Clinical Hypertension*. 2016;18(2):101-108. doi:10.1111/jch.12682.

879. Sharman JE, Howes FS, Head GA, et al. Home blood pressure monitoring: Australian Expert Consensus Statement. *Journal of Hypertension*. 2015;33(9):1721-1728. doi:10.1097/HJH.0000000000000673.

880. Campbell L. White coat hyperglycaemia. *BMJ*. 1993;306:208.

881. Sullivan C, Chambers T, Goldie D, Gillett M, Woods A. White coat hyperglycaemia. *BMJ*. 1993;306:208.

882. Desborough JP. The stress response to trauma and surgery. *British Journal of Anaesthesia*. 2000;85(1):109-117. doi:10.1093/bja/85.1.109.

883. Logan RW, Young JG, Taubman SL, et al. *GFORMULA SAS macro*. Version: April 2016. http://www.hsph.harvard.edu/causal/software/.

884. Mellers B, Hertwig R, Kahneman D. Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science*. 2001;12(4):269-275. doi:10.1111/1467-9280.00350.

885. Bateman I, Kahneman D, Munro A, Starmer C, Sugden R. Testing competing models of loss aversion: an adversarial collaboration. *Journal of Public Economics*. 2005;89(8):1561-1580. doi:10.1016/j.jpubeco.2004.06.013.

Appendices

Appendix A. (Chapter 5)

A.1. Full PubMed search string

PubMed search terms found to increase the chance of retrieving studies that would meet the criteria while decreasing the chance of other studies/articles were identified through trial and error. The final query used in PubMed that returned an initial sample of 1,871 articles on 14 June 2015:

```
2014[dp] AND humans[mh]
AND
(cohort[tiab] OR cohorts[tiab] OR cohort studies[mh] OR cross-over studies[mh] OR
follow-up[tiab] OR follow-up studies[mh] OR followup[tiab] OR longitudinal[tiab]
OR observational studies[tiab] OR observational study[pt] OR observational
study[tiab])
AND
(before and after[tiab] OR comparative study[pt] OR compared[tiab] OR
comparison[tiab] OR comparative[tiab] OR versus[tiab])
AND
("Acta Derm Venereol"[ta] OR "Acta Neuropathol"[ta] OR "Addict Biol"[ta] OR
"Addiction"[ta] OR "Adv Drug Deliv Rev"[ta] OR "Adv Nutr"[ta] OR "Ageing Res
Rev"[ta] OR "Aging Cell"[ta] OR "AIDS"[ta] OR "Aliment Pharmacol Ther"[ta] OR
"Allergy"[ta] OR "Alzheimers Dement"[ta] OR "Am J Clin Nutr"[ta] OR "Am J
Gastroenterol"[ta] OR "Am J Kidney Dis"[ta] OR "Am J Ophthalmol"[ta] OR "Am J
Pathol"[ta] OR "Am J Physiol Lung Cell Mol Physiol"[ta] OR "Am J Psychiatry"[ta]
OR "Am J Respir Cell Mol Biol"[ta] OR "Am J Respir Crit Care Med"[ta] OR "Am J
Respir Crit Care Med"[ta] OR "Am J Sports Med"[ta] OR "Am J Surg Pathol"[ta] OR
"Am J Transplant"[ta] OR "Anesthesiology"[ta] OR "Angiogenesis"[ta] OR "Ann Emerg
Med"[ta] OR "Ann Fam Med"[ta] OR "Ann Intern Med"[ta] OR "Ann Neurol"[ta] OR "Ann
Rheum Dis"[ta] OR "Ann Surg"[ta] OR "Annu Rev Immunol"[ta] OR "Annu Rev Med"[ta]
OR "Annu Rev Nutr"[ta] OR "Annu Rev Pathol"[ta] OR "Annu Rev Pharmacol"[ta] OR OR
"Annu Rev Public Health"[ta] OR "Antioxid Redox Signal"[ta] OR "Arch Neurol"[ta]
OR "Arch Ophthalmol"[ta] OR "Arch Pediatr Adolesc Med"[ta] OR "Arch Toxicol"[ta]
OR "Arterioscler Thromb Vasc Biol"[ta] OR "Arthritis Care Res (Hoboken)"[ta] OR
"Arthritis Res Ther"[ta] OR "Arthritis Rheumatol"[ta] OR "Atheroscler Suppl"[ta]
OR "Biol Psychiatry"[ta] OR "Blood Rev"[ta] OR "Blood"[ta] OR "BMC Med"[ta] OR
"BMJ"[ta] OR "Br J Anaesth"[ta] OR "Br J Dermatol"[ta] OR "Br J Psychiatry"[ta] OR
"Br J Sports Med"[ta] OR "Br J Surg"[ta] OR "Brain"[ta] OR "Bull World Health
```

Organ"[ta] OR "CA Cancer J Clin"[ta] OR "Cancer Cell"[ta] OR "Cancer Discov"[ta] OR "Cancer Res"[ta] OR "Cell Metab"[ta] OR "Chest"[ta] OR "Circ Cardiovasc Imaging"[ta] OR "Circ Cardiovasc Interv"[ta] OR "Circ Res"[ta] OR "Circulation"[ta] OR "Clin Exp Allergy"[ta] OR "Clin Gastroenterol Hepatol"[ta] OR "Clin Infect Dis"[ta] OR "Clin J Am Soc Nephrol"[ta] OR "Clin Microbiol Infect"[ta] OR "Clin Pharmacol Ther"[ta] OR "Clin Rev Allergy Immunol"[ta] OR "Cochrane Database Syst Rev"[ta] OR "Cold Spring Harb Perspect Med"[ta] OR "Crit Care Med"[ta] OR "Crit Care"[ta] OR "Crit Rev Food Sci Nutr"[ta] OR "Crit Rev Toxicol"[ta] OR "Curr Opin Infect Dis"[ta] OR "Curr Opin Lipidol"[ta] OR "Curr Opin Nephrol Hypertens"[ta] OR "Curr Opin Rheumatol"[ta] OR "Dent Mater"[ta] OR "Diabetes Care"[ta] OR "Diabetes"[ta] OR "Dis Model Mech"[ta] OR "Drug Resist Updat"[ta] OR "EMBO Mol Med"[ta] OR "Emerg Infect Dis"[ta] OR "Endocr Rev"[ta] OR "Endoscopy"[ta] OR "Environ Health Perspect"[ta] OR "Epidemiol Rev"[ta] OR "Epidemiology"[ta] OR "Eur Heart J"[ta] OR "Eur J Epidemiol"[ta] OR "Eur J Heart Fail"[ta] OR "Eur J Nucl Med Mol Imaging"[ta] OR "Eur Respir J"[ta] OR "Eur Urol"[ta] OR "Euro Surveill"[ta] OR "Exerc Immunol Rev"[ta] OR "Exerc Sport Sci Rev"[ta] OR "Exp Dermatol"[ta] OR "Fertil Steril"[ta] OR "Forensic Toxicol"[ta] OR "Front Neuroendocrinol"[ta] OR "Gastroenterology"[ta] OR "Gut"[ta] OR "Haematologica"[ta] OR "Health Aff (Millwood)"[ta] OR "Health Technol Assess"[ta] OR "Hepatology"[ta] OR "Hum Brain Mapp"[ta] OR "Hum Reprod Update"[ta] OR "Hum Reprod"[ta] OR "Hypertension"[ta] OR "Immunity"[ta] OR "Immunol Rev"[ta] OR "Inflamm Bowel Dis"[ta] OR "Int J Epidemiol"[ta] OR "Int J Obes (Lond)"[ta] OR "Intensive Care Med"[ta] OR "J Acquir Immune Defic Syndr"[ta] OR "J Allergy Clin Immunol"[ta] OR "J Am Acad Child Psychiatry"[ta] OR "J Am Acad Dermatol"[ta] OR "J Am Coll Cardiol"[ta] OR "J Am Coll Surg"[ta] OR "J Am Geriatr Soc"[ta] OR "J Am Med Assoc"[ta] OR "J Am Med Dir Assoc"[ta] OR "J Am Soc Nephrol"[ta] OR "J Antimicrob Chemother"[ta] OR "J Bone Joint Surg Am"[ta] OR "J Bone Joint Surg Am"[ta] OR "J Cachexia Sarcopenia Muscle"[ta] OR "J Cardiovasc Magn Reson"[ta] OR "J Cereb Blood Flow Metab"[ta] OR "J Clin Epidemiol"[ta] OR "J Clin Invest"[ta] OR "J Clin Oncol"[ta] OR "J Dent Res"[ta] OR "J Exp Med"[ta] OR "J Gerontol A Biol Sci Med Sci"[ta] OR "J Heart Lung Transplant"[ta] OR "J Heart Lung Transplant"[ta] OR "J Hepatol"[ta] OR "J Infect Dis"[ta] OR "J Invest Dermatol"[ta] OR "J Med Internet Res"[ta] OR "J Med Internet Res"[ta] OR "J Natl Cancer Inst"[ta] OR "J Neurol Neurosurg Psychiatry"[ta] OR "J Neuropathol Exp Neurol"[ta] OR "J Nucl Med"[ta] OR "J Nutr Biochem"[ta] OR "J Pathol"[ta] OR "J Pineal Res"[ta] OR "J Psychiatry Neurosci"[ta] OR "J Thorac Oncol"[ta] OR "J Thromb Haemost"[ta] OR "J Toxicol Environ Health B Crit Rev"[ta] OR "JACC Cardiovasc Imaging"[ta] OR "JACC Cardiovasc Interv"[ta] OR "JAMA Dermatol"[ta] OR "JAMA Intern Med"[ta] OR "JAMA Psychiatry"[ta] OR "Kidney Int"[ta] OR "Lancet Infect Dis"[ta] OR "Lancet Neurol"[ta] OR "Lancet Oncol"[ta] OR "Lancet"[ta] OR "Leukemia"[ta] OR "Med Res Rev"[ta] OR "Med Sci Sports Exerc"[ta] OR "Milbank Q"[ta] OR "Mod Pathol"[ta] OR "Mol Aspects Med"[ta] OR "Mol Psychiatry"[ta] OR "Mutat Res Rev Mutat Res"[ta] OR "N Engl J Med"[ta] OR "Nanotoxicology"[ta] OR "Nat Immunol"[ta] OR "Nat Med"[ta] OR "Nat Rev Cancer"[ta] OR "Nat Rev Cardiol"[ta] OR "Nat Rev Clin Oncol"[ta] OR "Nat Rev Drug Discov"[ta] OR "Nat Rev Endocrinol"[ta] OR "Nat Rev Gastroenterol Hepatol"[ta] OR "Nat Rev Immunol"[ta] OR "Nat Rev Nephrol"[ta] OR "Nat Rev Neurol"[ta] OR "Nat Rev Rheumatol"[ta] OR "Nat Rev Urol"[ta] OR "Neurobiol Aging"[ta] OR "Neuroimage"[ta] OR "Neurology"[ta] OR "Neuropathol Appl Neurobiol"[ta] OR "Neuropsychopharmacology"[ta] OR "Neurorehabil Neural

```
Repair"[ta] OR "Neuroscientist"[ta] OR "Nutr Rev"[ta] OR "Obes Rev"[ta] OR
"Obesity (Silver Spring)"[ta] OR "Obstet Gynecol"[ta] OR "Ocul Surf"[ta] OR
"Ophthalmology"[ta] OR "Osteoarthritis Cartilage"[ta] OR "Pain"[ta] OR "Part Fibre
Toxicol"[ta] OR "Pediatrics"[ta] OR "Pharmacol Rev"[ta] OR "Pharmacol Ther"[ta] OR
"Pigment Cell Melanoma Res"[ta] OR "PLoS Med"[ta] OR "PLoS Negl Trop Dis"[ta] OR
"Proc Nutr Soc"[ta] OR "Prog Lipid Res"[ta] OR "Prog Retin Eye Res"[ta] OR
"Psychother Psychosom"[ta] OR "Radiology"[ta] OR "Radiother Oncol"[ta] OR
"Rheumatology"[ta] OR "Schizophr Bull"[ta] OR "Sci Transl Med"[ta] OR "Semin
Immunopathol"[ta] OR "Sleep Med Rev"[ta] OR "Sports Med"[ta] OR "Stem Cells
Dev"[ta] OR "Stem Cells"[ta] OR "Stroke"[ta] OR "Surg Obes Relat Dis"[ta] OR
"Theranostics"[ta] OR "Thorax"[ta] OR "Thromb Haemost"[ta] OR "Tob Control"[ta] OR
"Toxicol Sci"[ta] OR "Trends Endocrinol Metab"[ta] OR "Trends Immunol"[ta] OR
"Trends Mol Med"[ta] OR "Trends Pharmacol Sci"[ta] OR "Ultraschall Med"[ta] OR
"World Psychiatry"[ta])
NOT
(2013[ppdat] OR 2015[ppdat] OR case series[tiab] OR cross-sectional studies[mh] OR
diagnosis[sh] OR economics[sh] OR genetics[sh] OR meta-analysis[pt] OR
prevalence[mh] OR randomised[tiab] OR randomized[tiab] OR randomized controlled
trial[pt] OR randomly[tiab] OR review[pt] OR systematic[sb])
```

A.2. Regular expressions used for full-text search

For the detailed manual review, the 288 PDF articles were automatically searched for words or word combinations using the following regular expressions (regex) in the full-text search software program FileLocator Pro (https://www.mythicsoft.com/filelocatorpro). The same process was repeated for each of the unscreened years 2014-2017.

Program search settings:

Multi-line RegEx with Match case on; Match Across whole file with Allow wildcards ticked.

```
/*** Search PDF text for: "propensity score" ***/
Regex: ((p|P)ropensity((?:)|.)(s|S)core).*((References)|(REFERENCES)|(Reference
List)|(Competing interest))

/*** Search PDF text for: "propensity score" AND "balance" ***/
Regex:
((((p|P)ropensity((?:)|.)(s|S)core).*(\<(b|B)alance\>))|((\<(b|B)alance\>).*((p|P)ropensity((?:)
|.)(s|S)core))).*((References)|(REFERENCES)|(Reference List)|(Competing interest))

/*** Search PDF text for: "Hosmer-Lemeshow" ***/
Regex: (Hosmer((?:)|.)Lemeshow).*((References)|(REFERENCES)|(Reference List)|(Competing
interest))

/*** Search PDF text for: "propensity score" AND "Hosmer-Lemeshow" ***/
Regex:
```

```
((((p|P)ropensity((?:)|.)(s|S)core).*(Hosmer((?:)|.)Lemeshow))|((Hosmer((?:)|.)Lemeshow).*(
(p|P)ropensity((?:)|.)(s|S)core))).*((References)|(REFERENCES)|(Reference List)|(Competing
interest))

/*** Search PDF text for: " c statistic" ***/
Regex: (\<(c|C)(\s|-)(s|S)tatistic((?:)|s)\>).*((References)|(REFERENCES)|(Reference
List)|(Competing interest))

/*** Search PDF text for: "propensity score" AND "c statistic" ***/
Regex:
((((p|P)ropensity((?:)|.)(s|S)core).*(\<(c|C)(\s|-)(s|S)tatistic((?:)|s)\>))|((\<(c|C)(\s|-
)(s|S)tatistic((?:)|s)\>).*((p|P)ropensity((?:)|.)(s|S)cor
e))).*((References)|(REFERENCES)|(Reference List)|(Competing interest))

/*** Search PDF text for: "standardi(s|z)ed difference" ***/
Regex: ((s|S)tandardi(s|z)ed((?:)|.)(d|D)ifference).*((References)|(REFERENCES)|(Reference
List)|(Competing interest))

/*** Search PDF text for: "propensity score" AND "standardi(s|z)ed difference" ***/
Regex:
((((p|P)ropensity((?:)|.)(s|S)core).*((s|S)tandardi(s|z)ed((?:)|.)(d|D)ifference))|(((s|S)t
andardi(s|z)ed((?:)|.)(d|D)ifference).*((p|P)ropensity((?:
)|.)(s|S)core))).*((References)|(REFERENCES)|(Reference List)|(Competing interest))

/*** Search PDF text for: "propensity score matching/matched" ***/
Regex:
(((p|P)ropensity((?:)|.)(s|S)core)((?:)|.)((m|M)atch(ed|ing))).*((References)|(REFERENCES)|
(Reference List)|(Competing interest))

/*** Search PDF text for: "propensity score" AND "matching/matched" ***/
Regex:
((p|P)ropensity((?:)|.)(s|S)core).*((m|M)atch(ed|ing)).*((References)|(REFERENCES)|(Referen
ce List)|(Competing interest))

/*** Search PDF text for: "greedy matching" ***/
Regex:
((((g|G)reedy).*((m|M)atch(ed|ing)))|(((m|M)atch(ed|ing)).*((g|G)reedy))).*((References)|(R
EFERENCES)|(Reference List)|(Competing interest))

/*** Search PDF text for: "propensity score" AND "greedy matching" ***/
Regex:
((((p|P)ropensity((?:)|.)(s|S)core).*((((g|G)reedy).*((m|M)atch(ed|ing)))|(((m|M)atch(ed|in
g)).*((g|G)reedy))))|(((((g|G)reedy).*((m|M)atch(ed|ing)))
|(((m|M)atch(ed|ing)).*((g|G)reedy))).*((p|P)ropensity((?:)|.)(s|S)core))).*((References)|(
REFERENCES)|(Reference List)|(Competing interest))

/*** Search PDF text for: "nearest neighb" AND "matching" ***/
Regex:
((((n|N)earest((?:)|.)(n|N)eighb).*((m|M)atch(ed|ing)))|(((m|M)atch(ed|ing)).*((n|N)earest((?:)|
.)(n|N)eighb))).*((References)|(REFERENCES)|(Referenc       e List)|(Competing interest))

/*** Search PDF text for: "propensity score" AND "nearest neighb" ***/
Regex:
((((p|P)ropensity((?:)|.)(s|S)core).*((((n|N)earest((?:)|.)(n|N)eighb).*((m|M)atch(ed|ing))
)|(((m|M)atch(ed|ing)).*((n|N)earest((?:)|.)(n|N)eighb))))
|(((((n|N)earest((?:)|.)(n|N)eighb).*((m|M)atch(ed|ing)))|(((m|M)atch(ed|ing)).*((n|N)eares
t((?:)|.)(n|N)eighb))).*((p|P)ropensity((?:)|.)(s|S)core))
).*((References)|(REFERENCES)|(Reference List)|(Competing interest))

/*** Search PDF text for: "high dimensional" ***/
Regex: ((h|H)igh((?:)|.)(d|D)imensional).*((References)|(REFERENCES)|(Reference
List)|(Competing interest))

/*** Search PDF text for: "propensity score" AND "high dimensional" ***/
Regex:
((((p|P)ropensity((?:)|.)(s|S)core).*((h|H)igh((?:)|.)(d|D)imensional))|(((h|H)igh((?:)|.)(d|D)i
mensional).*((p|P)ropensity((?:)|.)(s|S)core))).*((Re ferences)|(REFERENCES)|(Reference
List)|(Competing interest))

/*** Search PDF text for: "inverse probability" AND "weights weighting" ***/
Regex:
((i|I)nverse((?:)|.)(p|P)robability).*((w|W)eight).*((References)|(REFERENCES)|(Reference
List)|(Competing interest))

/*** Search PDF text for: "propensity score" AND "inverse probability" ***/
Regex:
((((p|P)ropensity((?:)|.)(s|S)core).*((i|I)nverse((?:)|.)(p|P)robability).*((w|W)eight))|(((i|I)
nverse((?:)|.)(p|P)robability).*((w|W)eight).*((p|P)r
opensity((?:)|.)(s|S)core))).*((References)|(REFERENCES)|(Reference List)|(Competing
interest))

/*** Search PDF text for: "marginal structural model" ***/
Regex:
((m|M)arginal((?:)|.)(s|S)tructural((?:)|.)(m|M)odel).*((References)|(REFERENCES)|(Referenc
e List)|(Competing interest))

/*** Search PDF text for: " g formula" ***/
Regex: (\<(g|G)(\s|-)(f|F)ormula\>).*((References)|(REFERENCES)|(Reference List)|(Competing
interest))

/*** Search PDF text for: " g estimation" ***/
Regex: (\<(g|G)(\s|-)(e|E)stimation\>).*((References)|(REFERENCES)|(Reference
List)|(Competing interest))

/*** Search PDF text for: "doubly robust" ***/
Regex: ((d|D)oubly((?:)|.)(r|R)obust).*((References)|(REFERENCES)|(Reference
List)|(Competing interest))

/*** Search PDF text for: "directed acyclic graph" ***/
Regex:
((d|D)irected((?:)|.)(a|A)cyclic((?:)|.)(g|G)raph).*((References)|(REFERENCES)|(Reference
List)|(Competing interest))

/*** Search PDF text for: "instrumental variable" ***/
Regex: ((i|I)nstrumental((?:)|.)(v|V)ariable).*((References)|(REFERENCES)|(Reference
List)|(Competing interest))

/*** Search PDF text for: "latent class" ***/
Regex: ((l|L)atent((?:)|.)(c|C)lass).*((References)|(REFERENCES)|(Reference
List)|(Competing interest))

/*** Search PDF text for: "structural equation" ***/
Regex: ((s|S)tructural((?:)|.)(e|E)quation).*((References)|(REFERENCES)|(Reference
List)|(Competing interest))
```

```
/*** Search PDF text for: "multiple imputation" ***/
Regex: ((m|M)ultiple((?:)|.)(i|I)mputation).*((References)|(REFERENCES)|(Reference
List)|(Competing interest))

/*** Search PDF text for: "sensitivity analysis" ***/
Regex: ((s|S)ensitivity((?:)|.)(a|A)nalys(i|e)s).*((References)|(REFERENCES)|(Reference
List)|(Competing interest))

/*** Search PDF text for: "machine learning" ***/
Regex: ((m|M)achine((?:)|.)(l|L)earning).*((References)|(REFERENCES)|(Reference
List)|(Competing interest))

/*** Search PDF text for: "Bayesian" ***/
Regex: (Bayesian).*((References)|(REFERENCES)|(Reference List)|(Competing interest))

/*** Search PDF text for: "Bayesian Information Criterion" ***/
Regex:
(Bayesian((?:)|.)(i|I)nformation((?:)|.)(c|C)riterion).*((References)|(REFERENCES)|(Referen
ce List)|(Competing interest))

/*** Search PDF text for: "regression discontinuity" ***/
Regex: ((r|R)egression((?:)|.)(d|D)iscontinuity).*((References)|(REFERENCES)|(Reference
List)|(Competing interest))

/*** Search PDF text for: "difference in difference" ***/
Regex:
((d|D)ifference((?:)|.)(i|I)n((?:)|.)(d|D)ifference).*((References)|(REFERENCES)|(Reference
List)|(Competing interest))

/*** Search PDF text for: "stepwise" ***/
Regex: ((s|S)tepwise).*((References)|(REFERENCES)|(Reference List)|(Competing interest))

/*** Search PDF text for: "SAS" (Note: used " SAS" in full-text review with 104/104) */
Regex: (\<SAS\>).*((References)|(REFERENCES)|(Reference List)|(Competing interest))

/*** Search PDF text for: "Stata" ***/
Regex: (\<Stata\>).*((References)|(REFERENCES)|(Reference List)|(Competing interest))

/*** Search PDF text for: " SPSS" ***/
Regex: (\<SPSS\>).*((References)|(REFERENCES)|(Reference List)|(Competing interest))

/*** Search PDF text for: "R" ***/
Regex: (\<R(\s|-)(((s|S)oftware)|((v|V)ersion))).*((References)|(REFERENCES)|(Reference
List)|(Competing interest))

/* found 53 files */
/*** Search PDF text for: "R" ***/
Regex: (\<R(\s|-
)(((s|S)oftware)|((v|V)ersion)|((p|P)ackage)|((s|T)atistic(s|al))|(Foundation)|(3.1))).*((R
eferences)|(REFERENCES)|(Reference List)|(Competing interest))

/*** Search PDF text for: "statistician" ***/
Regex: (\<(s|S)tatistician\>).*((References)|(REFERENCES)|(Reference List)|(Competing
interest))

/*** Search PDF text for: "alternative explanation" ***/
```

```
Regex: ((Discussion)|(DISCUSSION)).*(alternative
explanation).*((References)|(REFERENCES)|(Reference List)|(Competing interest))

/*** Search PDF text for: "bias analysis" ***/
Regex: ((b|B)ias((?:)|.)(a|A)nalys(i|e)s).*((References)|(REFERENCES)|(Reference
List)|(Competing interest))

/*** Search PDF text for: "bias list" ***/
Regex:
((((b|B)ias).{1,10}(\<(l|L)ist))|(\<((l|L)ist).{1,10}((b|B)ias))).*((References)|(REFERENCE
S)|(Reference List)|(Competing interest))

/*** Search PDF text for: "quantitative/probabilistic bias/sensitivity analys(i/e)s"*/
Regex:
((((q|Q)uantitative)|((p|P)robabilistic))((?:)|.)((((b|B)ias)|((s|S)ensitivity))((?:)|.)(a|A)nal
ys(i|e)s)).*((References)|(REFERENCES)|(Reference List)|(Competing interest))

/*** Search PDF text for: "Significant or Significantly" ***/
Regex: (\<(s|S)ignificant\>).*((References)|(REFERENCES)|(Reference List)|(Competing
interest))
```

A.3. Distinct statistical methods in some groups

Table 1. Distinct methods extracted and grouped as 'Any multivariable regression'

Articles: 257

Distinct methods recorded	References
accelerated failure time model	1
Andersen-Gill repeated-event model with robust variance	1
ANOVA	1
ARIMA regression model	1
binomial regression	3
competing risks regression model	11
competing risks regression, Fine and Gray method	6
Cox proportional hazards model	109
Cox proportional hazards model, time-varying	1
Cox proportional hazards model, weighted	3
Cox regression analysis stratified for matched pairs	10
Cox regression model with non-proportional hazards	7
Cox regression with heavyside functions	1
Cox regression, conditional	1
cumulative logit regression model	1
exact logistic regression	1
fixed-effects model	2

generalized additive model	1
generalized estimating equations	7
generalized estimating equations with an independent correlation matrix	5
generalized estimating equations with logit link	3
generalized least squares for serially correlated continuous data	1
generalized linear mixed model (GLMM)	2
generalized linear mixed model with log link	2
generalized linear model	2
generalized linear model with a log link function	1
generalized linear model with a logit link	1
generalized linear model with log link and gamma distribution	1
interrupted time-series model	2
joint model for longitudinal and survival data	1
linear mixed-effects model	12
linear regression	17
log-binomial logistic regression	1
log-binomial model	2
logistic regression	130
logistic regression, conditional	6
marginal structural Cox model	3
marginal structural model	2
mixed-effects Cox regression model	2
mixed-effects linear regression model	1
mixed-effects logistic regression model	6
mixed-effects model	3
mixed-effects pattern mixture model	1
multilevel Poisson regression model	1
multilevel random-effects logistic regression model	2
multilevel random-effects Poisson regression model	2
multinomial logit regression	5
multi-state model	1
negative binomial regression	7
Poisson generalized estimating-equation model	1
Poisson regression	14
Poisson regression model with Pearson adjustment for overdispersion	1
pooled logistic model	2
propensity score analysis using stratification	1
propensity score estimation using boosted regression trees	1
proportional odds logistic regression	1
proportional piecewise exponential survival model	1
random-effects model	2
zero-inflated negative binomial model	1

Table 2. Distinct methods extracted and grouped as ‘Multivariable regression NOT used’

Articles: 31

Distinct methods recorded	References
binomial test	1
Byar approximation to exact results based on the Poisson distribution	1
case-coverage method	1
chi-squared test	12
Cox regression analysis stratified for matched pairs*	1
crude odds ratio calculation	1
descriptive statistics only	1
Fisher’s exact test	8
Kaplan-Meier method with log-rank test	4
Kruskal-Wallis test	1
logistic regression*	2
Mann-Whitney U (Wilcoxon rank-sum) test	5
Mantel–Haenszel test	1
on-treatment analysis	1
Poisson regression, conditional*	1
standardized incidence ratio (SIR)	3
stratified analysis	2
Student's t-test	9
two-sample Z-test	1

* with a single explanatory variable

Table 3. Distinct methods extracted and grouped as ‘Propensity score (PS) methods’

Articles: 94

Distinct methods recorded	References
propensity score analysis using inverse probability of treatment weighting (IPTW)	14
propensity score analysis using stratification	9
propensity score analysis with bipartite weighting	1
propensity score as covariate	25
propensity score calculation using a high-dimensional propensity score	3
propensity score calculation with custom method	3
propensity score calculation, bivariate	1
propensity score estimation using boosted regression trees	1
propensity score for comparison of groups only	1
propensity score matching	54
propensity score matching of triads	1
propensity score matching using a greedy matching algorithm	20
propensity score matching using nearest neighbour matching	16
propensity score matching using Rubin’s Rules	2

Table 4. Multivariable methods used in articles using 'Propensity score (PS) methods'

Articles: 94

Distinct methods recorded	References
Andersen-Gill repeated-event model with robust variance	1
ANOVA	1
competing risks regression model	3
competing risks regression, Fine and Gray method	3
Cox proportional hazards model	42
Cox proportional hazards model, time-varying	1
Cox proportional hazards model, weighted	2
Cox regression analysis stratified for matched pairs	9
Cox regression model with non-proportional hazards	2
Cox regression with heavyside functions	1
cumulative logit regression model	1
generalized estimating equations	1
generalized estimating equations with an independent correlation matrix	3
generalized estimating equations with logit link	1
generalized linear mixed model (GLMM)	1
generalized linear mixed model with log link	1
generalized linear model with a log link function	1
generalized linear model with log link and gamma distribution	1
joint model for longitudinal and survival data	1
linear mixed-effects model	2
linear regression	5
logistic regression	70
logistic regression, conditional	5
marginal structural Cox model	2
mixed-effects Cox regression model	1
mixed-effects logistic regression model	1
mixed-effects model	1
multinomial logit regression	3
negative binomial regression	3
Poisson regression	5
Poisson regression model with Pearson adjustment for overdispersion	1
propensity score analysis using stratification	1
propensity score estimation using boosted regression trees	1

Table 5. Multivariable methods used in articles NOT using 'Propensity score (PS) methods'

Articles: 163

Distinct methods recorded	References
accelerated failure time model	1
ARIMA regression model	1
binomial regression	3

competing risks regression model	8
competing risks regression, Fine and Gray method	3
Cox proportional hazards model	67
Cox proportional hazards model, weighted	1
Cox regression analysis stratified for matched pairs	1
Cox regression model with non-proportional hazards	5
Cox regression, conditional	1
exact logistic regression	1
fixed-effects model	2
generalized additive model	1
generalized estimating equations	6
generalized estimating equations with an independent correlation matrix	2
generalized estimating equations with logit link	2
generalized least squares for serially correlated continuous data	1
generalized linear mixed model (GLMM)	1
generalized linear mixed model with log link	1
generalized linear model	2
generalized linear model with a logit link	1
interrupted time-series model	2
linear mixed-effects model	10
linear regression	11
log-binomial logistic regression	1
log-binomial model	2
logistic regression	60
marginal structural Cox model	1
marginal structural model	2
mixed-effects Cox regression model	1
mixed-effects linear regression model	1
mixed-effects logistic regression model	5
mixed-effects model	2
mixed-effects pattern mixture model	1
multilevel Poisson regression model	1
multilevel random-effects logistic regression model	2
multilevel random-effects Poisson regression model	2
multi-state model	1
negative binomial regression	3
Poisson generalized estimating-equation model	1
Poisson regression	9
pooled logistic model	1
proportional odds logistic regression	1
proportional piecewise exponential survival model	1
random-effects model	2
zero-inflated negative binomial model	1

Table 6. Multivariable methods in articles that claimed to do a ‘Sensitivity analysis’

Articles: 125; Note that all 3 articles that claimed to conduct a ‘sensitivity analysis’ yet did not use a multivariable method were vaccine studies

Distinct methods recorded	References
accelerated failure time model	1
Andersen-Gill repeated-event model with robust variance	1
competing risks regression model	2
competing risks regression, Fine and Gray method	5
Cox proportional hazards model	58
Cox proportional hazards model, time-varying	1
Cox proportional hazards model, weighted	2
Cox regression analysis stratified for matched pairs	7
Cox regression model with non-proportional hazards	5
Cox regression, conditional	1
cumulative logit regression model	1
custom matching procedure	2
fixed-effects model	2
generalized estimating equations	4
generalized estimating equations with an independent correlation matrix	5
generalized estimating equations with logit link	2
generalized linear mixed model (GLMM)	2
generalized linear mixed model with log link	2
generalized linear model	2
generalized linear model with a log link function	1
generalized linear model with a logit link	1
generalized linear model with log link and gamma distribution	1
interrupted time-series model	1
joint model for longitudinal and survival data	1
linear mixed-effects model	6
linear regression	5
log-binomial logistic regression	1
log-binomial model	2
logistic regression	63
logistic regression, conditional	4
marginal structural Cox model	3
marginal structural model	2
mixed-effects Cox regression model	1
mixed-effects linear regression model	1
mixed-effects logistic regression model	4
mixed-effects model	2
mixed-effects pattern mixture model	1
multilevel random-effects logistic regression model	1
multilevel random-effects Poisson regression model	1
multinomial logit regression	4
negative binomial regression	4
Poisson generalized estimating-equation model	1

Poisson regression	5
Poisson regression model with Pearson adjustment for overdispersion	1
pooled logistic model	2
proportional piecewise exponential survival model	1
random-effects model	2

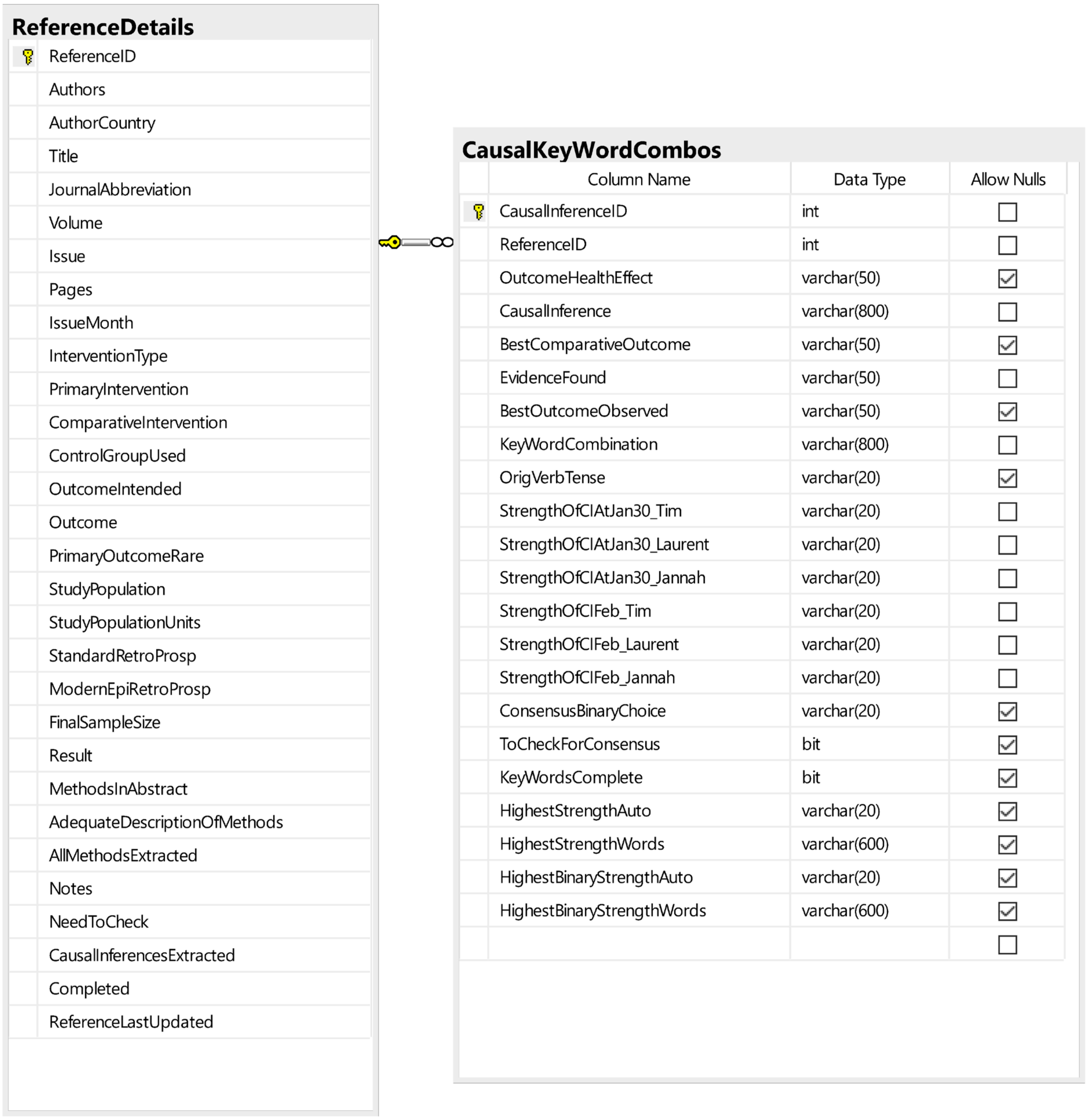
A.4. Software use in articles by journal category

Table 7. Software use in articles by journal category

Journal Category	SAS	SPSS	Stata	R	Other	Not spec.	Total
Cardiovascular	11 (55%)	1 (5%)	2 (10%)	2 (10%)	2 (10%)	2 (10%)	20
Critical Care Medicine	7 (39%)	7 (39%)	2 (11%)	0 (0%)	1 (6%)	1 (6%)	18
Gastroenterol. & Hep.	5 (21%)	8 (33%)	4 (17%)	5 (21%)	1 (4%)	1 (4%)	24
Infectious Diseases	6 (33%)	3 (17%)	3 (17%)	2 (11%)	2 (11%)	2 (11%)	18
Gen. & Internal Med.	17 (47%)	2 (6%)	4 (11%)	5 (14%)	1 (3%)	7 (19%)	36
Obstetrics & Gynec.	6 (16%)	15 (39%)	9 (24%)	1 (3%)	4 (11%)	3 (8%)	38
Other categories	47 (36%)	17 (13%)	22 (17%)	15 (11%)	7 (5%)	23 (18%)	131
Peripheral Vascular	11 (52%)	3 (14%)	2 (10%)	0 (0%)	2 (10%)	3 (14%)	21
Surgery	9 (20%)	13 (30%)	9 (20%)	3 (7%)	3 (7%)	7 (16%)	44
Urology & Nephrol.	10 (56%)	3 (17%)	3 (17%)	2 (11%)	0 (0%)	0 (0%)	18

Appendix B. (Chapter 6)

B.1. Database tables

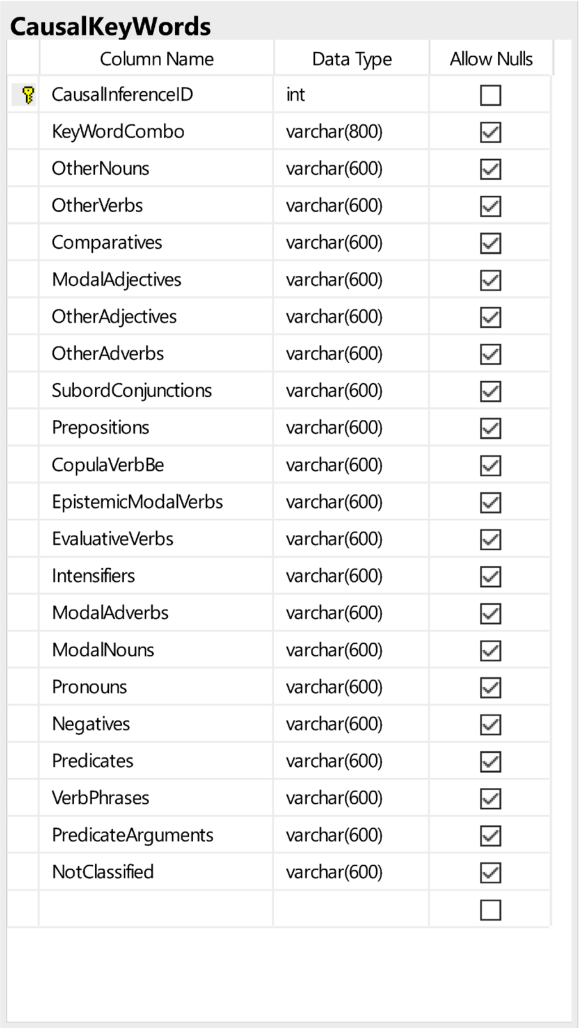


ReferenceDetails

🔑	ReferenceID
	Authors
	AuthorCountry
	Title
	JournalAbbreviation
	Volume
	Issue
	Pages
	IssueMonth
	InterventionType
	PrimaryIntervention
	ComparativeIntervention
	ControlGroupUsed
	OutcomeIntended
	Outcome
	PrimaryOutcomeRare
	StudyPopulation
	StudyPopulationUnits
	StandardRetroProsp
	ModernEpiRetroProsp
	FinalSampleSize
	Result
	MethodsInAbstract
	AdequateDescriptionOfMethods
	AllMethodsExtracted
	Notes
	NeedToCheck
	CausalInferencesExtracted
	Completed
	ReferenceLastUpdated

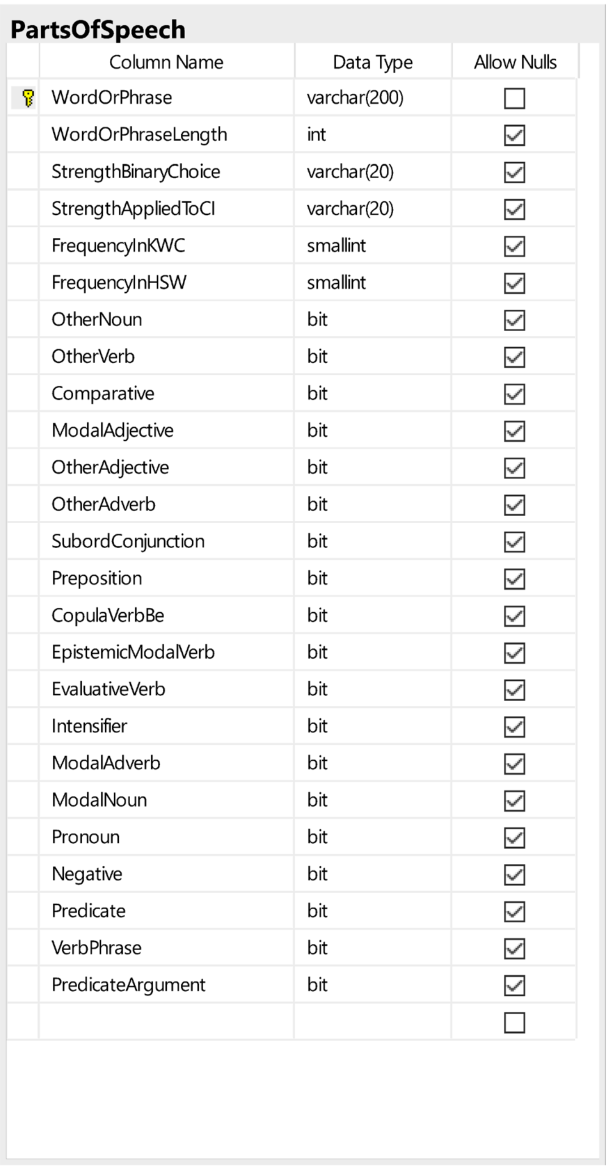
CausalKeyWordCombos

	Column Name	Data Type	Allow Nulls
🔑	CausalInferenceID	int	□
	ReferenceID	int	□
	OutcomeHealthEffect	varchar(50)	☑
	CausalInference	varchar(800)	□
	BestComparativeOutcome	varchar(50)	☑
	EvidenceFound	varchar(50)	□
	BestOutcomeObserved	varchar(50)	☑
	KeyWordCombination	varchar(800)	□
	OrigVerbTense	varchar(20)	☑
	StrengthOfCIAtJan30_Tim	varchar(20)	□
	StrengthOfCIAtJan30_Laurent	varchar(20)	□
	StrengthOfCIAtJan30_Jannah	varchar(20)	□
	StrengthOfCIFeb_Tim	varchar(20)	□
	StrengthOfCIFeb_Laurent	varchar(20)	□
	StrengthOfCIFeb_Jannah	varchar(20)	□
	ConsensusBinaryChoice	varchar(20)	☑
	ToCheckForConsensus	bit	☑
	KeyWordsComplete	bit	☑
	HighestStrengthAuto	varchar(20)	☑
	HighestStrengthWords	varchar(600)	☑
	HighestBinaryStrengthAuto	varchar(20)	☑
	HighestBinaryStrengthWords	varchar(600)	☑
			□



CausalKeyWords

	Column Name	Data Type	Allow Nulls
🔑	CausalInferenceID	int	□
	KeyWordCombo	varchar(800)	☑
	OtherNouns	varchar(600)	☑
	OtherVerbs	varchar(600)	☑
	Comparatives	varchar(600)	☑
	ModalAdjectives	varchar(600)	☑
	OtherAdjectives	varchar(600)	☑
	OtherAdverbs	varchar(600)	☑
	SubordConjunctions	varchar(600)	☑
	Prepositions	varchar(600)	☑
	CopulaVerbBe	varchar(600)	☑
	EpistemicModalVerbs	varchar(600)	☑
	EvaluativeVerbs	varchar(600)	☑
	Intensifiers	varchar(600)	☑
	ModalAdverbs	varchar(600)	☑
	ModalNouns	varchar(600)	☑
	Pronouns	varchar(600)	☑
	Negatives	varchar(600)	☑
	Predicates	varchar(600)	☑
	VerbPhrases	varchar(600)	☑
	PredicateArguments	varchar(600)	☑
	NotClassified	varchar(600)	☑
			□



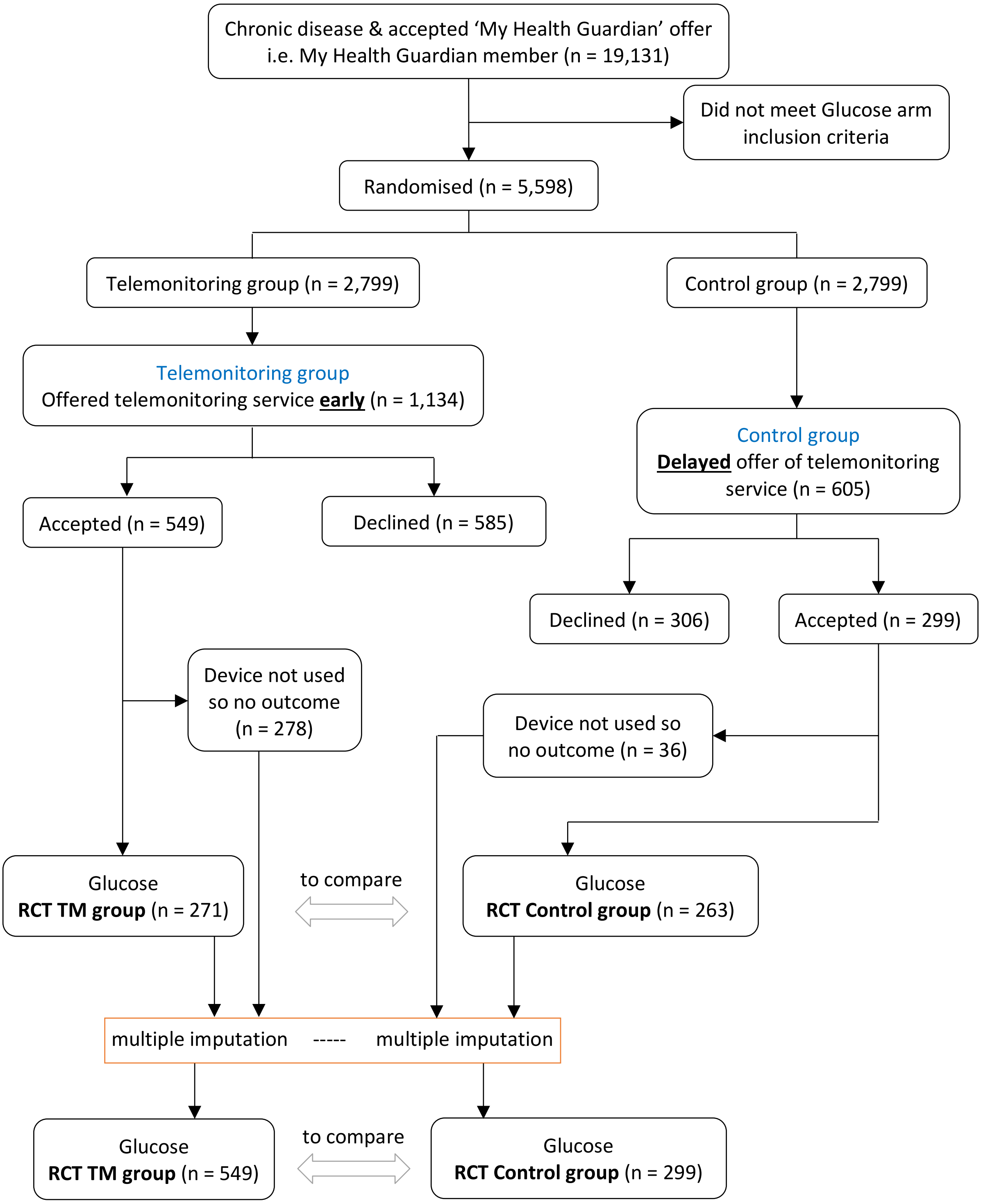
PartsOfSpeech

	Column Name	Data Type	Allow Nulls
🔑	WordOrPhrase	varchar(200)	□
	WordOrPhraseLength	int	☑
	StrengthBinaryChoice	varchar(20)	☑
	StrengthAppliedToCI	varchar(20)	☑
	FrequencyInKWC	smallint	☑
	FrequencyInHSW	smallint	☑
	OtherNoun	bit	☑
	OtherVerb	bit	☑
	Comparative	bit	☑
	ModalAdjective	bit	☑
	OtherAdjective	bit	☑
	OtherAdverb	bit	☑
	SubordConjunction	bit	☑
	Preposition	bit	☑
	CopulaVerbBe	bit	☑
	EpistemicModalVerb	bit	☑
	EvaluativeVerb	bit	☑
	Intensifier	bit	☑
	ModalAdverb	bit	☑
	ModalNoun	bit	☑
	Pronoun	bit	☑
	Negative	bit	☑
	Predicate	bit	☑
	VerbPhrase	bit	☑
	PredicateArgument	bit	☑
			□

Appendix C. (Chapter 7)

C.1. HCF Case Study

Figure 1. Glucose arm participant flow diagram



Chronic disease & accepted 'My Health Guardian' offer i.e. My Health Guardian member (n = 19,131)
Did not meet Glucose arm inclusion criteria
Randomised (n = 5,598)
Telemonitoring group (n = 2,799)
Control group (n = 2,799)
Telemonitoring group
Offered telemonitoring service **early** (n = 1,134)
Control group
Delayed offer of telemonitoring service (n = 605)
Accepted (n = 549)
Declined (n = 585)
Declined (n = 306)
Accepted (n = 299)
Device not used so no outcome (n = 278)
Device not used so no outcome (n = 36)
Glucose
RCT TM group (n = 271)
to compare
Glucose
RCT Control group (n = 263)
multiple imputation ----- multiple imputation
Glucose
RCT TM group (n = 549)
to compare
Glucose
RCT Control group (n = 299)

Figure 2. BP arm participant flow diagram

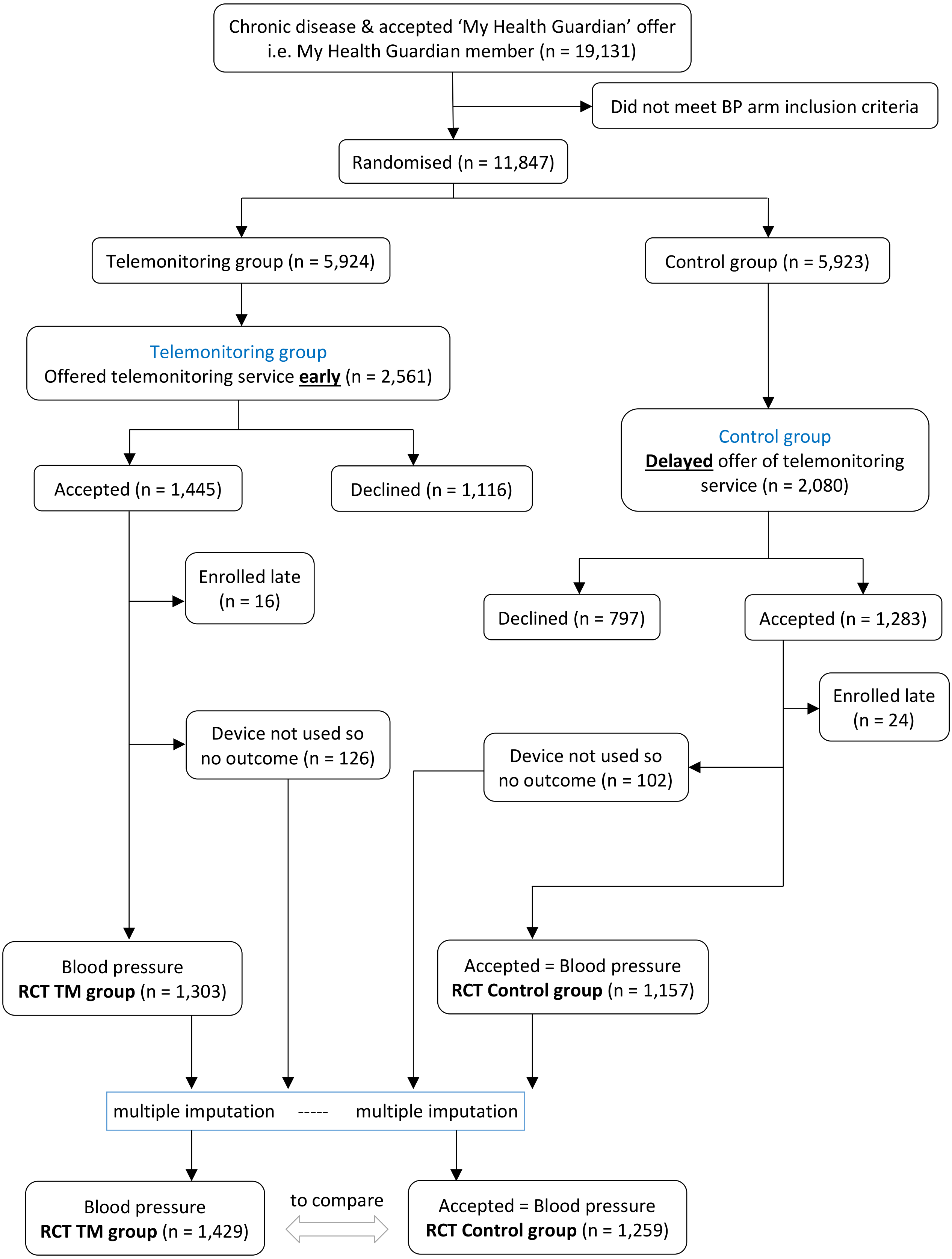

Chronic disease & accepted 'My Health Guardian' offer
i.e. My Health Guardian member (n = 19,131)
Did not meet BP arm inclusion criteria
Randomised (n = 11,847)
Telemonitoring group (n = 5,924)
Control group (n = 5,923)
Telemonitoring group
Offered telemonitoring service **early** (n = 2,561)
Control group
Delayed offer of telemonitoring service (n = 2,080)
Accepted (n = 1,445)
Declined (n = 1,116)
Enrolled late (n = 16)
Declined (n = 797)
Accepted (n = 1,283)
Enrolled late (n = 24)
Device not used so no outcome (n = 126)
Device not used so no outcome (n = 102)
Blood pressure **RCT TM group** (n = 1,303)
Accepted = Blood pressure **RCT Control group** (n = 1,157)
multiple imputation ----- multiple imputation
Blood pressure **RCT TM group** (n = 1,429)
to compare
Accepted = Blood pressure **RCT Control group** (n = 1,259)

Table 8. Glucose arm participant enrolments and use of glucometer each month

	2014						2015											
	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul†	Aug	Sep	Oct	Nov	Dec
Telemonitoring	Telemonitoring group							Total enrolled = 549										
Enrolled July	290	284	265	258	247	234	224	221	216	214	209	205	191	186	172	171	161	157
used glucometer	116	202	190	190	182	167	164	162	149	149	147	137	141	134	130	125	114	110
Enrolled August		92	88	82	77	70	66	66	64	63	60	58	55	49	48	46	44	42
used glucometer		24	55	50	56	48	43	46	42	40	37	34	36	36	31	30	27	27
Enrolled September			104	101	96	90	86	85	81	80	78	76	72	70	69	66	63	62
used glucometer			44	62	63	56	48	45	50	50	49	48	47	40	40	38	41	38
Enrolled October				43	42	41	39	39	38	37	37	35	31	28	27	25	24	24
used glucometer				27	28	28	24	23	23	22	22	20	16	15	15	15	14	12
Enrolled November					3	3	3	3	2	2	2	2	2	2	2	2	2	2
used glucometer					0	0	0	1	1	2	2	2	2	2	2	2	2	2
Enrolled January							5	5	5	5	5	5	5	5	5	4	4	4
used glucometer							1	4	3	3	4	4	5	5	5	4	4	3
Enrolled February								12	11	10	9	9	9	9	9	9	9	8
used glucometer								6	7	7	7	7	7	6	6	6	5	5
Total enrolled	290	376	457	484	465	438	423	431	417	411	400	390	365	349	332	323	307	299
used glucometer	116	226	289	329	329	299	280	287	275	273	268	252	254	238	229	220	207	197
% enrolled used gluco	40%	60%	63%	68%	71%	68%	66%	67%	66%	66%	67%	65%	70%	68%	69%	68%	67%	66%
% group total used gluco	21%	41%	53%	60%	60%	54%	51%	52%	50%	50%	49%	46%	46%	43%	42%	40%	38%	36%
Controls	Control group							Total enrolled = 299										
Enrolled July													152	152	150	146	137	133
used glucometer													27	95	110	105	102	83
Enrolled August														57	57	57	53	49
used glucometer														11	40	43	36	29
Enrolled September															69	69	65	65
used glucometer															11	43	55	48
Enrolled October																19	19	18
used glucometer																9	15	17
Enrolled November																	2	2
used glucometer																	1	2
Total enrolled													152	209	276	291	276	267
used glucometer													27	106	161	200	209	179
% enrolled used gluco													18%	51%	58%	69%	76%	67%
% group total used gluco													9%	35%	54%	67%	70%	60%

† the months highlighted in yellow were used to define the mean blood glucose primary outcome

Table 9. Glucose arm baseline characteristics before multiple imputation

For participants with ≥1 home blood glucose measurement from 1 July to 30 Nov 2015

Baseline characteristics	Telemonitoring N = 271 (49% of 549)	Controls N = 263 (88% of 299)	P-value
Sex			
Male	169 (62%)	171 (65%)	0.530
Female	102 (38%)	92 (35%)	
Age (years)			
Mean (SD)	68.8 (9.2)	65.7 (11.1)	0.001
Median (IQR)	69 (12)	67 (15)	
Ethnicity			
Missing (%)	40 (15%)	76 (29%)	
Caucasian	202 (87%)	165 (88%)	0.267
Asian	13 (6%)	12 (6%)	
Other	16 (7%)	10 (5%)	
HbA1c (DCCT %) (last from Jul13-Jun14)			
Missing (%)	191 (70%)	200 (76%)	
Mean (SD)	6.7 (1.2)	6.8 (1.2)	0.944
Median (IQR)	6.6 (1.2)	6.5 (1.4)	
BMI (last weight from Jul13-Jun14)			
Missing (%)	82 (30%)	82 (31%)	
Mean (SD)	30.5 (5.6)	30.4 (5.4)	0.838
Median (IQR)	29.6 (7.8)	29.9 (6.5)	
Diabetes type			
Type 1	8 (3%)	17 (6.5%)	0.074
Type 2	248 (92%)	237 (90.5%)	
Other/unspecified	13 (5%)	8 (3%)	
Hypertension	157 (58%)	57 (22%)	< .0001
Hyperlipidemia	80 (30%)	56 (22%)	0.037
Cardiovascular disease	145 (54%)	107 (41%)	0.003
Arthritis (any type)	131 (48%)	100 (38%)	0.018
Back pain (related diagnosis)	55 (20%)	58 (22%)	0.672
Walking pain (related diagnosis)	48 (18%)	36 (14%)	0.235
Eye problem (related diagnosis)	34 (13%)	27 (10%)	0.418
Insulin or Analogue	45 (17%)	41 (16%)	0.814
Pain relief drug	155 (57%)	122 (46%)	0.015
Number of Type 2 diabetes drugs			
0 drugs prescribed	71 (26%)	92 (35%)	0.263
1 drugs prescribed	127 (47%)	113 (43%)	
2 drugs prescribed	62 (23%)	48 (18%)	
3 drugs prescribed	9 (3%)	8 (3%)	
4 drugs prescribed	2 (1%)	2 (1%)	

Table 9. cont. Glucose arm baseline characteristics before multiple imputation

Baseline characteristics	TM N = 271 (49% of 549)	Controls N = 263 (88% of 299)	P-value
Employment status (before Jul14)			
Missing (%)	223 (82%)	210 (80%)	
Full-time	4 (8%)	4 (8%)	0.734
Part-time	2 (4%)	6 (11%)	
Self-employed	2 (4%)	2 (4%)	
No employment	15 (31%)	13 (25%)	
Retired	25 (52%)	28 (53%)	
Moderate exercise			
Missing (%)	230 (85%)	238 (90%)	
Yes (before Jul 2014)	9 (22%)	4 (16%)	0.752
Smoking status			
Missing (%)	120 (44%)	118 (45%)	
Never smoker	88 (58%)	89 (61%)	0.860
Past smoker	56 (37%)	50 (34%)	
Current smoker	7 (5%)	6 (4%)	
Risk level (last from Jul13-Jun14)			
Extreme Risk	11 (4%)	11 (4%)	0.011
High Risk	63 (23%)	49 (19%)	
Medium Risk	17 (6%)	10 (4%)	
Low Risk	100 (37%)	77 (29%)	
Self-Care	80 (30%)	116 (44%)	

Table 10. BP arm baseline characteristics before multiple imputation

Baseline characteristics	Analyses 5 & 6			Analysis 7		
	TM N = 1,429	Controls N = 1,259	P	TM N = 773	Controls N = 617	P
Sex						
Male	727 (51%)	661 (52%)	0.40	426 (55%)	370 (60%)	0.07
Female	702 (49%)	598 (48%)		347 (45%)	247 (40%)	
Age (years)						
Mean (SD)	70.6 (9.9)	69.1 (9.5)	<.0001	70.6 (9.1)	69.4 (9.0)	0.01
Median (IQR)	72 (13)	70 (13)		72 (12)	71 (12)	
Ethnicity						
Missing (%)	298 (21%)	365 (29%)		143 (19%)	148 (24%)	
Caucasian	1,036 (73%)	809 (64%)	0.58	577 (75%)	424 (69%)	0.23
Asian	34 (2%)	27 (2%)		20 (3%)	13 (2%)	
Other	61 (4%)	58 (5%)		33 (4%)	32 (5%)	
BMI (last weight Jul13-Jun14)						
Missing (%)	549 (38%)	483 (38%)		283 (37%)	213 (35%)	
Mean (SD)	29.4 (6.3)	29.3 (5.3)	0.74	29.2 (5.9)	28.8 (4.5)	0.38
Median (IQR)	28.6 (7.1)	28.7 (6.0)		28.4 (6.6)	28.6 (5.2)	
Diabetes type						
Type 1	7 (0.5%)	2 (0.2%)	0.009	3 (0.4%)	1 (0.2%)	0.04
Type 2	139 (10%)	145 (12%)		70 (9%)	46 (7%)	
Other/unspecified	18 (1%)	4 (0.3%)		11 (1%)	1 (0.2%)	
No diabetes	1,265 (89%)	1,108 (88%)		689 (89%)	569 (92%)	
Systolic BP (last Jul13-Jun14)						
Missing (%)	659 (46%)	602 (48%)		321 (42%)	260 (42%)	
Mean (SD)	132.6 (13.7)	132.2 (13.2)	0.57	132.3 (13.4)	132.4 (13.2)	0.88
Median (IQR)	130.0 (17.0)	130.0 (16.3)		130.0 (15.8)	130.0 (15.0)	
Diastolic BP (last Jul13-Jun14)						
Missing (%)	683 (48%)	622 (49%)		333 (43%)	273 (44%)	
Mean (SD)	75.1 (9.4)	76.0 (8.7)	0.08	75.0 (8.9)	76.2 (8.8)	0.06
Median (IQR)	75.0 (10.0)	76.3 (10.3)		75.1 (10.0)	76.0 (10.8)	
Cholesterol (last Jul13-Jun14)						
Missing (%)	1,309 (92%)	1,164 (93%)		695 (90%)	563 (91%)	
Mean (SD)	4.5 (1.6)	4.5 (1.3)	0.80	4.4 (1.4)	4.4 (1.2)	0.92
Median (IQR)	4.2 (1.4)	4.4 (2.0)		4.1 (1.4)	4.4 (2.1)	
Hyperlipidemia						
Diagnosis before Jul 2014	504 (35%)	373 (30%)	0.002	283 (37%)	199 (32%)	0.09
Cardiovascular disease						
Diagnosis before Jul 2014	616 (43%)	543 (43%)	0.99	359 (46%)	279 (45%)	0.65
Arthritis (any type)						
Diagnosis before Jul 2014	712 (50%)	562 (45%)	0.007	393 (51%)	295 (48%)	0.26
Back pain (related diagnosis)						
Diagnosis before Jul14	342 (24%)	257 (20%)	0.03	196 (25%)	132 (21%)	0.08

Table 10. cont. BP arm baseline characteristics before multiple imputation

	Analyses 5 & 6			Analysis 7		
Baseline characteristics	**TM** **N = 1,429**	**Controls** **N = 1,259**	**P**	**TM** **N = 773**	**Controls** **N = 617**	**P**
Walking pain (related diagnosis)						
Diagnosis before Jul14	166 (12%)	147 (12%)	0.96	91 (12%)	88 (14%)	0.17
Eye problem (related diagnosis)						
Diagnosis before Jul14	159 (11%)	107 (9%)	0.02	89 (12%)	55 (9%)	0.11
Insulin or Analogue						
Prescribed before Jul14	229 (16%)	164 (13%)	0.03	113 (15%)	85 (14%)	0.66
Pain relief drug						
Prescribed before Jul14	801 (56%)	580 (46%)	<.0001	447 (58%)	318 (52%)	0.02
Employment status						
Missing (%)	650 (46%)	663 (53%)		345 (45%)	278 (45%)	
Full-time	69 (5%)	57 (5%)	0.34	33 (4%)	36 (6%)	0.03
Part-time	47 (3%)	50 (4%)		25 (3%)	32 (5%)	
Self-employed	43 (3%)	26 (2%)		25 (3%)	16 (3%)	
No employment	409 (29%)	295 (23%)		241 (31%)	156 (25%)	
Retired	211 (15%)	168 (13%)		104 (13%)	99 (16%)	
Moderate exercise						
Missing (%)	731 (51%)	717 (57%)		384 (50%)	307 (50%)	
Yes (before Jul 2014)	388 (27%)	345 (27%)	0.004	223 (29%)	211 (34%)	0.004
Smoking status						
Missing (%)	806 (56%)	775 (62%)		427 (55%)	344 (56%)	
Never smoker	380 (27%)	299 (24%)	0.82	218 (28%)	163 (26%)	0.33
Past smoker	231 (16%)	178 (14%)		122 (16%)	108 (18%)	
Current smoker	12 (0.8%)	7 (0.6%)		6 (0.8%)	2 (0.3%)	
Risk level (last Jul13-Jun14)						
Missing (%)	86 (6%)	94 (7%)		48 (6%)	38 (6%)	
Extreme Risk	68 (5%)	35 (3%)	<.0001	35 (5%)	11 (2%)	0.009
High Risk	284 (20%)	178 (14%)		140 (18%)	87 (14%)	
Medium Risk	102 (7%)	107 (9%)		59 (8%)	54 (9%)	
Low Risk	496 (35%)	467 (37%)		281 (37%)	243 (39%)	
Self-Care	393 (28%)	378 (30%)		210 (27%)	184 (30%)	

Table 11. Both arms: Diagnosis variable definitions

With an assumption that some diagnoses are data entry mistakes, e.g. Diabetes insipidus.

Variable	Variable values	ICD10Code	Diagnosis from Healthways database
Diabetes Type	Other or unspecified	E09	Impaired glucose regulation
	Type 1	E10	Type 1 diabetes mellitus
	Type 1	E1040	Type 1 diabetes mellitus with unspecified neuropathy
	Type 1	E1043	Type 1 diabetes mellitus with diabetic autonomic neuropathy
	Type 1	E108	Type 1 diabetes mellitus with unspecified complication
	Type 2	E11	Type 2 diabetes mellitus
	Type 2	E1131	Type 2 diabetes mellitus with background retinopathy
	Type 2	E1140	Type 2 diabetes mellitus with unspecified neuropathy
	Type 2	E1142	Type 2 diabetes mellitus with diabetic polyneuropathy
	Type 2	E1164	Type 2 diabetes mellitus with hypoglycaemia
	Type 2	E1172	Type 2 diabetes mellitus with features of insulin resistance
	Type 2	E1173	Type 2 diabetes mellitus with foot ulcer due to multiple causes
	Type 2	E119	Type 2 diabetes mellitus without complication
	Other or unspecified	E13	Other specified diabetes mellitus
	Other or unspecified	E1336	Other specified diabetes mellitus with diabetic cataract
	Other or unspecified	E1340	Other specified diabetes mellitus with unspecified neuropathy
	Other or unspecified	E1373	Other specified diabetes mellitus with foot ulcer - multiple causes
	Other or unspecified	E14	Unspecified diabetes mellitus
	Other or unspecified	E1434	Unspecified diabetes mellitus with other retinopathy
	Other or unspecified	E1440	Unspecified diabetes mellitus with unspecified neuropathy
	Other or unspecified	E1472	Unspecified diabetes mellitus with features of insulin resistance
	Other or unspecified	E232	Diabetes insipidus
	Other or unspecified	HCF9	Diabetes - unconfirmed
Hypertension	Yes/No	EM258	Hypertension
		I10	Essential (primary) hypertension
		I11	Hypertensive heart disease
		I158	Other secondary hypertension
Hyperlipidemia	Yes/No		High cholesterol
		E780	Pure hypercholesterolemia
		E781	Pure hyperglyceridaemia
		E784	Other hyperlipidemia
		E785	Hyperlipidemia, unspecified
Cardiovascular Disease	Yes/No		Arrhythmia
			Atrial fibrillation
			Atrial fibrillation2006
			Bilateral varicose veins operation
			Blockage in 1 valve
			Cardiac stent placed x 7
			Mitral valve stenosis
			Pacemaker and AICD replaced
			Systemic stroke
		AM034	Cerebrovascular disorders except transient ischemic attacks
		B70C	Stroke w/o other cc
		EM249	Acute myocardial infarction
		EM253	Deep vein thrombophlebitis
		EM254	Peripheral vascular disorders
		EM270	Angina pectoris
		EP223	Cardiac valve procedures
		EP224	Coronary bypass
		EP236	Perm cardiac pacemaker implant
		EP238	Cardiac pacemaker device replacement
		F15Z	Percutaneous coronary angioplasty w/o AMI W stent implantation
		HCF3	Atrial fibrillation - unconfirmed

		HCF6	Coronary artery disease (CAD) - unconfirmed
		I058	Other mitral valve diseases
		I083	Combined disorders of mitral, aortic and tricuspid valves
		I088	Other multiple valve diseases
		I089	Multiple valve disease, unspecified
		I20	Angina pectoris
		I209	Angina pectoris, unspecified
		I21	Acute myocardial infarction
		I219	Acute myocardial infarction, unspecified
		I25	Chronic ischaemic heart disease
		I250	Atherosclerotic cardiovascular disease, so described
		I251	Atherosclerotic heart disease
		I252	Old myocardial infarction
		I259	Chronic ischaemic heart disease, unspecified
		I26	Pulmonary embolism
		I30	Acute pericarditis
		I319	Disease of pericardium, unspecified
		I350	Aortic (valve) stenosis
		I359	Aortic valve disorder, unspecified
		I390	Mitral valve disorders in diseases classified elsewhere
		I42	Cardiomyopathy
		I455	Other specified heart block
		I460	Cardiac arrest with successful resuscitation
		I471	Supraventricular tachycardia
		I472	Ventricular tachycardia
		I48	Atrial fibrillation and flutter
		I499	Cardiac arrhythmia, unspecified
		I50	Heart failure
		I500	Congestive heart failure
		I516	Cardiovascular disease, unspecified
		I519	Heart disease, unspecified
		I64	Stroke, not specified as haemorrhage or infarction
		I7020	Atherosclerosis of arteries of extremities, unspecified
		I712	Thoracic aortic aneurysm, without mention of rupture
		I73	Other peripheral vascular diseases
		I730	Raynaud's syndrome
		I738	Other specified peripheral vascular diseases
		I739	Peripheral vascular disease, unspecified
		I82	Other venous embolism and thrombosis
		I829	Embolism and thrombosis of unspecified vein
		R00	Abnormalities of heart beat
		R000	Tachycardia, unspecified
		R001	Bradycardia, unspecified
		R002	Palpitations
		R01	Cardiac murmurs and other cardiac sounds
		R011	Cardiac murmur, unspecified
		R110	Neurological stroke
Arthritis (any type)	Yes/No		Psoriatic arthritis
		HCF19	Osteoarthritis - unconfirmed
		M0125	Arthritis in Lyme disease, pelvic region and thigh (A69.2+)
		M06	Other rheumatoid arthritis
		M0680	Other specified rheumatoid arthritis, multiple sites
		M0689	Other specified rheumatoid arthritis, site unspecified
		M069	Rheumatoid arthritis unspecified
		M0690	Rheumatoid arthritis, unspecified, multiple sites
		M0699	Rheumatoid arthritis, unspecified, site unspecified
		M0900	Juvenile arthritis in psoriasis, multiple sites (L40.5+)
		M13	Other arthritis

		M130	Polyarthritis, unspecified
		M138	Other specified arthritis
		M139	Arthritis unspecified
		M1390	Arthritis, unspecified, multiple sites
		M1393	Arthritis, unspecified, forearm
		M1394	Arthritis, unspecified, hand
		M1396	Arthritis, unspecified, lower leg
		M1397	Arthritis, unspecified, ankle and foot
		M1398	Arthritis, unspecified, other site
		M1399	Arthritis, unspecified, site unspecified
		M150	Primary generalised (osteo)arthrosis
Back Pain	Yes/No		Back pain
			Back problems
			Degenerative spine
			Laminectomy - spinal fusion
			Lower back pain
			Ruptured disc
			Scoliosis of spine
			Spinal surgery
			Upper back pain
		AP025	Spinal procedures
		HCF16	Low back pain - unconfirmed
		HM432	Medical back problems
		HP447	Back and neck procedures with spinal fusion
		HP448	Back and neck procedures without spinal fusion
		M41	Scoliosis
		M4326	Other fusion of spine, lumbar region
		M4506	Ankylosing spondylitis, lumbar region
		M480	Spinal stenosis
		M4802	Spinal stenosis, cervical region
		M4807	Spinal stenosis, lumbosacral region
		M51	Other intervertebral disc disorders
		M513	Other specified intervertebral disc degeneration
		M518	Other specified intervertebral disc disorders
		M519	Intervertebral disc disorder, unspecified
		M54	Dorsalgia
		M543	Sciatica
		M545	Low back pain
		M546	Pain in thoracic spine
		S1316	Dislocation of C6/C7 cervical vertebrae
Other Walking Pain	Yes/No		Cervical fractures
			Dorsal and plantar spur on both feet
			Femoral bypass and graft surgery
			Hip replacement
			Knee replacement
			Poor circulation of the leg
			Right ankle injury
			Stenting on the left leg
			Toe amputation
		HM425	Fracture of femur
		HM437	Tendonitis, myositis and bursitis
		HP403	Hip and femur procedures except major joint
		HP413	Knee procedures
		HP416	Foot procedures
		HP422	Arthroscopy
		JP520	Amputation of lower limb for endocrine, nutrit and metabol disorders
		K41	Femoral hernia

		M0737	Other psoriatic arthropathies, ankle and foot (L40.5+)
		M10	Gout
		M109	Gout unspecified
		M1096	Gout, unspecified, lower leg
		M353	Polymyalgia rheumatica
		M6265	Muscle strain, pelvic region and thigh
		M6797	Unspecified disorder of synovium and tendon, ankle and foot
		M706	Trochanteric bursitis
		M707	Other bursitis of hip
		M710	Abscess of bursa
		M7115	Other infective bursitis, pelvic region and thigh
		M7117	Other infective bursitis, ankle and foot
		M712	Synovial cyst of popliteal space [Baker]
		M7136	Other bursal cyst, lower leg
		M7141	Calcium deposit in bursa, shoulder region
		M7156	Other bursitis, not elsewhere classified, lower leg
		M722	Plantar fascial fibromatosis
		M797	Fibromyalgia
		M7970	Fibromyalgia, multiple sites
		M8437	Stress fracture, not elsewhere classified, ankle and foot
		M955	Acquired deformity of pelvis
		R235	Orthopaedic - other joint replacement
		S720	Fracture of neck of femur
		S7205	Fracture of base of neck of femur
		S821	Fracture of upper end of tibia
		S8241	Fracture of upper end of fibula
		S825	Fracture of medial malleolus
		S832	Tear of meniscus, current
		S837	Injury to multiple structures of knee
		S860	Injury of achilles tendon
		S870	Crushing injury of knee
		S96	Injury muscle tendon at ankle foot level
		W01	Fall same lvl from slip trip & stumble
		W010	Fall on same level from slipping
		W06	Fall involving bed
		W109	Fall on & frm oth and unspec stair step
		W135	Fall from or through floor
		W138	Fall from, out of or through other specified building or structure
		W18	Other fall on same level
		W189	Unspecified fall on same level
		Z441	Fitting and adjustment of artificial leg (complete)(partial)
		Z740	Reduced mobility
		Z9664	Presence of hip implant
		Z9665	Presence of knee implant
Eye Problems	Yes/No		Bleeding in the back of eye Blindness Cataract r eye Cataract surgery and glaucoma Cataracts Cateracts removed Cva and loss of eyesight in right remaining eye Detached retina right eye Left eye with very limited eyesight. Macular degeneration Vision impairment
		BM085	Other disorders of the eye
		BP073	Lens procedures with or without vitrectomy
		H183	Changes in corneal membranes

	H25	Senile cataract
	H259	Senile cataract, unspecified
	H26	Other cataract
	H262	Complicated cataract
	H264	After-cataract
	H269	Cataract, unspecified
	H28	Cataract & oth disrd lens in dis cl/e
	H33	Retinal detachments and breaks
	H350	Background retinopathy and retinal vascular changes
	H353	Degeneration of macula and posterior pole
	H40	Glaucoma
	H409	Glaucoma, unspecified
	H544	Blindness, one eye
	H57	Other disorders of eye and adnexa
	Z947	Corneal transplant status
	Z961	Presence of intraocular lens

Table 12. Both arms: Medication variable definitions

Variable	Variable values	Medicine Class from Healthways database
Insulin or Analogue	Yes/No	Alpha glucosidase inhibitors Biguanides Comb.sulfonamides & trimethoprim incl. derivatives Combinations of oral blood glucose lowering drugs Intermediate-acting sulfonamides Oral blood glucose lowering drugs Other oral blood glucose lowering drugs Sulfonamides, plain Sulfonamides, urea derivatives Thiazolidinediones
Number of Type 2 Drugs	0 to 4	Alpha glucosidase inhibitors Biguanides Comb.sulfonamides & trimethoprim incl. derivatives Combinations of oral blood glucose lowering drugs Intermediate-acting sulfonamides Oral blood glucose lowering drugs Other oral blood glucose lowering drugs Sulfonamides, plain Sulfonamides, urea derivatives Thiazolidinediones
Pain Relief Drugs	Yes/No	Acetic acid derivatives and related substances Anilides Corticosteroids Corticosteroids and antiinfectives in combination Corticosteroids and mydriatics in combination Corticosteroids, plain Corticosteroids, potent (group III) Corticosteroids, weak (group I)

	Corticosteroids, weak, comb with antiseptics
	Glucocorticoids
	Natural opium alkaloids
	Opioids
	Opium alkaloids and derivatives
	Oripavine derivatives
	Other antiinfl./antirheumatic agents, non-steroids
	Other opioids
	Oxicams
	Preparations inhibiting uric acid production
	Preparations w. no effect on uric acid metabolism
	Propionic acid derivatives
	Salicylic acid and derivatives

Figure 3. BP arm distribution of all telemonitoring device SBP and DBP measurements

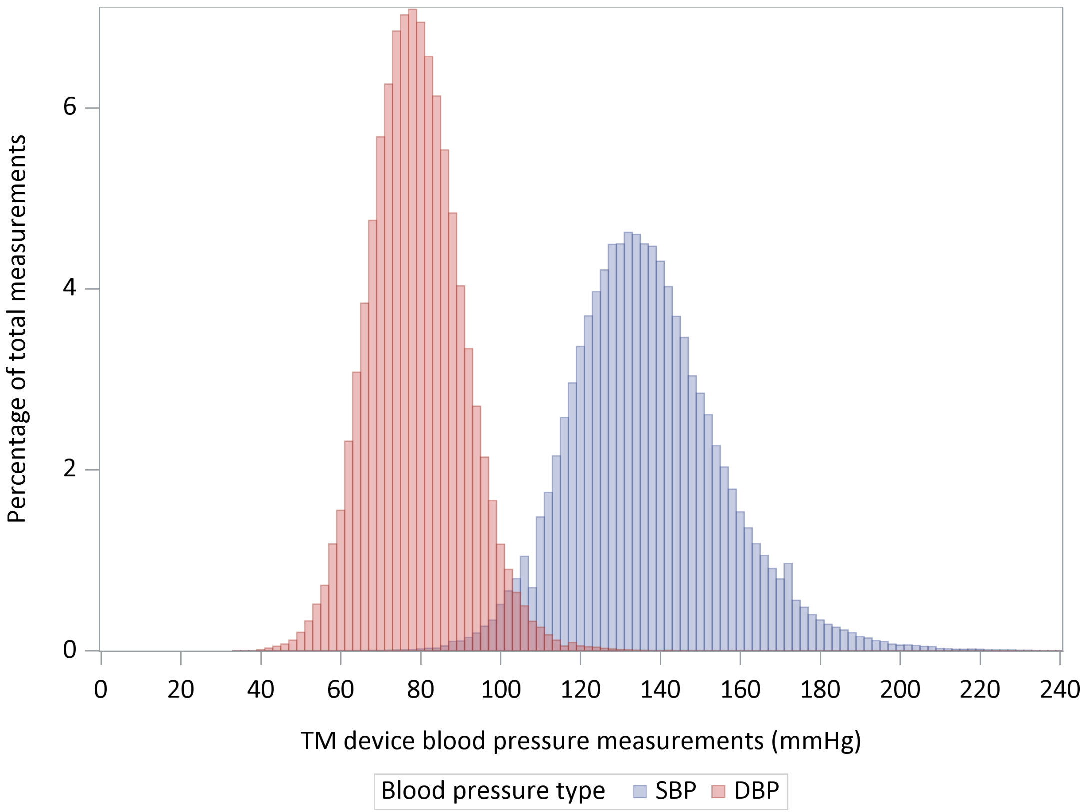


Figure 4. Causal diagram for the Glucose arm modified for a presentation

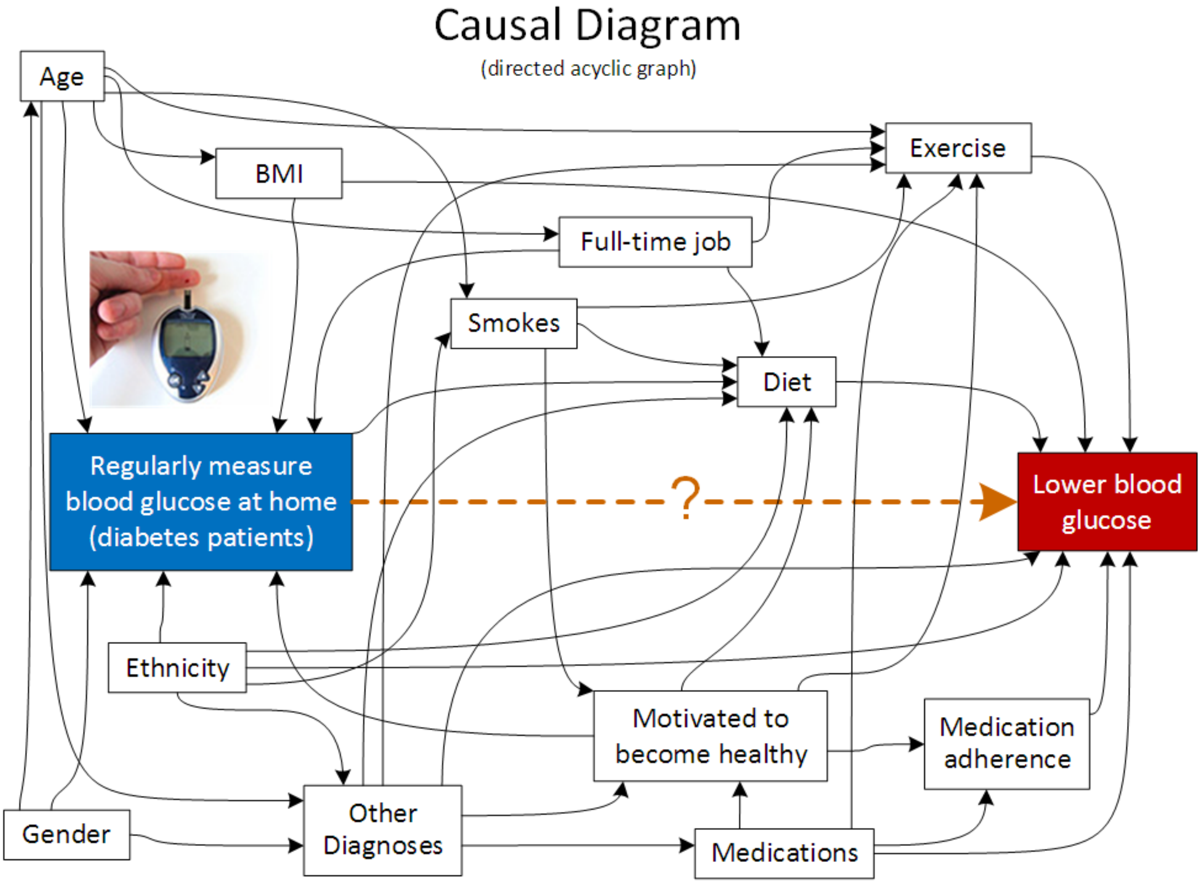


Figure 5. BP arm participants and total number of weeks with BP measured

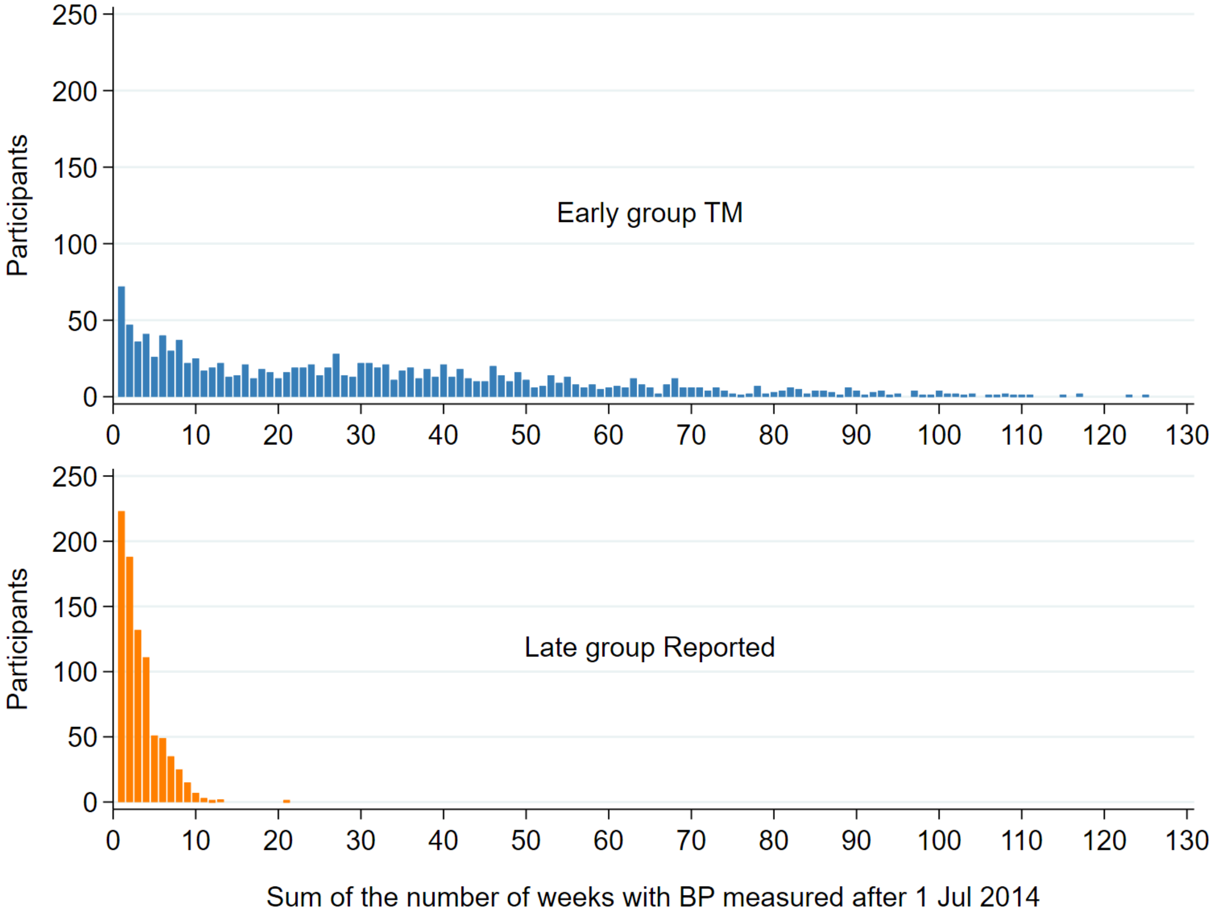


Figure 6. Participants in BP arm with total weeks between first and last measurement

Telemonitoring device measurements from the Telemonitoring group; reported over the phone measurements from Control group

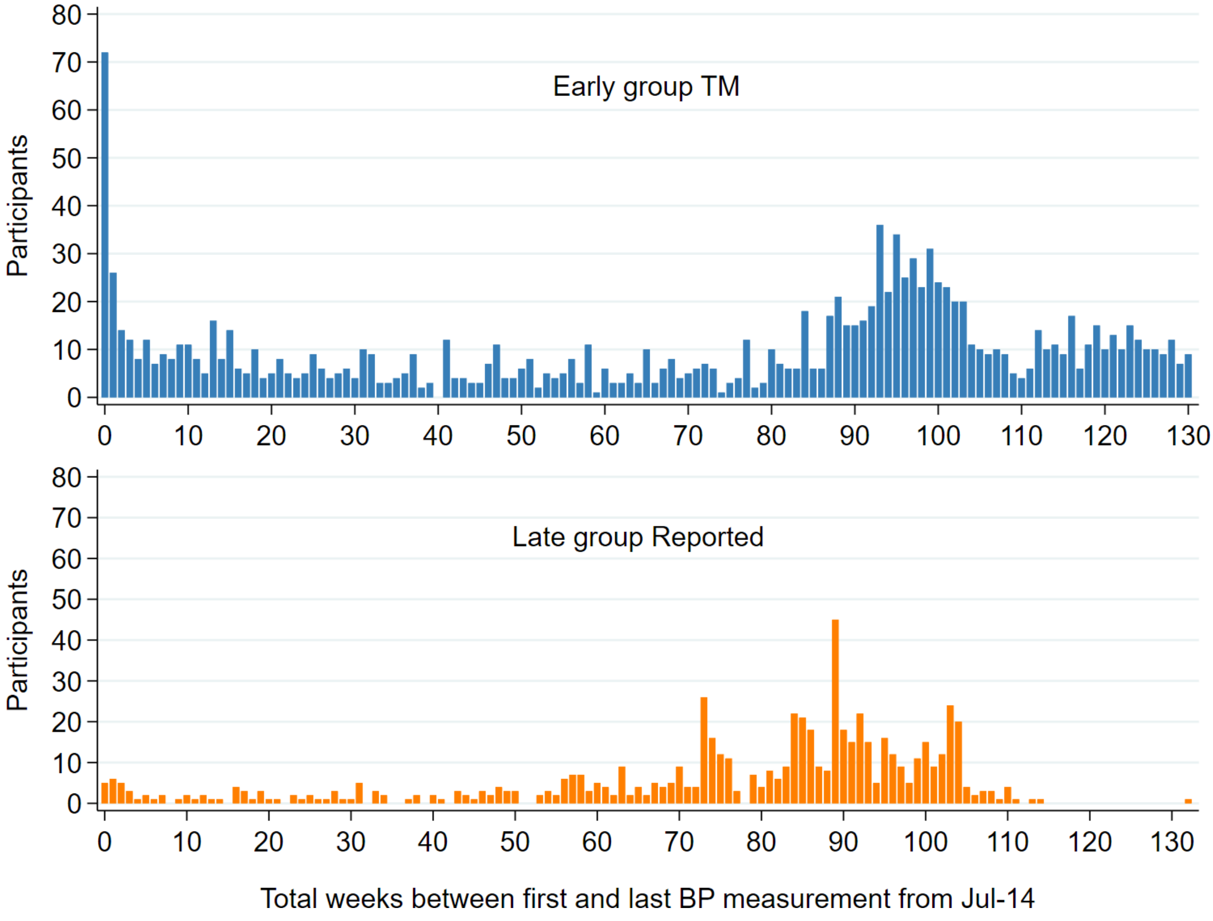


Table 13. Glucose arm baseline characteristics for the full dataset after multiple imputation

All enrolled participants are included.

Baseline characteristics		Telemonitoring N = 549	Controls N = 299	P-value
Sex	*N imputed*	*0*	*0*	
	Male	322 (59%)	193 (65%)	0.093
	Female	227 (41%)	106 (35%)	
Age (years)	*N imputed*	*0*	*0*	
	Mean (SD)	67.9 (10.8)	65.9 (11.4)	0.010
	Median (IQR)	69 (13)	67 (15)	
Ethnicity	*N imputed*	*115/549*	*88/299*	
(missing 22%)	Caucasian	481 (88%)	259 (87%)	0.753
	Asian	28 (5%)	22 (7%)	
	African	0	1	
	Pacific Islander	0	1	
	Aboriginal	1	1	
	Other	39 (7%)	15 (5%)	
HbA1c	*N imputed*	*396/549*	*225/299*	
(missing 73%)	Mean (SD)	6.8 (1.3)	6.9 (1.3)	0.723
	Median (IQR)	6.8 (1.6)	6.8 (1.7)	
BMI	*N imputed*	*183/549*	*98/299*	
(missing 31%)	Mean (SD)	30.4 (5.8)	30.1 (5.8)	0.546
	Median (IQR)	29.8 (7.6)	29.6 (7.0)	
Diabetes type	*N imputed*	*3/549*	*1/299*	
	Type 1	30 (5%)	21 (7%)	0.129
	Type 2	494 (90%)	270 (90%)	
	Type 1 & 2	1	1	
	Other/unspecified	24 (4%)	7 (2%)	
Hypertension	*N imputed*	*0*	*0*	
	Diagnosis before Jul 2014	321 (58%)	64 (21%)	< .0001
Hyperlipidemia	*N imputed*	*0*	*0*	
	Diagnosis before Jul 2014	162 (30%)	60 (20%)	0.003
Cardiovascular disease	*N imputed*	*0*	*0*	
	Diagnosis before Jul 2014	258 (47%)	115 (38%)	0.017
Arthritis (any type)	*N imputed*	*0*	*0*	
	Diagnosis before Jul 2014	249 (45%)	113 (38%)	0.033
Back pain related diagnosis	*N imputed*	*0*	*0*	
	Diagnosis before Jul14	108 (20%)	64 (21%)	0.549
Walking pain related diagnosis	*N imputed*	*0*	*0*	
	Diagnosis before Jul14	77 (14%)	44 (15%)	0.784

Table 13 cont. Glucose arm baseline characteristics after multiple imputation

Baseline		Telemonitoring N = 271	Controls N = 263	P-value
Eye problem	*N imputed*	*0*	*0*	
related diagnosis	Diagnosis before Jul14	59 (11%)	33 (11%)	0.897
Insulin or Analogue	*N imputed*	*0*	*0*	
prescribed before Jul14	Prescribed	119 (22%)	50 (17%)	0.085
Pain relief drug	*N imputed*	*0*	*0*	
prescribed before Jul14	Prescribed	292 (53%)	139 (46%)	0.062
Number of Type 2	*N imputed*	*0*	*0*	
diabetes drugs	0 drugs prescribed	162 (30%)	106 (35%)	0.209
prescribed before Jul14	1 drugs prescribed	241 (44%)	126 (42%)	
	2 drugs prescribed	113 (21%)	56 (19%)	
	3 drugs prescribed	31 (6%)	9 (3%)	
	4 drugs prescribed	2	2	
Employment status	*N imputed*	*449/549*	*240/299*	
(missing 81%)	Full-time	46 (8%)	30 (10%)	0.221
	Part-time	62 (11%)	46 (16%)	
	Self-employed	29 (5%)	15 (5%)	
	No employment	122 (22%)	72 (24%)	
	Retired	289 (53%)	137 (46%)	
Moderate exercise	*N imputed*	*486/549*	*273/299*	
(missing 88%)	Yes before Jul 2014	135 (24%)	49 (16%)	0.183
Smoking status	*N imputed*	*270/549*	*141/299*	
(missing 45%)	Never smoker	329 (60%)	177 (59%)	0.972
	Past smoker	185 (34%)	106 (35%)	
	Current smoker	35 (6%)	16 (5%)	
Risk level	*N imputed*	*0*	*0*	
(last recorded Jul13-Jun14)	Extreme Risk	20 (4%)	12 (4%)	0.008
	High Risk	127 (23%)	57 (19%)	
	Medium Risk	41 (7%)	14 (5%)	
	Low Risk	191 (35%)	87 (29%)	
	Self Care	170 (31%)	129 (43%)	