Dealing with partially observed covariates in propensity score analysis of observational data

Helen Abigail Blake

Thesis submitted in accordance with the requirements for the degree of

Doctor of Philosophy of the University of London

September 2019

Department of Medical Statistics

Faculty of Epidemiology and Population Health

LONDON SCHOOL OF HYGIENE & TROPICAL MEDICINE

Funded by the Economic and Social Research Council

Research group affiliations: Electronic Health Records group
Missing Data Interest Group

I, Helen Abigail Blake, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Observational data, such as electronic health records, are a valuable source of information for researchers seeking to answer health-related questions. Since treatment allocation is not typically randomized in studies using observational data, there is confounding – systematic differences in the characteristics of patients in different treatment groups. Propensity score analysis (PSA) can be used to handle confounding by modelling the probability of being allocated to a particular treatment, based on patient characteristics. However, a common issue in analyses of observational data is missing data. In general, not dealing appropriately with missing data can lead to loss of efficiency and biased estimates of the treatment effect. Furthermore, having partially observed covariate data can complicate the estimation of the propensity score.

The missingness pattern approach (MPA) has been proposed to handle partially observed covariate data in PSA. One key objective of my thesis is to understand when the approach is appropriate, by exploring its underlying assumptions. I began by comparing different statements of the MPA's underlying assumptions given in the literature. I considered the plausibility of the MPA's assumptions in simple scenarios, finding that they are separate to the conventional classification of missingness mechanisms. I used d-separation (a rule for testing conditional independence statements) with single world intervention graphs, representing a variety of scenarios, in order to develop guidance for when the assumptions seem plausible.

I also explored the connection between using the MPA and using missing indicators in the context of PSA, finding that the use of missing indicators is a simplification of the MPA. I extended this work to outcome regression, mathematically proving that using missing indicators is valid under the MPA's assumptions as well as an additional simplifying assumption. I also conducted simulation studies to assess bias when using missing indicators to handle partially observed covariate data in outcome regression.

Acknowledgements

Firstly, huge thanks to my supervisors Elizabeth Williamson, Clémence Leyrat and James Carpenter for their invaluable advice, support, and encouragement given to me during the course of my PhD. They always made time for me and were very patient.

I would like to also thank my other co-authors for their guidance and useful comments when developing the two papers in this thesis, as well as for their support in using the electronic health records data, and I would also like to thank my upgrading committee for their insightful feedback.

I would also like to acknowledge the support of my fellow PhD students with whom I shared an office during the course of this research.

I would like to thank Economic and Social Research Council and the Bloomsbury Doctoral Teaching Centre for funding this research.

I have appreciated all of the support from my parents, sisters and brothers-in-law throughout this research and indeed all of my studies. My little niece Louise has been a wonderful distraction from the thesis.

Contents

List of Tables

List of Figures

List of Abbreviations

ACEI	Angiotensin converting enzyme inhibitor
AKI	Acute kidney injury
ARB	Angiotensin receptor blocker
ATC	Average treatment effect in the control group
ATE	Average treatment effect
ATT	Average treatment effect in the treatment group
BMI	Body mass index
CIO	Conditionally independent outcomes assumption
CIT	Conditionally independent treatment assumption
CKD	Chronic kidney disease
CPRD	UK Clinical Practice Research Datalink
CRA	Complete records analysis
eGFR	Estimated glomerular filtration rate
EHR	Electronic health record
HES	Hospital episode statistics
ICD-10	International statistical classification of diseases and related health problems – 10th revision
IPTW	Inverse probability of treatment weighting
MAR	Missing at random
MCAR	Missing completely at random
MI	Multiple imputation
MIA	Missing indicator approach
MNAR	Missing not at random
MPA	Missingness pattern approach
mSITA	Missingness strongly ignorable treatment allocation assumption
OPCS	Office of Population Censuses and Surveys
PSA	Propensity score analysis
SITA	Strongly ignorable treatment allocation assumption
SWIT	Single world intervention graph template

Chapter 1

Introduction

Establishing the efficacy and safety of commonly used drugs remains a challenge in pharmacoepidemiological research. Real world evidence — arising from observational data obtained outside the context of highly controlled randomized clinical trials, typically data generated during routine clinical practice — plays an increasingly important role in a wide range of pharmacoepidemiological investigations. The expectation that routinely collected health data will be used to measure medication effects – both harms and benefits – is now written into EU legislation [1]. In the US, legislation has noted the potential benefits of using routinely collected health data for regulatory decisions [2].

Important questions that can be usefully addressed using large scale routine health data include the investigation of long-term and rare effects of medications, treatment interactions, efficacy of drugs in patients with rare conditions, long-term resistance to treatments such as antibiotics, and establishing optimal treatment policies for chronic conditions [3].

These questions are all, at heart, causal questions. Causal inference is the process of drawing conclusions about questions regarding causal relationships, such as the comparative effect of different treatments on a health outcome [4].

A framework for the formal definition and estimation of causal effects, based on the idea of counterfactuals — the idea of what would have happened had a different treatment been prescribed — has been proposed and is widely used in

pharmacoepidemiological research to address causal questions [5].

Randomized controlled trials are commonly considered to be the 'gold-standard' for the estimation of causal effects. Evidence from randomized trials has a number of limitations. The patients recruited to randomized controlled trials are often not representative of the general population. More specifically, they tend to include patients who are younger, more often male and who have fewer comorbidities. Trials often exclude the very patients who tend to be treated with the treatments under investigation in clinical practice. Also, the treatment administration and monitoring in trials is highly controlled, often leading to much higher adherence to the prescribed treatment. As such, patients' use of treatments in trials may not be reflective of use in clinical practice.

Real-world evidence, reflecting the effectiveness of treatments in routine clinical practice for large samples of the general population over long periods of time, can be obtained by the analysis of routinely collected health data with limited exclusion criteria [6]. Hence, observational studies can be used to answer research questions surrounding the effectiveness and safety of treatments in long-term routine clinical practice; such research questions may be difficult or infeasible to address through randomized controlled trials.

Routinely collected health-related information is increasingly been stored electronically. These electronic health records (EHRs) offer rich opportunities for pharmacoepidemiological research investigating treatments in real-world settings. Examples of EHR databases in the UK include the Clinical Practice Research Datalink, Hospital Episode Statistics, and The Health Improvement Network [7–9]. These EHR databases contain large numbers of anonymised patient records, with information on demographic characteristics, as well as prescriptions, diagnostic tests and procedures.

Since EHR data are not collected for the purposes of research, but rather as part of routine clinical or administrative practice, treatment allocation is not random and is instead dependent on a range of factors including age, sex and comorbidities. So, characteristics for one treatment group (say, the active treatment) may system-

atically differ from those of another treatment group (say, the control treatment). If these characteristics are risk factors for the outcome under study, then these characteristics may 'confound' the causal relationship between treatment and control and lead to biased results. Hence, observational studies using EHR must use strategies for dealing with confounding bias.

While multivariable regression has a long history of use as a method to account for confounding in observational data, methods based on the propensity score have been increasingly applied, particularly to the analysis of large-scale health data. The popularity of propensity score methods in this context is in part due to the ability to estimate marginal population-level effects, which are typically more relevant for policy makers. Further, propensity score methods can more readily handle large amounts of potential confounding data, which is a major advantage when investigating rare outcomes.

Propensity score analysis compares patients in the active treatment group to patients from the control group with the same propensity for being allocated to the active treatment group [10]. The basic premise relies on the idea that patients who have a similar 'likelihood' of receiving the treatment — whether or not they actually do receive a prescription for that treatment — are, on average, similar. Therefore, propensity score analysis compares outcomes of patients with similar propensity scores to obtain estimates of treatment effect.

Whatever analytic approach is used to account for confounding bias, the problem of missing data is likely to arise. In EHR data, while health outcomes and treatment prescriptions are usually well recorded, variables that potentially lead to confounding bias are less so. For example, these potential 'confounders' include patient characteristics, such as ethnicity or BMI. Further, the extent of missing data in EHR is typically much greater than is likely to arise in more traditional study designs. Therefore, the problem of missing confounder data in studies using EHR is likely to pose considerable challenges.

To deal with the problem of missing confounder data, patients records are often excluded from analysis. However, this leads to a loss of information and can result

in biased estimates of the treatment effect.

This thesis focuses on propensity score based methods, due to its popularity in the analysis of EHR data. In such studies, there are a number of methods for dealing with missing confounder data, some of which have been specifically proposed for the context of propensity score analysis. One method that has been proposed is the missingness pattern approach (MPA), which incorporates information about the pattern of missing variables into the propensity score. The assumptions underlying the MPA have been discussed in the literature, however, the MPA has not been used much, possibly due to lack of understanding regarding these assumptions.

1.1 Motivating example

Renin angiotensin system blocking using ACE inhibitors (ACEI) and angiotensin receptor blockers (ARBs) is a common treatment for a wide range of conditions, including hypertension and heart failure. However, some patients may experience adverse effects. For instance, acute kidney injury (AKI) — a sudden decline in kidney function — is thought to be associated with use of ACEI/ARBs [11]. This relationship is biologically plausible, however evidence to support a causal link is limited: randomized evidence is scarce due to insufficient or no reporting of renal events in randomized trials of ACEI/ARBs [11]. Despite this limited evidence, guidelines recommend reducing or ceasing ACEI/ARB use during acute illness [11].

Mansfield et al. (2016) used data taken from UK primary care linked data, from the Clinical Practice Research Datalink (CPRD), to investigate the relationship between use of ACEI/ARBs and the risk of AKI [11]. The large amount of missing data in two potential confounders was handled using a missing category approach. The assumptions under which this approach would have provided valid inference had not yet been clearly outlined in the literature, thus the validity of the assumptions underlying this approach could not be fully explored.

I obtained ethics approval to use this data from The Medicines and Healthcare products Regulatory Agency, CPRD division and the London School of Hygiene and Tropical Medicine (Appendix A).

1.1.1 The Clinical Practice Research Datalink

The CPRD, formerly known as the General Practice Research Database, contains anonymised primary care records from over 1600 general practices for 11 million registered, alive patients as of 4th September 2019 [12]. These patients are fairly representative of the general UK population [7,13,14]. Data are collected by general practice staff as part of routine clinical care [7]. Data are recorded in a number of ways in the CPRD: clinical measures such as symptoms and diagnoses can be classified using Read codes [15] or recorded numerically [7], and prescription data is recorded with British National Formulary codes and dosage information [7]. Additional notes can be recorded as free-text, but are not available to researchers as standard.

CPRD data can be linked to other data sources, including Hospital Episode Statistics (HES) data, to provide more complete information about the patient pathway. Hospital Episodes Statistics (HES) contains records of all patients admitted to NHS hospitals in England, covering every hospital stay. Data for HES are recorded by clinicians and entered into an electronic database by dedicated clinical coding departments [8]. Data are recorded in a number of ways in HES, including: ICD-10 codes for the classification of diagnoses [16], OPCS codes to classify operations and procedures [17].

Linkage of CPRD and other datasets is carried out by a trusted third party, NHS Digital [18]. Linkage uses pseudonymised identifiers and a deterministic linkage algorithm that produces a ranking variable that indicates the quality of links [19].

1.1.2 Study design

Mansfield et al. (2016) used data taken from the CPRD linked to HES to investigate the relationship between use of ACEI/ARBs and the risk of AKI, using a cohort of new users of antihypertensive drugs to limit confounding by indication [11] (i.e. to avoid confounding bias arising from a comparison of ACEI/ARBs users and healthy patients with no antihypertensive prescriptions).

More than 500,000 new users of antihypertensives between 1997-2014 were in-

cluded in Mansfield et al.'s study. Being an observational comparative effectiveness study, strategies to deal with confounding were required: the authors chose to use multivariable Poisson regression, adjusting for age, sex, various chronic comorbidities, time exposed to other antihypertensive drugs and calendar period [11]. In this study, many of the baseline characteristics were not balanced across the treatment groups, indicating potential confounding bias (see Table 5.1 in the research paper pre-print in Chapter 5). By using propensity score analysis, potential confounders can be balanced using the propensity score to summarise all of the covariates, replacing the need to include all covariates separately in a regression model.

A key comorbidity and potential confounder, baseline chronic kidney disease (CKD) stage, had missing values in over 50% of patients. Ethnicity also had over 50% missing data. Restricting analysis to patients with complete records would lead to a large loss of data; only a fifth of the patients in the study had both ethnicity and baseline CKD stage recorded. Mansfield et al. opted to use a 'missing baseline CKD stage' category to minimize selection bias and performed a sensitivity analysis excluding patients with missing baseline CKD stage [11]. They also used sensitivity analysis to compare the main results for the cohort to the results for a subset of patients with known ethnicity, finding that neither of the sensitivity analyses had much effect on the results of the study [11]. However, they do not comment beyond this on the assumptions about missing data inherent in their analyses.

1.2 Aims and objectives

The overall aim of this thesis is to investigate missing data methods incorporating missingness information to deal with partially observed confounder data when using causal inference methods in observational studies, with a focus on gaining a clear understanding of the assumptions required and providing practical guidance for assessing their plausibility. The specific objectives are listed below.

Objective 1: to explore the assumptions underlying the missingness pattern approach. The missingness pattern approach (MPA) has been proposed as a method to handle missing confounder data in propensity score analysis. I will

explore the assumptions under which the MPA would provide valid inference, by: (i) investigating the connection between the MPA and the conventional classification of missing data proposed by Rubin (1976) [20], (ii) identifying settings where the assumptions are likely to be plausible, and (iii) developing ways of assessing the assumptions using causal diagrams.

Objective 2: to develop guidance for assessing the assumptions underlying the missingness pattern approach By developing the work from Objective 1, I will develop practical guidance for assessing the MPA's assumptions in a given setting and I will demonstrate the practical guidance on the motivating example using electronic health data.

Objective 3: to investigate the missing indicator approach for propensity score analysis. I will explore the relationship between the MPA and the missing indicator approach in the context of propensity score analysis, in particular investigating the implications of this relationship on the assumptions under which the missing indicator approach can provide valid inference.

Objective 4: to investigate the missing indicator approach for outcome regression. I will extend the work from Objective 3 to investigate the use of the missing indicator approach in the context of outcome regression.

Objective 5: to investigate variance estimation for missing confounder methods incorporating missingness patterns for propensity score analysis. I will derive a variance estimator for inverse probability of treatment weighting after using the MPA to deal with partially observed confounder data.

1.3 Thesis overview

I begin in Chapter 2 with an overview of the potential outcome framework for causal inference and the principles of propensity score analysis. I also describe missing data methods for propensity score analysis, introduce causal diagrams for representing causal relationships and review the use of the MPA for propensity score analysis in health research.

In Chapter 3, I explore the assumptions underlying the validity of the MPA, by

considering connections with prior work from the literature. I also devise ways of assessing the assumptions using causal diagrams.

In Chapter 4, I consider ways to communicate how to assess the assumptions in practice and provide the initial guidance developed for assessing the MPA's assumptions.

In Chapter 5, I present a research paper pre-print that explores the assumptions underlying the MPA and provides the current guidance for assessing the MPA's assumptions. This research paper has been submitted for publication in Statistics in Medicine.

In Chapter 6, I derive a variance estimator inverse probability of treatment weighting after using the MPA and discuss potential for future simple simulation studies to empirically assess the performance of this estimator.

In Chapter 7, I explore the connection between the MPA and the approach where a missingness category for partially observed characteristics is added to the propensity score model. I then extend these findings to investigate the use of missing indicators in standard outcome regression.

In Chapter 8, I present a research paper pre-print that explores the use of the missing indicator approach in standard outcome regression. This research paper has been provisionally accepted for publication by the Biometrical Journal.

In Chapter 9, I conclude with a discussion and propose new avenues for research in this area.

Chapter 2

Background

In this chapter, I provide some background information on the methodology used in my PhD. I begin with an overview of causal inference and the potential outcome framework, including propensity score analysis. Next, I describe missing confounder data methods, introducing key concepts in missing data methodology. I then introduce causal diagrams, which are a way of visually representing relationships in a scenario of interest. Finally, I review the existing literature using the missingness pattern approach.

2.1 Causal inference and the potential outcome framework

Causal inference is the process of drawing conclusions about questions regarding cause and effect. Causal questions in pharmacoepidemiological research concern the effects of drugs, medical devices or other medical interventions in a large population [3]. For example, my motivating example is a comparative effectiveness study that investigates the association between use of renin-angiotensin system blocking drugs, compared to other antihyperhensive drugs, and the risk of AKI in a cohort of over 500,000 adults [11]. Causal questions in medical research typically concern the effect of a treatment, exposure or intervention on health outcomes [21]; in this thesis I discuss causal effects in terms of treatments, to correspond with the motivating

example.

The Neyman-Rubin framework was developed to make inferences about causal effects and relies on the concept of counterfactuals: what would have happened had the cause not been present [10]. For example, suppose individuals in a study are allocated to one of two treatment groups, say an active treatment or a control. Each individual has an observed outcome and a counterfactual outcome (i.e. the outcome that would have happened if, counter to fact, the individual had a different treatment allocation). We refer to these collectively as potential outcomes. Each individual then has two potential outcomes: the outcome that would have been observed if they were allocated to the active arm, and the outcome that would have been observed if they were allocated to the control arm [10, 22]. Thus a contrast of potential outcome values for an individual gives the causal effect of treatment for this individual. However, the 'fundamental problem of causal inference' [22] is that this comparison cannot be made directly since only one of these potential outcomes can be observed and the other is counterfactual [23].

Instead, inferences are made by considering a group of individuals, some of whom are allocated to the active arm and others allocated to the control arm. The average outcomes from each treatment arm are compared to estimate the average treatment effect. This estimate is unbiased when the individuals in the active treatment arm are comparable with the individuals in the control arm [24]. An example of when the treatment arms are comparable is when treatment allocation is randomized [22].

When randomization is not feasible, observational data can be used to estimate treatment effects. However, in non-randomized settings, obtaining unbiased estimates of the treatment effect is more complex and relies on assumptions that are untestable in practice.

2.1.1 Notation and estimands of causal effect

Consider a group of n patients, with information on p characteristics represented by a row vector $X_i = (X_{i1}, \ldots, X_{ip})^\mathsf{T}$ where $i = 1, \ldots, n$ and X_i is fully observed. Throughout this thesis, treatment allocation is assumed to be binary, denoted by

$Z_i = 1$ if patient i is in the treatment group or $Z_i = 0$ if they are in the control group. Correspondingly, the two potential outcomes for patient i are denoted as $Y_i(z)$, where $z = 0, 1$. The observed outcome for patient i is denoted as Y_i. Henceforth, the subscripts are omitted where unambiguous.

An estimand is a quantity that we want to make inferences about. In this thesis, the estimand of interest is the average treatment effect (ATE): $E[Y(1) - Y(0)]$. Restricting attention to binary outcomes, this estimand is the risk difference [25,26]. The risk difference is often of interest in public health questions and is easy to interpret (as it is an absolute measure) [27]. In addition, the risk difference has the desirable property of being collapsible, unlike the odds ratio [27]. An alternative estimand would be the marginal risk ratio, $E[Y(1)/Y(0)]$, which is also collapsible [25].

While I focus on causal inference for the whole population, sometimes the population of interest in a research question is the subgroup of patients in the treatment arm, for which the corresponding estimand is called the average treatment effect in the treatment group (ATT), $E[Y(1) - Y(0)|Z = 1]$ [25, 28]. Similarly, if the population of interest is the subgroup of patients in the control arm, the corresponding estimand is the average treatment effect in the control group (ATC), $E[Y(1) - Y(0)|Z = 0]$. The choice between the ATE, ATT and ATC depends on the context of the research [26, 28, 29]; analogous definitions of the estimands can be made in terms of risk ratios or odds ratios as required [25]. In this thesis, attention is restricted to the ATE in terms of the risk difference.

2.1.2 Identification of causal effects

Although the Neyman-Rubin framework was developed for randomized controlled trials [30], we can use observational data to obtain estimates of the ATE under certain 'identifiability' assumptions [31]. In this thesis, we assume the following hold: the consistency assumption, the strongly ignorable treatment assignment assumption, the 'no interference' assumption, and the positivity assumption. Note that different identifiability assumptions may be used in other causal inference ap-

proaches, such as analyses using instrumental variables [10, 32].

The consistency assumption states that, if an individual is assigned a particular treatment then the corresponding potential outcome will be observed for that individual, irrespective of the way in which they were assigned to that treatment group [31]. This can be expressed as:

$$Y_i = Y_i(1) \times Z_i + Y_i(0) \times (1 - Z_i).$$

The 'no interference' assumption states that the treatment received by one patient does not affect the potential outcomes of another patient: [33–35]

$$Y_i(z_1, \ldots, z_i, \ldots, z_n) = Y_i(z_i),$$

where $Y_i(z_1, \ldots, z_i, \ldots, z_n)$ is the hypothetical potential outcome where Z_i is set to z_i for all values of $i = 1, \ldots, n$.

The strongly ignorable treatment assignment (SITA) assumption is that there is no unmeasured confounding: [33]

$$\big(Y_i(1), Y_i(0)\big) \perp Z_i | X_i \forall i \ . \tag{2.1}$$

The SITA assumption has also been referred to as 'conditional exchangeability' since the treatment and control groups are exchangeable based on the observed covariate information [36].

Finally, the positivity assumption states that, on the basis of their characteristics, it must be possible for each individual to be allocated to treatment or to control [23, 37], and can be expressed as:

$$0 < P(Z_i = 1 | X_i) < 1 \quad \forall \ i \ .$$

Throughout, we assume that these assumptions hold in the complete data.

2.1.3 Treatment effect estimation in observational studies

Randomized controlled trials are considered to be the gold-standard for causal inference. When treatment allocation is randomized, patient characteristics are balanced on average across the two treatment groups (i.e. have similar distributions in the two groups) and so the ATE can be identified as:

$$\text{ATE} = E[Y|Z = 1] - E[Y|Z = 0] \ .$$

When randomization is not feasible, observational data can be used to estimate treatment effects. Furthermore, observational studies can be used to answer research questions that could not feasibly be addressed using randomized controlled trials, such as the long-term efficacy and safety of treatments in routine practice. However, a common issue with observational data is confounding bias: systematic differences in patient characteristics between treatment groups. Since patient characteristics are not balanced across treatment groups, the SITA assumption does not hold and the ATE cannot be identified. One solution is to identify a set of confounders which satisfy the SITA assumption and thus identify the ATE using strategies to account for those confounders.

Conventionally, a variable is considered to be a 'confounder' in the epidemiological sense if it (i) is associated with treatment allocation, (ii) is associated with the outcome, and (iii) does not lie on the causal pathway between treatment and outcome [38]. In this thesis, we use the more formal definition: a variable is a confounder if it is a member of some set of variables that is sufficient to control for confounding [39]. Causal diagrams provide a way of identifying a set of measured variables which satisfies the SITA assumption of no unmeasured confounding.

When estimating the treatment effect in observational studies, two key approaches to deal with confounding are: (i) outcome regression models conditioning on confounders; or (ii) propensity score methods, as described below. In this thesis, I focus on propensity score methods for estimating marginal treatment effects.

2.1.4 Propensity score analysis

The propensity score $e(x)$ is the probability of being assigned to the treatment group, as opposed to the control group, given a set of observed characteristics:

$$e_i(x_i) = P(Z_i = 1 | X_i = x_i), \tag{2.2}$$

for patient i $(i = 1, \ldots, n)$ with a vector of confounder values $X_i = x_i$.

The propensity score is a balancing score: at each level of the propensity score, the distributions of observed characteristics are the same for treated individuals as for controls on average [33]. Rosenbaum and Rubin (1983) showed that at each value of the propensity score, the difference in mean outcomes for the treated and control groups is an unbiased estimate of the ATE at that value of the propensity score under the identifiability assumptions described above [33].

Typically, propensity scores are unknown and must be estimated from the data. Often, they are estimated using a logistic regression model for the treatment, with observed confounders as covariates [28]. The predictions obtained from this model are the individual estimated propensity scores. Sometimes factors that are not necessarily confounders but are associated with the outcome of interest can also be included as covariates to increase precision [25]. Alternative strategies for the estimation of propensity scores, such as classification trees, random forests and generalised boosted modelling, are discussed elsewhere [28, 40] but are not considered further here.

Rosenbaum and Rubin (1983) showed that, provided the above identifiability assumptions hold, matching, stratification and adjustment on the estimated propensity score can give unbiased estimates of the ATE [33]. In propensity score matching, treated and control individuals are 'matched', according to their propensity score and, for each matched pair, the difference in their observed outcomes are calculated. The average of these differences then provides an estimate of the ATT [33]. To estimate the ATE, each individual in the sample must be matched, which means that some individuals will appear more than once in the matched sample [23]. Al-

ternatively, the ATC can be estimated by matching each control individual with a treated individual [23, 25].

Propensity score stratification involves separating individuals into strata (e.g. quintiles), based on their propensity scores [33,41]. The treatment effect is estimated in each of the strata and a weight for each stratum is calculated, corresponding to the size of the stratum. Then a weighted average of the treatment effects is calculated, providing an estimate of the ATE [23]. Estimates of average treatment effects in the treatment and control subgroups, the ATT and ATC, can be obtained by using weights corresponding to the proportions of treated individuals in each stratum, or weights representing the proportions of control individuals respectively [23].

In propensity adjustment, an outcome regression model (e.g. a logistic regression model for a binary outcome) is fitted including treatment and the propensity score as covariates [28, 33], and often also including potential confounders as covariates [42]. The resulting treatment coefficient is often reported as an estimate of the treatment effect. In this report, an extension of the propensity adjustment method will be considered, as follows. Potential outcomes for each individual can be predicted using the outcome regression model with treatment and propensity score as covariates, and the difference between potential outcomes can be calculated for each individual and averaged to estimate the ATE [43]. This method can also be used to estimate the average treatment effect in the treatment group and the average treatment effect in the control group by restricting to the appropriate subset of individuals as required [23].

Another propensity score method that can be used to estimate the causal effect of treatment is inverse probability of treatment weighting (IPTW). IPTW uses the estimated propensity scores as weights to construct 'pseudo-populations' in which the distributions of observed confounders are balanced across treatment groups: the pseudo-population where everyone had treatment and the pseudo-population where everyone had control [28, 36]. The mean outcome in each is calculated, and the

difference between these, provides an estimate of the ATE: [23]

$$\widehat{\text{ATE}}_{\text{IPTW}} = \frac{\sum_i \frac{Y_i Z_i}{\hat{e}_i}}{\sum_i \frac{Z_i}{\hat{e}_i}} - \frac{\sum_i \frac{Y_i(1-Z_i)}{(1-\hat{e}_i)}}{\sum_i \frac{(1-Z_i)}{(1-\hat{e}_i)}} \ . \tag{2.3}$$

Different weights can be applied to construct pseudo-populations which reflect the distribution of observed confounders in the treatment group or control group to obtain the ATT or ATC, respectively [23].

Throughout this thesis, IPTW is used to estimate treatment effects.

2.2 Missing confounder data

So far, the discussion of propensity score methods has assumed that data is fully observed. However, in practice, observational studies suffer from large amounts of missing data. For example, a valuable source of observational data for investigating treatments in routine clinical practice is data from electronic health records (EHRs). Whilst health outcomes and treatment prescriptions are usually well recorded, EHRs tend to suffer from missing data in recording of patient characteristics. This can be seen in the motivating example introduced in Chapter 1, where two key confounders, ethnicity and baseline chronic kidney disease (CKD) stage, each had over 50% of missing data.

Having missing data is problematic as there is a loss of information, or efficiency, from the available data [44]. Missing data can also lead to bias if the assumptions underlying a chosen missing data method are not satisfied [44, 45].

2.2.1 Further notation and concepts related to missing data

We assume throughout that Y and Z are fully observed, and that any missing data is in the confounders — often the case in EHRs. Let R_{ij} be a missing indicator indicating whether the confounder j ($j = 1, \ldots, p$) for patient i ($i = 1, \ldots, n$) is observed ($R_{ij} = 1$) or not ($R_{ij} = 0$). Following D'Agostino and Rubin (2000) and other established missing data literature [45–47], the set of confounder values X_{ij} ($i = 1, \ldots, n; j = 1, \ldots, p$) can be partitioned into those that are observed and those

that are missing, where X_{obs} represents the set of values that are observed and X_{mis} represents the set of values that are missing:

$$X = \{X_{obs}, X_{mis}\} \text{ where } X_{obs} = \{X_{ij}|R_{ij} = 1\} \text{ and } X_{mis} = \{X_{ij}|R_{ij} = 0\}.$$

We will use $R_i = (R_{i1}, ..., R_{ip})$ to refer to the vector of missing indicators for patient i, omitting the subscript i where unambiguous.

We can use missing indicators to define missingness patterns, which are a way of representing the knowledge of which characteristics are observed or unobserved. Subjects can be separated into sets according to the possible combinations of being observed or missing, i.e. the missingness patterns.

Suppose two covariates, A and B, are measured for a group of individuals and that there is missing data present in both. We denote the respective missing indicators as R_A and R_B. In this case, there are four possible combinations of being observed or missing, and hence 4 distinct missingness patterns defined by:

(i) the set of patients for whom both A and B are observed (i.e. $R_A = 1$ and $R_B = 1$),

(ii) the set of patients for whom only A is observed ($R_A = 1$ and $R_B = 0$),

(iii) the set of patients for whom only B is observed ($R_A = 0$ and $R_B = 1$),

(iv) the set of patients for whom neither A nor B are observed ($R_A = 0$ and $R_B = 0$).

2.2.2 Taxonomy of missingness mechanisms

Missingness mechanisms refer to the process by which data become missing, corresponding to the relationship between the reason for missingness in a particular sample and the actual values of the observed and missing data [44]. The most common classification of missingness mechanisms is Rubin's taxonomy, in which data are missing completely at random (MCAR), missing at random (MAR) or missing not at random (MNAR). [20, 44, 48]

The first missingness mechanism, where data are MCAR, means that the probability of being missing does not depend on the observed data or the unobserved data, i.e.

$$P(R|Y, Z, X_{obs}, X_{mis}) = P(R) \tag{2.4}$$
$$\text{or } R \perp Y, Z, X_{obs}, X_{mis} \ .$$

Data are MAR if the probability of being missing depends on the observed values of data but not on the missing values:

$$P(R|Y, Z, X_{obs}, X_{mis}) = P(R|Y, Z, X_{obs}) \tag{2.5}$$
$$\text{i.e. } R \perp X_{mis}|Y, Z, X_{obs} \ .$$

Finally, data are MNAR if the probability of being missing depends on the unobserved data, after conditioning on the observed data:

$$R \not\perp X_{mis}|Y, Z, X_{obs} \ . \tag{2.6}$$

In other words, the probability of being missing depends on the missing value itself.

Information about the missingness mechanisms are an important factor when considering whether an missing data method is appropriate for a particular dataset.

2.2.3 Methods to handle missing data

There has been much methodological research into missing data [44, 45, 49]. Common ad hoc methods for handling missing data include excluding patients with missing data [50] or excluding variables with missing data [51]. Other simple methods include using missing indicators, and replacing missing observations with fixed values [52]; imputing missing values with the mean of observed values [49]; or, in the context of longitudinal studies, imputing missing values by carrying forward the last observation observed [45]. Alternative methods include multiple imputation [44, 53], likelihood-based methods use models based on observed data [45] and

inverse probability weighting [54]. Multiple imputation is an increasingly popular approach to missing data, where missing values are imputed multiple times with plausible values in order to create multiple 'complete' imputed datasets, and results from each dataset are combined using Rubin's Rules to obtain an overall treatment effect estimate [20, 44]. Standard implementations of multiple imputation require data to be MAR [44, 49].

In the context of propensity scores, insights from the general methodological research into missing data cannot directly be used because the aims of regression modelling in propensity score analysis are different (i.e. to achieve balance rather than to estimate parameters [46]) and so the assumptions underlying the validity of missing data methods may be different. Thus a chosen missing data technique may need either stronger SITA-type assumptions or assumptions regarding the missingness mechanism to ensure that using subsequent propensity score methods will achieve balance between treatment groups and obtain valid inferences [51].

2.2.4 Common methods to handle missing confounder data

A common approach to dealing with missing confounders is complete record analysis (CRA), also known as complete case analysis, where individuals with missing data are discarded before analysis. This approach leads to loss of efficiency as information individuals with partial information is discarded. Also, this approach often leads to biased estimates of the treatment effect when missingness depends on both outcome and treatment [50].

The missing indicator approach, a simple method for handling missing confounder data, adds a 'missing' category to partially observed categorical confounders. Equivalently, for continuous confounders, missing values are set to a fixed value, say 0, and both the confounder and its corresponding missing indicator are included in the propensity score model. Although the missing indicator approach has been suggested as a missing data method for propensity score analysis [25, 55], the use of missing indicators is generally considered to be an ad hoc method [56, 57] that yields biased results [58, 59].

A popular alternative for handling missing confounder data is multiple imputation. Similarly to dealing with missing data in general, multiple imputation imputes missing covariates with plausible values several times by drawing from the predictive distribution of the missing covariates given observed data, thus creating a number of imputed datasets. The full analysis (estimation of the propensity score then estimation of the treatment effect) is performed separately in each imputed dataset [60]. The results are then combined using Rubin's rules to obtain an overall estimate of the treatment effect and standard error [20,44,60]. Multiple imputation is very powerful but also can be fairly complex. Guidelines regarding optimal use of multiple imputation in conjunction with propensity score analysis have been proposed [60,61].

Another method that has been proposed is the missingness pattern approach, which incorporates information about the pattern of missing variables into the propensity score. My thesis focuses on this method, which avoids discarding information on individuals with missing confounder data and is relatively simple to understand.

2.2.5 Using missingness patterns to handle missing confounder data in propensity score analysis

Rosenbaum and Rubin (1984) and D'Agostino and Rubin (2000) proposed a generalized propensity score that additionally took into account information on missingness [46, 62]. The generalized propensity score is defined as the probability of being assigned to treatment Z, given the observed covariates X_{obs} and the missing indicator [46]:

$$e^*(X) = P(Z = 1|X_{obs}, R),$$

where $e^*(X)$ denotes the generalized propensity score variable.

Rosenbaum and Rubin (1984) proved that adjusting for the generalized propensity score balances on average the observed covariates and the observed-data indicator (but not the unobserved data) [62], i.e.

$$X_{obs}, R \perp Z|e^*.$$

To estimate the generalized propensity scores, Rosenbaum and Rubin (1984) suggested fitting regression models for $e^*(X)$ in each missingness pattern. Once the estimated scores have been obtained and collected into one variable, the usual propensity score methods can be used for analysis.

D'Agostino and Rubin (2000) stated that, assuming $P(Z|X, Y(0), Y(1), R) = P(Z|X_{obs}, R)$, this missingness pattern approch (MPA) can obtain valid causal inferences and provide unbiased estimates of the ATE [46]. However, no explicit proof is provided.

Mattei (2009) instead give the following assumptions for valid inference from the MPA [63]:

$$\begin{aligned} P(Z|X, Y(0), Y(1), R) &= P(Z|X, R), \\ \text{and either } P(Z|X, R) &= P(Z|X_{obs}, R), \\ \text{or } P(Y(0), Y(1)|X, R) &= P(Y(0), Y(1)|X_{obs}, R). \end{aligned}$$

An extension of the MPA, suggested by D'Agostino et al. [64], is to estimate the propensity scores in each missingness pattern for all subjects with data observed for that pattern, but only retaining the scores for subjects who actually had that specific pattern. Consequently, some subjects are used more than once in the estimation procedure. However, they do not take into account the resulting correlation in the data at the analysis stage.

2.3 Causal diagrams

A variable is defined a confounder if it is a member of some set of variables that is sufficient to control for confounding [39]. When considering what variables could be confounders, it can be helpful to draw a causal diagram to visualise the assumptions being made in a given setting. Causal diagrams are useful for making explicit the assumptions about a scenario's underlying causal structure as well as identifying a set of measured variables (if such a set exists) that would be sufficient to account for confounding.

2.3.1 Introduction to causal diagrams

Causal diagrams are a visual representation of the causal relationships between variables in a scenario of interest and are a useful tool for assessing conditional independence statements under an assumed causal structure [65].

In graphs, nodes represent variables and directed arrows indicate causal relationships between variables. We can visualise direct and indirect relationships between variables by considering paths. A path between two variables is a unbroken sequence of arrows between the variables, irrespective of arrow direction. Different paths between treatment and outcome variables, for example, represent the different ways that treatment is associated with outcome, either directly or via other variables. A directed path is a path where each arrow is in the same direction. We can describe relationships between variables on a path using the concept of descendants. If there is a directed path $P \rightarrow M \rightarrow Q$, we say that M and Q are descendants of P. Directed acyclic graphs are graphs where all arrows are directed and there are no cycles, i.e. directed paths that start and finish at the same variable. Causal directed acyclic graphs, also known as causal diagrams, are directed acyclic graphs which include all variables that are associated with two or more variables already included in the graph.

For illustration, we consider a simplified version of our cohort study, introduced in Chapter 1, looking at the relationship between prescription of ACEI/ARBs and risk of AKI, where baseline CKD stage is the sole confounder (Figure 2.1). In this simplified example, we assume (temporarily for illustrative purposes) that baseline CKD stage is fully observed.

An example of a path in Figure 2.1 is the sequence of arrows from baseline CKD stage to ACEI/ARB prescription, and from ACEI/ARB prescription to AKI. As each arrow in this path is following the same direction, this is a directed path between baseline CKD stage and AKI. In this path, the ACEI/ARB node and the AKI node are descendants of the baseline CKD stage node, and AKI is also a descendant of ACEI/ARB.

In Figure 2.1, there are two paths between treatment and outcome: the direct

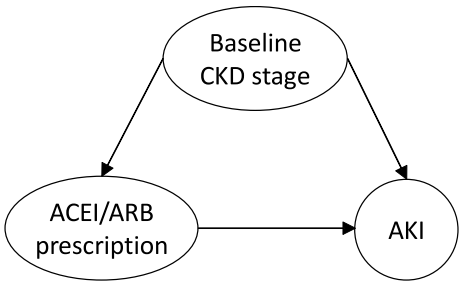


Figure 2.1: A causal directed acyclic graph representing confounding of the relationship between prescription of angiotensin-converting enzyme inhibitors or angiotensin receptor blockers (ACEI/ARBs) and risk of acute kidney injury (AKI), by baseline chronic kidney disease (CKD) stage. All variables are fully observed.

arrow from ACEI/ARB prescription to AKI (the causal effect of interest), and the sequence of arrows from ACEI/ARB prescription to baseline CKD stage, and from baseline CKD stage to AKI.

After creating a causal diagram that represents our scenario of interest, using clinical knowledge of the scenario, we can determine whether conditional independence statements hold in that causal diagram by applying the d-separation rule.

2.3.2 The d-separation rule

The d-separation rule determines if two sets of variables (say A and B) are independent when conditioning on a third set of variables (C) under an assumed causal structure [66]. In order to define the d-separation rule, we first define colliders and blocked paths.

For a particular path in a graph, a variable which has two incoming arrows is called a collider for that path. In Figure 2.1, the AKI node is a collider for the path from baseline CKD stage to ACEI/ARB prescription via AKI (i.e. the two arrows 'collide' at the AKI node).

A path may be blocked in two ways [4,66]: either (i) the path contains a collider that is not in the conditioning set, and does not have any descendants in the conditioning set; or (ii) the path contains a non-collider that is in the conditioning set. If a path from A to B is not blocked, we say that this path is open, in which case, A and B are associated. We say that C 'd-separates' A and B if every path from A to B is blocked by C. If C d-separates A and B, then we have $A \perp B|C$, i.e. A is

conditionally independent of B given C.

2.3.3 Extensions of causal diagrams

Causal diagrams represent relationships between observed variables. However, researchers often need to assess assumptions involving potential outcomes, rather than observed outcomes. For example, the SITA assumption involves $Y(0)$ and $Y(1)$ instead of Y. In order to incorporate potential outcomes into causal diagrams, we consider Richardson and Robin's single world intervention graphs [67] and Balke and Pearl's twin networks [68].

2.3.3.1 Single world intervention graphs

A single world intervention graph is obtained from a causal directed acyclic graph by 'splitting' the treatment variable into two components, separating the random variable Z from the possible fixed values z treatment can take (eg. 0 or 1 for a binary treatment variable) [67]. The random variable part keeps the arrows entering the original variable and the fixed value part keeps the arrows leaving the original variable. A new graph is constructed for each possible treatment value and descendants of the fixed treatment part are relabelled to reflect the effect of that particular treatment. For example, when $z = 0$, Y — a descendant of Z in the original causal diagram — becomes $Y(0)$ and when $z = 1$, $Y = Y(1)$.

Returning to our simplified example in Figure 2.1, we can split the ACEI/ARB prescription variable to obtain the two single world intervention graphs in Figure 2.2. In Figure 2.2a, the ACEI/ARB prescription variable has been split into two hemispheres: the random hemisphere (ACEI/ARB) and the fixed hemisphere representing the fixed value of no prescription. For convenience we can represent all values in a single world intervention template (SWIT), where instead of setting treatment to one particular value, we fix treatment equal to some value z, where z may be any possible treatment value. A SWIT for our simple fully observed example is given in Figure 2.3.

The d-separation rule can be applied to SWITs [67]. For example, we can apply

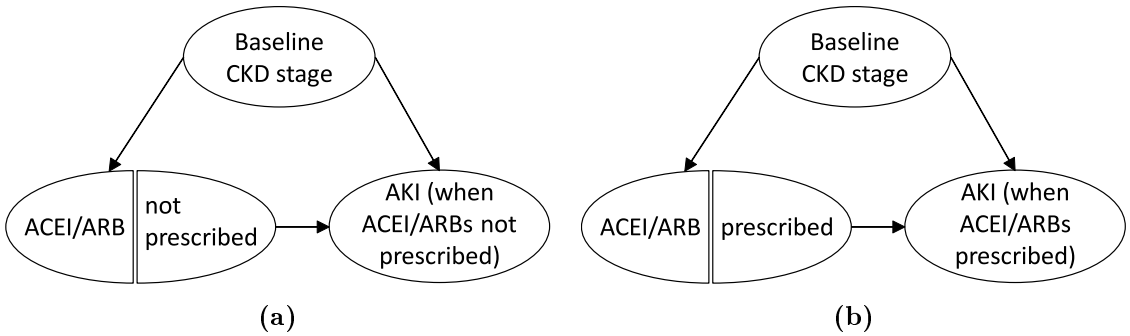


Figure 2.2: The single world intervention graphs resulting from splitting the treatment variable in the graph in Figure 2.1 and intervening to: (a) not prescribe ACEI/ARBs, and (b) prescribe ACEI/ARBs, with all variables fully observed.

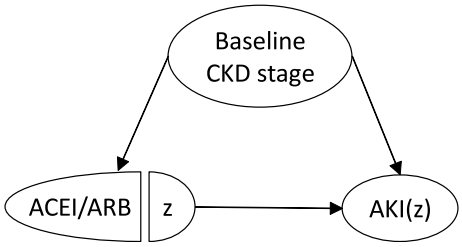


Figure 2.3: The single world intervention template resulting from splitting the treatment variable in the graph in Figure 2.1 and intervening to set the ACEI/ARB variable equal to $z = 0, 1$, where $z = 1$ represents treatment (prescription of ACEI/ARBs) and $z = 0$ represents control (no ACEI/ARB prescription). All variables are fully observed.

the d-separation rule to Figure 2.3 to determine if AKI(z) $\perp$ ACEI/ARB | baseline CKD stage for $z = 0, 1$. The SWIT in Figure 2.3 contains only one path from ACEI/ARB prescription to AKI(z) via baseline CKD stage. Conditioning on baseline CKD stage blocks this path (since it is not a collider). Since the confounder d-separates the treatment and outcome, ACEI/ARB prescription is conditionally independent of the corresponding potential AKI outcome, given baseline CKD stage. Note that we cannot actually assess the SITA assumption as previously defined, as each graph includes only one of the two outcomes, whereas the SITA assumption involves the joint distribution of $Y(1)$ and $Y(0)$ — we are considering a weaker version of the SITA assumption:

$$Z \perp Y(z)|X \text{ for } z = 0, 1. \tag{2.7}$$

The SITA assumption implies our weaker version in equation (2.7), but the converse is not true [26].

2.3.3.2 Twin networks

Researchers may sometimes need to assess whether conditional independence statements involving both observed and counterfactual values of a variable are satisfied in a particular scenario. In order to do this, an alternative to SWITs called twin networks can be used, as described by Balke and Pearl (1994) [68] and Shpitser and Pearl (2007) [69]. In brief, a twin network can be constructed from a directed acyclic graph, which involves real world variables and relationships, by adding counterparts of variables and relationships in the counterfactual world where treatment has been intervened upon to be set to some realisation of the random variable Z.

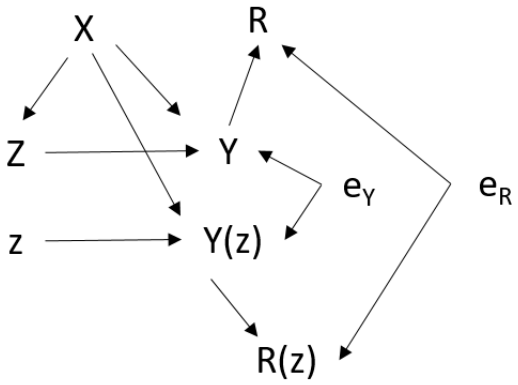


Figure 2.4: A simple twin network.
X: partially observed confounder. Z: observed treatment allocation. Y: observed outcome. $Y(z)$: potential outcome resulting from intervening to set treatment to value z. R: observed missingness indicator (=1 if X observed, =0 if X is missing). $R(z)$: potential missingness indicator (=1 if X observed in counterfactual world, =0 if X is missing in counterfactual world). e_Y: unobserved error term between Y and $Y(z)$. e_R: unobserved error term between R and $R(z)$.

Figure 2.4 gives an example of a twin network in a simple scenario with a single partially observed confounder, where treatment has a causal effect on missingness and there are no unobserved common causes with missingness and other variables. We can determine apply the d-separation rule to twin networks in the same way as SWITs, conditioning only on the variables in the conditioning set of the assumptions.

2.4 The MPA in the literature

Before investigating the statistical properties of the MPA and providing practical recommendations for its use, it is important to understand when and how researchers currently use this approach. Therefore, I initially planned to perform a systematic review of the literature to provide a description of the methods used by researchers to address the issue of missing confounder data in epidemiological studies analysed using propensity scores. Specific objectives were to:

- Estimate the proportion of papers reporting the use of the MPA — and alternative approaches— for propensity score analysis when some confounders are partially observed
- Assess whether the assumptions underlying the validity of the methods employed were explicitly stated, and their plausibility discussed
- Among papers reporting the use of the MPA, describe the implementation of the method (e.g. standard vs D'Agostino's extension of the approach [64], missingness patterns pooling [70]) and the method used to estimate the variance of the treatment effect.

However, a systematic review with a similar scope was published by Malla et al. [71] in the Journal of Comparative Effectiveness Research while I was at the screening stage. To avoid duplication, it has been decided not to perform the systematic review, but the protocol and the search algorithm are detailed in Appendix B together with the results of the screening strategy.

Malla et al. screened Embase (OvidSP) and Medline (OvidSP) to retrieve publications using propensity score methods in comparative effectiveness research between 1 January 2007 and 30 June 2017. Of the 167 papers included in their systematic review, 118 (71%) retrospectively analysed routine datasets, which emphasises the importance of the development of guidance for handling missing data in this setting.

Although missing data are almost systematic in routinely collected data, only

86 articles (51%) provided information about how missing data were handled, and 62 (37%) reported the amount of missing data. Among papers reporting the use of a missing data method, the most popular approach was complete record analysis (n=53 (62%)), followed by multiple imputation (n=16 (19%)). These results were expected given that complete record analysis is also the most common approach to handle missing covariate data in multivariable regression. As for multiple imputation, it has been the focus of several methodological papers in the context of propensity score analysis in the past few years [60, 61, 72], confirming an increasing interest for this method.

The MPA was used in 3 articles only (3%) and the related missing indicator approach was reported in 1 article. The remaining papers reported the use of a range of single imputation methods. This distribution confirms the scarcity of studies using the MPA in practice. Furthermore, the reasons for missingness were discussed in 12 papers, among which only 3 linked these reasons to the missingness mechanism. This highlights a suboptimal reporting of missing data and missing data methods for propensity score analysis, despite the availability of the STROBE guidelines for the reporting of observational studies which includes specific items related to missing data [73]. This could be explained by the difficulty to assess the assumptions underlying the validity of the different approaches in real-life scenarios, especially for the MPA, since its assumptions have not yet been explored in practice. Causal diagrams could facilitate this investigation and encourage researchers to carefully consider and report the reasons for missingness and the rationale for the choice of the missing data methods.

Chapter 3

Understanding the assumptions underlying the missingness pattern approach

One potential explanation why the MPA is not often used by researchers is that strategies for assessing assumptions underlying the MPA in practice have not been discussed in prior literature. By using the causal diagrams introduced in Chapter 2 to visually represent scenarios of interest and the assumptions that we have made, we can apply the d-separation rule to determine whether the MPA's assumptions hold in a particular clinical scenario.

In this chapter, I explore the connections between the MPA and prior literature. I then discuss how my use of causal diagrams evolved, driven by the nuances of the MPA's assumptions.

3.1 Connections between the MPA's assumptions and prior literature

In Chapter 2, our statement of the MPA's assumptions follows Mattei's statement of assumptions sufficient for valid inference using the MPA [63]. Whereas Mattei (2009) used conditional independence statements that hold jointly for the potential

outcomes, we use weaker statements that hold separately for each potential outcome. D'Agostino and Rubin (2000) instead asserted that, under the following assumption, using generalized propensity scores can provide unbiased estimates of the ATE [46].

$$P(Z|X, Y(0), Y(1), R) = P(Z|X_{obs}, R) \ . \tag{3.1}$$

We can also express Mattei's assumptions (i.e. the stronger versions of the mSITA, CIT and CIO assumptions) using conditional probabilities:

$$\text{strong mSITA: } P(Z|X, Y(0), Y(1), R) = P(Z|X, R) \tag{3.2}$$

$$\text{CIT: } P(Z|X, R) = P(Z|X_{obs}, R) \tag{3.3}$$

$$\text{strong CIO: } P(Y(0), Y(1)|X, R) = P(Y(0), Y(1)|X_{obs}, R) \tag{3.4}$$

Note that Mattei's 'stronger' version of the CIT assumption (3.3) is the same as our CIT assumption, since it does not involve potential outcomes.

Substituting the right-hand side of equation (3.3) into the right-hand side of equation (3.2) gives the assumption in equation (3.1) as defined by D'Agostino and Rubin [46]. Thus Mattei's statement of the mSITA and CIT assumption imply the assumption given by D'Agostino and Rubin (2000).

Furthermore, assumptions (3.2) and (3.4) can hold whilst assumption (3.1) is violated. Thus Mattei (2009) gives a wider set of assumptions than D'Agostino and Rubin (2000) under which the MPA can give valid inference.

Other work exploring non-systematic monitoring of time-varying covariates [74, 75] suggest a version of the "no unmeasured confounding assumption" which, in the single time-point exposure setting, can be written as:

$$Z \perp Y(z)|X_{obs}, R. \tag{3.5}$$

If the D'Agostino and Rubin assumption (3.1) holds, then assumption (3.5) holds. Furthermore, if either the mSITA and CIT assumptions hold, or the mSITA and CIO assumptions hold, then assumption (3.5) holds. The mSITA, CIT and CIO

assumptions can be seen as a wider set of assumptions under which variants of missingness-pattern-type approaches can produce valid inference. All three sets of assumptions seem likely to be satisfied in a setting where missing confounder values are unavailable to the individual making the treatment decision and thus do not affect treatment. However, only Mattei's statement of the assumptions, with CIT and CIO as two separate sub-assumptions, makes it clear that there is another quite different set of scenarios in which missingness-pattern-type methods may provide valid inference.

3.2 Using weaker versions of the MPA's assumptions

Mattei (2009) states three assumptions under which the MPA leads to valid inference [63]. I present weaker versions of these assumptions and prove that, under these assumptions, the MPA still gives a consistent estimator of the ATE. The first assumption is an extension of the SITA assumption (equation (2.1)), which I call the Missingness Strongly Ignorable Treatment Assignment (mSITA) assumption due to its similarities with the SITA assumption (equation (2.1)):

$$\text{mSITA:} \quad Z \perp Y(z)|X, R \quad \text{for} \quad z = 0, 1. \tag{3.6}$$

A key difference comparing assumption (3.6) with the weaker version of equation (2.1) is the inclusion of information about the missingness pattern, represented by R, in the conditioning set. We assume that SITA holds in the full data, thus this assumption states that additionally conditioning on R does not introduce bias.

I call the two further assumptions [63]: the conditionally independent treatment (CIT) assumption and the conditionally independent outcomes (CIO) assumption.

$$\text{CIT:} \quad Z \perp X_{mis}|X_{obs}, R$$

$$\text{CIO:} \quad Y(z) \perp X_{mis}|X_{obs}, R \quad \text{for} \quad z = 0, 1.$$

If mSITA holds, and either CIT or CIO holds, then the MPA provides a consistent estimate of the treatment effect. I refer to these assumptions as the 'MPA's assumptions'. To prove that the weaker versions of the MPA's assumptions lead to a consistent estimator of the ATE, I first show that, under these assumptions, $E[(1-Z)Y/(1-e^*)] = E[Y(0)]$.

First, using the consistency assumption and conditioning on the missing indicator and observed confounder values, we have that:

$$\begin{aligned} E\left[\frac{ZY}{e^*}\right] &= E\left[\frac{ZY(1)}{e^*}\right] \\ &= E\left[E\left[\frac{ZY(1)}{e^*}|X_{obs}, R\right]\right] \\ &= E\left[\frac{1}{e^*}E\big[ZY(1)|X_{obs}, R\big]\right]. \end{aligned} \tag{3.8}$$

Switching briefly to summation notation:

$$\begin{aligned} &E\big[ZY(1)|X_{obs}, R\big] \\ &= \sum_z \sum_y zyP(Z=z|X_{obs}, R)P(Y(1)=y|Z, X_{obs}, R) \\ &= \sum_z \sum_y zyP(Z=z|X_{obs}, R) \sum_x P(Y(1)=y, X_{mis}=x|Z, X_{obs}, R) \\ &= \sum_z \sum_y \sum_x zyP(Z=z|X_{obs}, R)P(Y(1)=y|Z, X_{mis}, X_{obs}, R) \\ &\qquad \times P(X_{mis}=x|Z, X_{obs}, R) \end{aligned}$$

Using mSITA ($Z \perp Y(z)|X, R$ for $z = 0, 1$) and CIT ($Z \perp X_{mis}|X_{obs}, R$), we have that $P(Y(1)=y|Z, X_{mis}, X_{obs}, R) = P(Y(1)=y|X_{mis}, X_{obs}, R)$ and $P(X_{mis}=x|Z, X_{obs}, R) = P(X_{mis}=x|X_{obs}, R)$ respectively. Hence:

$$\begin{aligned} &E\big[ZY(1)|X_{obs}, R\big] \\ &= \sum_z \sum_y \sum_x zyP(Z=z|X_{obs}, R)P(Y(1)=y|X_{mis}, X_{obs}, R)P(X_{mis}=x|X_{obs}, R) \\ &= \sum_z \sum_y zyP(Z=z|X_{obs}, R) \sum_x P(Y(1)=y, X_{mis}=x|X_{obs}, R) \end{aligned}$$

$$= \sum_z \sum_y zyP(Z = z|X_{obs}, R)P(Y(1) = y|X_{obs}, R)$$
$$= \sum_z zP(Z = z|X_{obs}, R) \sum_y yP(Y(1) = y|X_{obs}, R)$$
$$= E\big[Z|X_{obs}, R\big]E\big[Y(1)|X_{obs}, R\big]$$

We can also show that $E\big[ZY(1)|X_{obs}, R\big] = E\big[Z|X_{obs}, R\big]E\big[Y(1)|X_{obs}, R\big]$ using mSITA with CIO ($Y(z) \perp X_{mis}|X_{obs}, R$ for $z = 0, 1$):

$$\begin{aligned}
&E\big[ZY(1)|X_{obs}, R\big] \\
&= \sum_z \sum_y zyP(Y(1) = y|X_{obs}, R)P(Z = z|Y(1), X_{obs}, R) \\
&= \sum_z \sum_y zyP(Y(1) = y|X_{obs}, R) \sum_x P(Z = z, X_{mis} = x|Y(1), X_{obs}, R) \\
&= \sum_z \sum_y \sum_x zyP(Y(1) = y|X_{obs}, R)P(Z = z|Y(1), X_{mis}, X_{obs}, R) \\
&\qquad \times P(X_{mis} = x|Y(1), X_{obs}, R)
\end{aligned}$$

Under mSITA, $P(Z = z|Y(1), X_{mis}, X_{obs}, R) = P(Z = z|X_{mis}, X_{obs}, R)$, and under CIO, $P(X_{mis} = x|Y(1), X_{obs}, R) = P(X_{mis} = x|X_{obs}, R)$. Hence:

$$\begin{aligned}
&E\big[ZY(1)|X_{obs}, R\big] \\
&= \sum_z \sum_y \sum_x zyP(Y(1) = y|X_{obs}, R)P(Z = z|X_{mis}, X_{obs}, R)P(X_{mis} = x|X_{obs}, R) \\
&= \sum_z \sum_y zyP(Y(1) = y|X_{obs}, R) \sum_x P(Z = z, X_{mis} = z|X_{obs}, R) \\
&= \sum_z \sum_y zyP(Y(1) = y|X_{obs}, R)P(Z = z|X_{obs}, R) \\
&= \sum_z zP(Z = z|X_{obs}, R) \sum_y yP(Y(1) = y|X_{obs}, R) \\
&= E\big[Z|X_{obs}, R\big]E\big[Y(1)|X_{obs}, R\big]
\end{aligned}$$

Thus, the MPA's assumptions enable us to rewrite equation 3.8 as follows:

$$E\left[\frac{ZY}{e^*}\right] = E\left[\frac{1}{e^*}E\big[Z|X_{obs}, R\big]E\big[Y(1)|X_{obs}, R\big]\right].$$

Since $e^* = E[Z|X_{obs}, R]$:

$$E\left[\frac{ZY}{e^*}\right] = E\left[E\big[Y(1)|X_{obs}, R\big]\right] = E[Y(1)].$$

This proof is reproduced in Appendix A of the research paper pre-print in Chapter 5.

It can similarly be shown that $E[(1-Z)Y/(1-e^*)] = E[Y(0)]$ under the MPA's assumptions.

3.2.1 The MPA's connection to Rubin's taxonomy of missing data

The assumptions underlying the MPA are separate from Rubin's taxonomy of missing data (i.e. classification of missingness mechanisms into: missing completely at random, missing at random, and missing not at random [20]) in the sense that classifying data according to Rubin's taxonomy does not give us any information as to whether the MPA's assumptions would hold. For instance, one might intuitively expect that if data are missing completely at random, the MPA's assumptions would hold, as many missing data methods are appropriate under this assumption. However this is not true, as can be seen in the counterexample in Figure 3.1a where, although the confounder data is missing completely at random since the observed-confounder indicator (R) is not affected by any other variables, the MPA is not appropriate because the confounder values directly affect both treatment and outcome for patients with $R = 0$, violating both the CIT and CIO assumptions.

Conversely, one might expect that having confounder data missing not at random would mean that the MPA is not appropriate. This is also not true, as can be seen in the counterexample in Figure 3.1b. In this graph, missingness of the confounder depends on the confounder itself and so data are missing not at random. Applying d-separation to this scenario, we find that the mSITA assumption holds (since Z and $Y(z)$ are not associated given X and R) and that the CIT assumption holds (since Z and X_{mis} are not associated given X_{obs} and R), and thus the MPA is

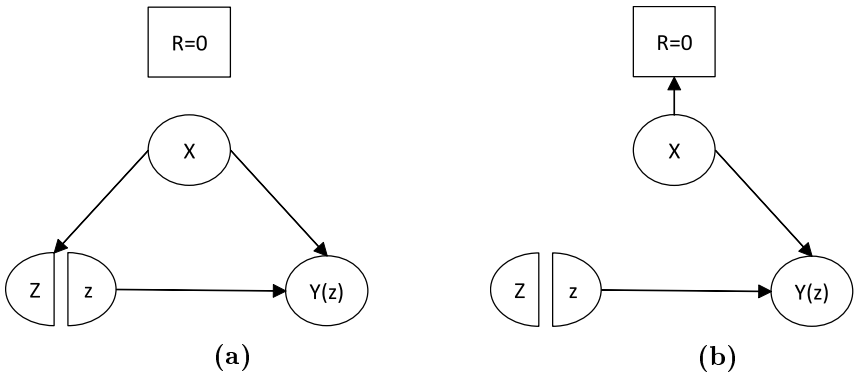


Figure 3.1: (a) An example of a single world intervention graph template, conditioning on $R = 0$, under a missing completely at random mechanism. In this example, the MPA's assumptions do not hold. (b) An example of a single world intervention graph template, conditioning on $R = 0$, under a missing not at random mechanism. In this example, the MPA's assumptions hold.

appropriate here. Despite the data being missing not at random, since the mSITA and CIT assumptions hold here, the MPA is appropriate in this scenario. Thus, classification of the missingness mechanism according to Rubin's taxonomy does not provide information as to whether the MPA's assumptions will hold; instead, the plausibility of the MPA's assumptions depends on which relationships between variables exist in the subgroup of patients with missing confounder values.

3.3 The evolution of causal diagrams for assessing the MPA's assumptions

In Section 2.3.3.1, we demonstrated how to apply d-separation in a simple single world intervention template (SWIT) representing a simplified version of the motivating example with a single fully observed confounder, baseline chronic kidney disease (CKD) stage (Figure 2.3). We now consider the situation where baseline CKD stage is partially observed, letting R_{CKD} denote the corresponding missing indicator variable.

By simply incorporating R_{CKD} as another variable in the SWIT, we obtain the SWIT in Figure 3.2, which additionally assumes that R_{CKD} is associated with ACEI/ARB prescription via an unobserved common cause (denoted U). Applying d-separation to this template will allow us to assess the mSITA assumption, i.e.

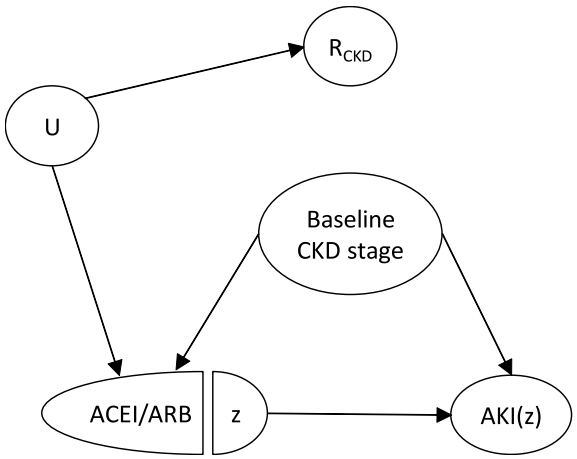


Figure 3.2: Extension of the single world intervention template in Figure 2.3 where the missingness of baseline chronic kidney disease stage is associated with treatment. R_{CKD} denotes the missing indicator for baseline CKD stage. U denotes an unobserved common cause of R_{CKD} and treatment. $z = 0, 1$, where $z = 1$ represents prescription of ACEI/ARBs and $z = 0$ represents no ACEI/ARB prescription.
ACEI/ARB: angiotensin-converting enzyme inhibitor or angiotensin receptor blocker, AKI: acute kidney injury. CKD: chronic kidney disease.

to determine whether — after conditioning on baseline CKD stage and R_{CKD} — ACEI/ARB prescription and AKI(z) are conditionally independent (for $z = 0, 1$). Since the only path from ACEI/ARB to AKI(z) in Figure 3.2 is blocked by baseline CKD stage and R_{CKD}, the mSITA assumption holds.

However, we cannot use the SWIT in Figure 3.2 to check whether the CIT and CIO assumptions hold, as these assumptions involve the observed and missing values of the confounder separately. Indeed, a necessary condition for one of the CIT and CIO assumptions to hold is that baseline CKD stage is a confounder only when it is observed, and so relationships between confounder values and treatment or outcome must differ depending on whether the confounder values are observed or missing.

We considered two strategies to be able to assess the CIT and CIO assumptions in SWITs. We initially split the confounder node into two separate nodes: a node representing the observed baseline CKD stage values and a node representing the missing baseline CKD stage values. Our second strategy was to construct a SWIT restricted to the missingness pattern for the subgroup of patients with missing values.

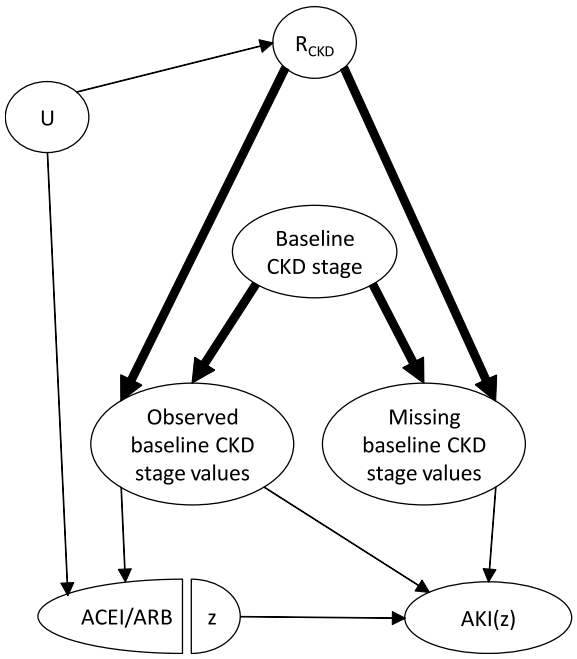


Figure 3.3: Extension of the single world intervention template in Figure 3.2 where the missingness of baseline CKD stage is associated with treatment, modified to separate the observed and missing components of baseline CKD stage. R_{CKD} denotes the missing indicator for baseline CKD stage. U denotes an unobserved common cause of measurement of baseline CKD stage (i.e. R_{CKD}) and treatment. $z = 0, 1$, where $z = 1$ represents prescription of ACEI/ARBs and $z = 0$ represents no ACEI/ARB prescription. Bold arrows indicate deterministic relationships.
ACEI/ARB: angiotensin-converting enzyme inhibitor or angiotensin receptor blocker, AKI: acute kidney injury. CKD: chronic kidney disease.

3.3.1 SWITs with separate confounder nodes

Our first strategy incorporated the observed and missing baseline CKD stage values separately into the SWIT, using bold arrows to represent their deterministic relationships with the full baseline CKD stage variable and R_{CKD}. For example, the bold arrows in Figure 3.3 indicate that the variables representing observed and missing values are each fully determined by R_{CKD} and the baseline CKD stage variable by construction. Figure 3.3 also encodes the assumption that the missing values of baseline CKD stage do not affect prescription of ACEI/ARBs (represented by the absence of an arrow from the missing baseline CKD stage node to the ACEI/ARB node). This seems plausible as when baseline CKD stage is not available to the General Practitioner, it cannot be used to decide whether or not to prescribe ACEI/ARBs. We also assume that the missing values of baseline CKD stage directly affect risk of AKI (represented by the presence of an arrow from the missing baseline CKD stage

node to the AKI node) since baseline CKD stage is associated with risk of kidney disease, whether it is observed or missing.

When using SWITs with separate confounder nodes, the presence of deterministic relationships means that extra conditional independencies hold that are not implied by the graph. Thus d-separation is not complete: d-separation does not identify all possible conditional independencies [67, 76]. In our simplified example, this means that if, for instance, we determine that treatment and the missing confounder values are not d-separated in the SWIT (given the observed confounder values and the observed-confounder indicator), then we would conclude that CIT does not hold. However, lack of completeness implies the CIT assumption might still hold. Consequently, caution must be exercised when we do not find that a particular conditional independence holds, and we must consider the deterministic relationships and clinical knowledge to decide if the assumption truly does or does not hold.

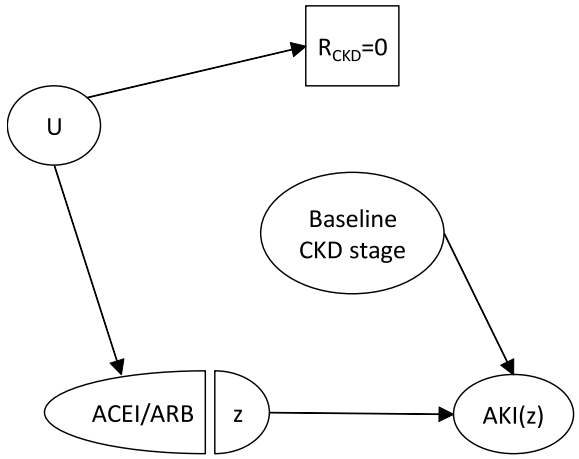


Figure 3.4: Extension of the single world intervention template in Figure 3.2 where the missingness of baseline CKD stage is associated with treatment, restricted to the missingness pattern with missing baseline CKD stage values (i.e. $R_{\text{CKD}} = 0$). U denotes an unobserved common cause of measurement of baseline CKD stage (i.e. R_{CKD}) and treatment. $z = 0, 1$, where $z = 1$ represents prescription of ACEI/ARBs and $z = 0$ represents no ACEI/ARB prescription.
ACEI/ARB: angiotensin-converting enzyme inhibitor or angiotensin receptor blocker, AKI: acute kidney injury. CKD: chronic kidney disease.

3.3.2 SWITs by missingness pattern

In order to avoid the difficulties in applying d-separation to SWITs with deterministic arrows, we developed an alternative strategy for constructing SWITs to be used for assessing the CIT and CIO assumptions. Instead of splitting the confounder node, we constructed a SWIT for the subgroup of patients with missing confounder values (i.e. for the missingness pattern $R = 0$), using a square node to denote restriction to this missingness pattern (Figure 3.4).

We can apply the d-separation rule to this modified SWIT in the same way as a normal SWIT. This will allow us to assess the CIT and CIO assumptions. To explain this, we can rewrite the CIT and CIO assumptions separately for each missingness pattern in a simple situation with a single partially observed confounder X. For the subgroup of patients with X observed, the CIT and CIO assumptions, $Z \perp \emptyset | X, R = 1$, and $Y(z) \perp \emptyset | X_{obs}, R = 1$, respectively, are trivially true because X_{mis} is empty ($= \emptyset$) given $R = 1$. In the subgroup of patients with X missing, the assumptions become: $Z \perp X | R = 0$, and $Y(z) \perp X | R = 0$, respectively. Thus, we can construct SWITs restricted to the missingness pattern $R = 0$ in order to assess the CIT and CIO assumptions under an assumed causal structure.

In Figure 3.4, the only path connecting Z and X passes through $Y(z)$, a collider on the path; thus applying the d-separation rule shows that Z and X are conditionally independent (in the subgroup with $R = 0$). Hence the CIT assumption holds. However, the CIO assumption does not hold as there is a direct arrow from X to $Y(z)$. Thus, in this example, the mSITA and CIT assumptions hold and hence the MPA can obtain valid inference.

We have demonstrated how to construct SWITs for a given scenario and applied the d-separation rule to the SWITs to determine whether the mSITA, CIT and CIO assumptions were plausible. We now develop guidance to investigate when the MPA's underlying assumptions hold.

Chapter 4

Guidance for assessing the assumptions underlying the missingness pattern approach

A key purpose of this thesis was to create guidance for assessing the assumptions under which valid inference can be obtained from missing confounder methods incorporating missingness information. In this section, I discuss the process of developing this guidance. We have developed guidance for researchers seeking to decide whether the MPA's assumptions are plausible in a particular clinical setting, and thus whether the MPA is an appropriate method for dealing with missing confounder data in the setting of interest. We developed our guidance by considering a variety of scenarios and applying d-separation to SWITs representing each of these scenarios.

Initially, we developed a framework in the form of a decision-tree (Figure 4.1). The intended purpose of this framework was to elucidate the clinical assumptions underlying the validity of the MPA. In this early framework, we considered the temporal order of the missingness relative to the confounder, treatment and outcome variables, however, we found that this was too restrictive. Our current guidance, given in a step-by-step format, instead focuses on considering the plausibility of key violations and constructing a causal diagram to help assess the validity of the assumptions. We will first describe the development of the decision-tree framework.

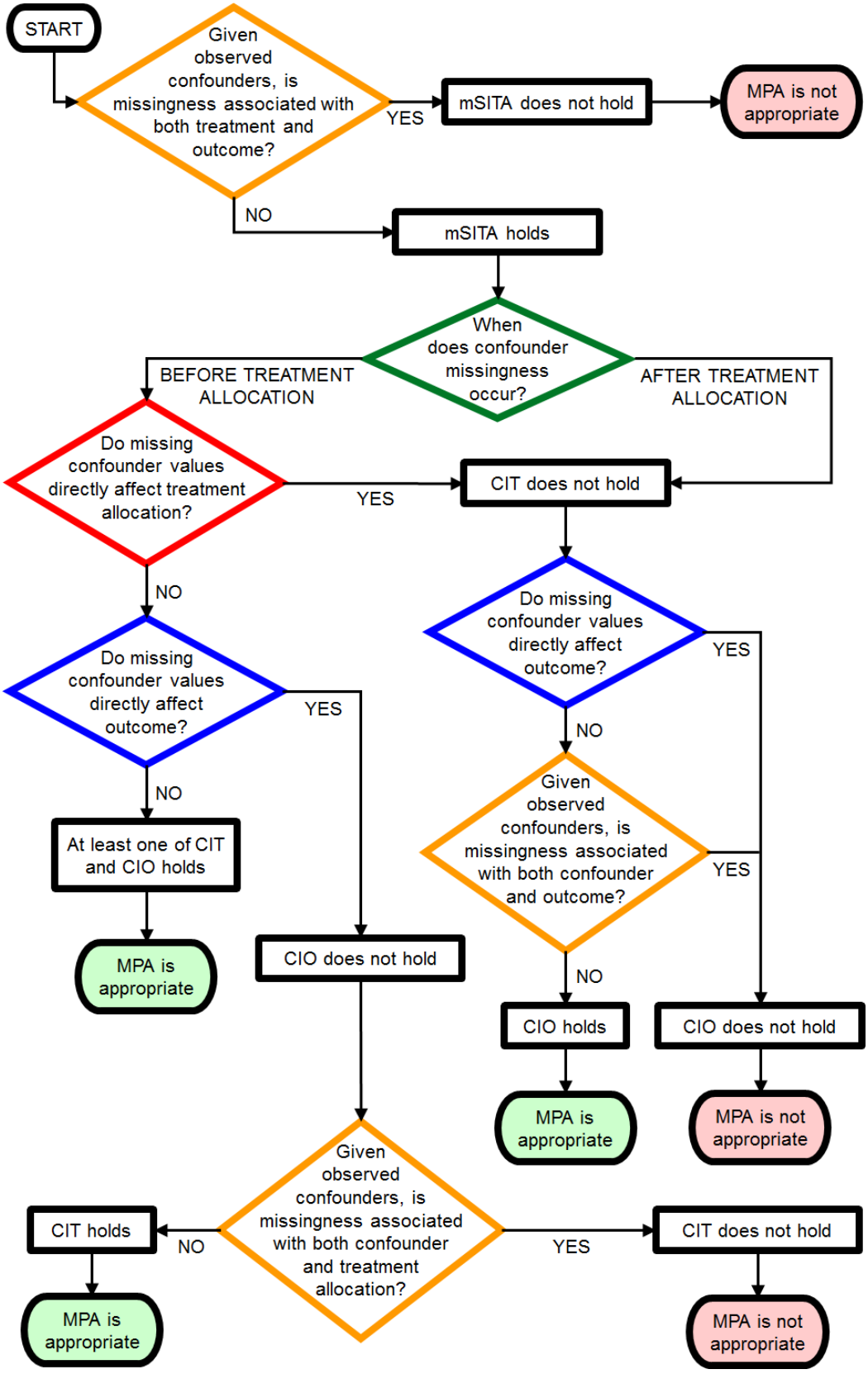


Figure 4.1: Framework to decide if the missingness pattern approach (MPA) is appropriate when treatment allocation and outcome are fully observed and there is one partially observed confounder.
mSITA: missingness strongly ignorable treatment allocation assumption. CIT: conditionally independent treatment assumption. CIO: conditionally independent outcomes assumption.

4.1 Development of the early framework for investigating the MPA's assumptions in practice

Our initial attempts to develop guidance were restricted to settings with a fully observed binary treatment, a fully observed outcome and a single partially observed confounder.

In these initial explorations, to assess the mSITA assumption, we applied d-separation to conventional SWITs. In order to assess the CIT and CIO assumptions, we applied d-separation to our modified SWITs which incorporated the observed and missing parts of the confounder separately and thus involve deterministic relationships.

We considered a variety of scenarios to allow us to produce general conclusions about when the MPA's underlying assumptions can be expected to hold. In order to achieve this, we considered three key issues, namely: (i) relationships between missing confounder values and treatment or outcome, (ii) temporal order of variables, and (iii) relationships between missingness and confounder, treatment or outcome.

4.1.1 Relationships between missing confounder values and treatment or outcome

Recall that, in order for at least one of the CIT and CIO assumptions to hold, we require X to be a confounder only when it is observed. Thus, for the MPA to be an appropriate method, relationships between confounder values and either treatment or outcome must differ depending on whether the confounder values are observed or missing.

4.1.2 Temporal order of variables

A requirement for treatment to have a causal effect on outcome is that treatment temporally precedes outcome [34]. We have also been implicitly assuming that our

confounders precede treatment. We now consider where missingness appears in this temporal order.

Missingness is often considered to be external to the causal structure, i.e. the relationships between confounders, treatment and outcome are the same whether or not the values of these variables are measured. However, as previously mentioned, for the CIT or CIO assumptions to hold we require that relationships between treatment or outcome with confounder values must differ depending on missingness. We assume that: for the relationship between treatment and confounder values to differ by missingness, R must occur temporally prior to Z, and for the relationships between outcome and confounder values to differ by missingness, R must occur prior to $Y(z)$. Thus, we assume that R always occurs temporally before $Y(z)$.

Furthermore, we assume that R occurs after X, since a patient's stage of CKD at the baseline visit exists prior to the assessment of that stage being made (or not made).

4.1.3 Relationships between missingness and confounder, treatment or outcome

Missingness can be associated with values of the confounder, treatment and outcome. When missingness is considered to be external to the causal structure, such association is thought to arise either through the variables directly causing the missingness, or via shared common causes of missingness and the variables under study. In our case, as discussed above, missingness is part of the causal structure and can affect the values of, and the relationships between, study variables. This creates a third possibility: that association arises due to missingness affecting confounder, treatment or outcome values. Thus the way in which the association between missingness and study variables arises can impact on the validity of the MPA's assumptions.

There are three ways in which associations between missingness and treatment may arise. First, missingness has a direct effect on treatment (when missingness occurs temporally prior to treatment). Second, treatment has a direct effect on missingness (when treatment precedes missingness). Third, treatment and missing-

ness share a common cause.

There are two ways in which associations may arise between missingness and confounder values (under our assumption that R occurs temporally after X). First, the confounder values have a direct effect on missingness. Second, X and R share a common cause.

There are two ways in which associations may arise between missingness and outcome (under our assumption that missingness temporally precedes outcome). First, missingness has a direct effect on outcome. Second, missingness and outcome share a common cause.

Hence we need to carefully consider the relationships between missingness and each of the confounder, treatment and outcome when deciding whether the MPA's assumptions seem plausible.

4.1.4 Application of the early framework to the illustrative example

To demonstrate how the framework can be used in practice, we now apply the framework in Figure 4.1 to the cohort study described in Section 1.1.

Given observed confounders, is missingness associated with both treatment and outcome?

The first decision in the framework is to decide whether or not missingness is associated with both treatment and outcome, given observed confounders. In other words, to decide if we expect the missingness reason to have unobserved common causes with each of treatment and outcome.

CKD stage: Missingness in baseline CKD stage is more likely for patients expected to have a higher risk of kidney disease due to age, chronic comorbidities or prescription of medicines that may interfere with renal function. Whilst these risk factors are associated with missingness and treatment or outcome, they are already captured and accounted for in the electronic health records. So with re-

spect to baseline CKD stage, it seems plausible that missingness is not associated with both treatment and outcome, given observed confounders.

Ethnicity: With respect to ethnicity, missingness may be caused by service-level factors such as the circumstances at the time patients are admitted [13]. We believe that these factors are unlikely to also be determinants of treatment or risk factors for AKI, and so it seems plausible that missingness of ethnicity is not associated with both treatment and outcome.

Thus we can follow the "No" arrow from the first yellow decision box in Figure 4.1, finding that mSITA is expected to hold for our case study.

When does confounder missingness occur?

CKD stage: CKD stage, a measure of kidney function, was defined at baseline, i.e. prior to treatment by definition.

Ethnicity: Ethnicity is also recorded (or not recorded) before treatment is allocated, for example when patients register at a general practice.

So we follow the "Before treatment allocation" arrow to the red decision box.

Do missing confounder values directly affect treatment allocation?

CKD stage: If baseline CKD stage is not available, this unobserved information cannot be used to determine the General Practitioner's treatment decision whether or not to prescribe ACEI/ARBs.

Ethnicity: If the General Practitioner believes that ethnicity is important, it is more likely to be recorded.

It seems plausible that baseline CKD stage and ethnicity affect treatment allocation only when they are measured and so we follow the "No" arrow in the decision-tree to the left-hand blue decision box.

Do missing confounder values directly affect outcome?

CKD stage: Baseline CKD stage (whether observed or missing) is associated with risk of kidney disease.

Since at least one of the confounders affect outcome even when not measured, we follow the "Yes" arrow from the left-hand blue decision box and we find that here, we do not expect the CIO to hold.

Given observed confounders, is missingness associated with both confounder and treatment allocation?

Again, we expect that common causes of missingness and prescription of ACEI/ARBs are already accounted for in our analysis and thus we follow the "No" arrow from the lowest yellow decision box, finding that we expect that the CIT assumption holds.

Result of applying the framework to the cohort study: Based on the assumptions regarding the clinical scenario described above, we conclude that it is plausible that (although the CIO assumption does not hold) the mSITA and CIT assumptions hold and thus the MPA is appropriate.

4.2 Current guidance for assessing the MPA's assumptions in practice

In our early guidance, we considered the temporal order of the missingness relative to the confounder, treatment and outcome variables, considering settings where missingness occurs after confounder and before outcome. However, this was too restrictive and did not allow for scenarios where data is collected retrospectively but the reason for missingness affected treatment assignment and/or outcome. Another limitation of the early framework was ambiguity in Figure 4.1 regarding the text descriptions of violations that could occur. This early guidance was developed using single world intervention graph templates where the confounder node is split into its observed and missing components.

When developing our current guidance, we decided against using graphs with separate missing and observed confounder nodes, in order to avoid difficulties when applying the d-separation rule in the presence of deterministic relationships, and chose to use graphs which condition on missingness pattern. This change in the type of causal diagram used to develop guidance lead to a change in the presentation of the guidelines, resulting in a step-by-step format. In our current guidance, we forego the temporal restriction, and focus on considering the plausibility of key violations and constructing causal diagrams to help assess the validity of the assumptions. We also endeavoured to provide clearer explanations of violations, aided by causal diagrams. This guidance is given in the Chapter 5 pre-print, in Section 5.9.

Chapter 5

Research paper: Propensity scores using missingness pattern information: a practical guide

Helen A. Blake[1,2], Clémence Leyrat[1,3], Kathryn E. Mansfield[3], Shaun Seaman[4], Laurie A. Tomlinson[3], James Carpenter[1,5], and Elizabeth J. Williamson[1,6]

1. Department of Medical Statistics, London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT, UK
2. Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine, 15-17 Tavistock Place, London, WC1H 9SH, UK
3. Department of Non-communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT, UK
4. MRC Biostatistics Unit, School of Clinical Medicine, Cambridge, UK
5. MRC Clinical Trials Unit at UCL, Institute of Clinical Trials and Methodology, 90 High Holborn, London, WC1V 6LJ, UK
6. Health Data Research UK, 215 Euston Road, London, NW1 2BE, UK

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT
www.lshtm.ac.uk

Registry
T: +44(0)20 7299 4646
F: +44(0)20 7299 4656
E: registry@lshtm.ac.uk

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

RESEARCH PAPER COVER SHEET

PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.

SECTION A – Student Details

Student	Helen A Blake
Principal Supervisor	Elizabeth Williamson
Thesis Title	Dealing with partially observed covariates in propensity score analysis of observational data

If the Research Paper has previously been published please complete Section B, if not please move to Section C

SECTION B – Paper already published

Where was the work published?			
When was the work published?			
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Choose an item.	Was the work subject to academic peer review?	Choose an item.

**If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.*

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	Statistics in Medicine
Please list the paper's authors in the intended authorship order:	Helen Blake, Clemence Leyrat, Kathryn Mansfield, Shaun Seaman, Laurie Tomlinson, James Carpenter, Elizabeth Williamson.
Stage of publication	**Undergoing revision**

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I undertook the novel research being presented in the manuscript, including developing the causal diagram theory and performing the simulations. I wrote the first draft of the manuscript and critically revised it following comments from my supervisors and other co-authors.

Student Signature: [redacted] Date: 24/9/19

Supervisor Signature: [redacted] Date: 24/9/19

5.1 Overview of the research paper pre-print: Propensity scores using missingness pattern information: a practical guide

In the following research paper pre-print, we explore assumptions under which the MPA can obtain valid inference. After introducing the MPA and the method's underlying assumptions, we discuss the plausibility of the CIT and CIO assumptions using two illustrative examples. The first example describes a clinical scenario where the CIT assumption seems plausible, whilst the second considers a scenario where the CIO assumption seems plausible. We describe how we can use causal diagrams to assess the plausibility of the MPA's assumptions and provide guidance for assessing these assumptions in a given clinical setting. We then illustrate this guidance in detail using our motivating study.

5.2 Abstract

Electronic health records are a valuable data source for investigating health-related questions, and propensity score analysis has become an increasingly popular approach to address confounding bias in such investigations. However, because electronic health records are typically routinely recorded as part of standard clinical care, there are often missing values, particularly for potential confounders. In our motivating study – using electronic health records to investigate the effect of renin-angiotensin system blockers on the risk of acute kidney injury – two key confounders, ethnicity and chronic kidney disease stage, have 59% and 53% missing data, respectively.

The missingness pattern approach (MPA), a variant of the missing indicator approach, has been proposed as a method for handling partially observed confounders in propensity score analysis. In the MPA, propensity scores are estimated separately for each missingness pattern present in the data. Although the assumptions underlying the validity of the MPA are stated in the literature, it can be difficult in

practice to assess their plausibility.

In this paper, we explore the MPA's underlying assumptions by using causal diagrams to assess their plausibility in a range of simple scenarios, drawing general conclusions about situations in which they are likely to be violated. We present a framework providing practical guidance for assessing whether the MPA's assumptions are plausible in a particular setting and thus deciding when the MPA is appropriate. We apply our framework to our motivating study, showing that the MPA's underlying assumptions appear reasonable, and we demonstrate the application of MPA to this study.

5.3 Introduction

Observational data are an important source of information for investigating the effect of treatments or interventions on health outcomes. In observational data, confounding is often an issue, as characteristics of treated patients can systematically differ from those of untreated patients. Propensity score methods aim to take account of confounding by achieving balance of patient characteristics across the treatment groups being compared. [33] However, observational studies may suffer from large amounts of missing data, which can lead to biased treatment effect estimates if the missing data are not handled appropriately. [50] We focus on scenarios where the outcome and treatment of interest are fully observed, but data are missing on potential confounders. This is a common occurrence, for example, in studies using electronic health record data and insurance claims data, where prescriptions and diagnoses tend to be well recorded but potential confounders, such as smoking status, may be less well recorded. [7]

The 'missingness pattern' approach (MPA) is a way of handling missing confounder data that has been proposed in propensity score analysis. [46,62] It accounts for missing data by incorporating information about which confounders are missing into the estimation of the propensity score itself. [46,62] Despite being easy to implement, the MPA has not been widely used in practice. This might be explained by the lack of guidance about its use in the literature. In particular, while the

assumptions required for the validity of the MPA have been described formally in terms of conditional independence, [46, 62, 63] how these mathematical statements relate to real clinical scenarios remains unclear. Our aim is therefore to investigate the assumptions underlying the MPA in order to provide practical guidance for researchers about how to identify whether these assumptions hold in a given clinical scenario.

We start by introducing our motivating example which investigates the association between renin-angiotensin system drugs and risk of acute kidney injury in Section 5.4. We review propensity score methods for complete data (Section 5.5) and approaches to handle missing confounder data in propensity score analysis, with a particular focus on the MPA and the related missing indicator approach (Section 5.6). We discuss the plausibility of the assumptions underlying the MPA in Section 5.7. We use causal diagrams to evaluate the assumptions in Section 5.8 and present a framework giving practical guidance for assessing these assumptions in Section 5.9. We illustrate our results on our motivating example (Section 5.10) and conclude with a discussion (Section 5.11).

5.4 Motivating Example

We consider data from a cohort study using electronic health records to investigate the association between use of angiotensin-converting enzyme inhibitors or angiotensin receptor blockers (ACEI/ARBs) and risk of acute kidney injury (AKI) in new users of antihypertensive drugs. [11]

Data were obtained from the UK Clinical Practice Research Datalink linked to the Hospital Episode Statistics database for adults who were new users of antihypertensive drugs between 1997-2014. Follow-up began at the first prescription of any of the antihypertensive drugs: ACEI/ARBs, beta blockers, calcium channel blockers or diuretics. Our treatment of interest is ACEI/ARB prescription at the start of follow-up, and the outcome is AKI within 5 years. Potential confounders considered are: gender; age; ethnicity; prescription of other antihypertensive drugs at start of follow-up; and status of chronic comorbidities at start of follow-up, including

Table 5.1: Patient characteristics by prescription of ACEI/ARBs

Baseline Characteristic		Prescribed ACEI/ARB			
		Yes (n (%)) (Total = 159,389)		No (n (%)) (Total = 411,197)	
Age (years)	18 to 42	16,616	(10.4%)	94,265	(22.9%)
	43 to 53	39,541	(24.8%)	77,224	(18.8%)
	54 to 62	36,325	(22.8%)	77,985	(19.0%)
	63 to 71	30,667	(19.2%)	75,141	(18.3%)
	$\geq$ 72	36,240	(22.7%)	86,582	(21.1%)
Sex	Female	62,652	(39.3%)	236,296	(57.5%)
Chronic Kidney Disease Stage	$\leq$ Stage 2	88,826	(55.7%)	146,825	(35.7%)
	Stage 3a	10,535	(6.6%)	15,489	(3.8%)
	Stage 3b	2,728	(1.7%)	3,127	(0.8%)
	Stage 4	457	(0.3%)	551	(0.1%)
	Missing	56,843	(35.7%)	245,205	(59.6%)
Ethnicity	White	63,791	(40.0%)	153,747	(37.4%)
	South Asian	3,072	(1.9%)	4,734	(1.2%)
	Black	1,065	(0.7%)	3,905	(0.9%)
	Mixed	237	(0.1%)	681	(0.2%)
	Other	814	(0.5%)	1,623	(0.4%)
	Missing	90,410	(56.7%)	246,507	(59.9%)
Comorbidities:					
Diabetes Mellitus	Yes	44,727	(28.1%)	38,714	(9.4%)
Ischaemic Heart Disease	Yes	42,214	(26.5%)	76,013	(18.5%)
Arrhythmia	Yes	17,494	(11.0%)	39,094	(9.5%)
Cardiac Failure	Yes	18,647	(11.7%)	13,074	(3.2%)
Hypertension	Yes	124,340	(78.0%)	240,135	(58.4%)
Other anti-hypertensives:					
Beta-blocker	Yes	14,666	(9.2%)	205,156	(49.9%)
Calcium Channel Blocker	Yes	3,501	(2.2%)	91,912	(22.4%)
Diuretic	Yes	21,950	(13.8%)	129,582	(31.5%)

Abbreviations:
ACEI/ARBs: Angiotensin-converting enzyme inhibitors or angiotensin receptor blockers
Diuretic: Thiazide diuretics, Loop diuretics or Potassium sparing diuretics

chronic kidney disease (CKD) stage. Table 5.1 shows the baseline characteristics of the cohort. Of the 570,586 patients included in the cohort, 159,389 (27.9%) were prescribed an ACEI/ARB. Many characteristics are not balanced across the treatment groups, indicating potential for confounding. Propensity score analysis is a popular method for taking account of confounding in analysis of electronic health records. However, two potential confounders have missing data: ethnicity (59.0% missing) and baseline CKD stage (52.9% missing). Only 121,527 (21%) of patients have complete data for both variables.

5.5 Propensity score methods for complete data

5.5.1 Notation and assumptions

Suppose we have a group of n patients, each with a row vector X_i of p confounders: $X_i = (X_{i1}, ..., X_{ip})^\top$, where X_{ij} is the value of confounder j for patient i ($i = 1, \ldots, n$ and $j = 1, \ldots, p$). Throughout the paper, we will assume that in the full data (i.e. with no missing confounder data) the X_i are sufficient to control for confounding. [39] In this paper, we restrict our attention to a binary treatment (or exposure) and a binary outcome. Patient i receives either treatment $Z_i = 1$ or control $Z_i = 0$. Each patient has two potential outcomes: $Y_i(1)$ denotes the outcome that would have been observed for patient i if they had received treatment, and $Y_i(0)$ denotes the outcome value that would have been observed if patient i had received control. [10] Y_i denotes the outcome value that was actually observed. Henceforth, we omit the i and j subscripts where unambiguous. Our estimand is the average treatment effect (ATE): $E[Y(1) - Y(0)]$. [25, 26] While the odds ratio suffers from non-collapsibility, the risk ratio does not and provides an alternative relative measure; results in this paper follow similarly for this estimand.

To estimate the treatment effect we make four standard assumptions: consistency, no interference, strongly ignorable treatment assignment (SITA), and positivity. Consistency states that, for a patient who receives a particular treatment level z, their observed outcome Y is the corresponding potential outcome $Y(z)$, irre-

spective of the way in which they were assigned to that treatment level. [31] Under the assumption of no interference, the treatment received by one patient does not affect the potential outcomes of another patient. [33–35] SITA implies that there are no unmeasured confounders and can be expressed as [23, 33]:

$$\text{SITA}: \quad Z \perp \big(Y(1), Y(0)\big)|X. \tag{5.1}$$

where $\perp$ denotes independence. Finally, positivity states that, given their individual characteristics, all patients have a non-zero probability of receiving either treatment or control. [23, 37] Throughout this paper, we assume these four assumptions hold for the complete data.

5.5.2 Propensity scores

The propensity score $e(x)$ is the probability of receiving treatment, conditional on observed confounders X [33]:

$$e_i(x_i) = P(Z_i = 1|X_i = x_i),$$

for patient i $(i = 1, \ldots, n)$ with a vector of confounder values $X_i = x_i$. Under the four assumptions described above, Rosenbaum and Rubin showed that at each value of the propensity score the confounders X are balanced across treatment groups. [33]

Typically, propensity scores are unknown and must be estimated from the data, often as the predictions, $\hat{e}_i$, from a logistic regression of treatment allocation on potential confounders. [28] We use inverse probability of treatment weighting (IPTW), which creates weights from the estimated propensity scores to construct 'pseudo-populations' [36] in which the distribution of observed confounders are balanced across treatment groups, resulting in the following estimator [28]:

$$\widehat{\text{ATE}} = \left(\frac{\sum_{i=1}^{n} \frac{Y_i Z_i}{\hat{e}_i}}{\sum_{i=1}^{n} \frac{Z_i}{\hat{e}_i}} \right) - \left(\frac{\sum_{i=1}^{n} \frac{Y_i (1-Z_i)}{(1-\hat{e}_i)}}{\sum_{i=1}^{n} \frac{(1-Z_i)}{(1-\hat{e}_i)}} \right). \tag{5.2}$$

5.6 Propensity score methods with missing confounder data

In practice, observational studies can suffer from large amounts of missing data, potentially leading to both loss of efficiency and biased estimates. [44] The magnitude of bias will depend on the extent to which the probability of missing confounder data is related to outcome and exposure. [44] The most common classification of missingness mechanisms is Rubin's taxonomy, in which data are missing completely at random (MCAR), missing at random (MAR) or missing not at random (MNAR). [44,48] Under a MCAR mechanism, the probability of being missing does not depend on the observed or missing data. Missing data are MAR when the probability of being missing depends on observed data values but, given these, does not depend on missing values. If the probability of being missing depends on the unobserved values of data then data are MNAR.

The simplest way of handling missing confounder data in propensity score analysis is a complete records (or complete case) analysis, which restricts the analyses to patients with full data on all variables. [50] This approach provides unbiased estimates of the conditional average treatment effect as long as missingness does not depend on both the treatment and the outcome. [50]

The missing indicator approach is another simple method. For partially observed categorical confounders, a 'missing' category is added before including the confounder in the propensity score model. For continuous confounders, missing values are set to a particular value, say 0, and both the confounder and a missing indicator (a variable indicating whether that variable is observed or not) are included in the propensity score model. Applied to standard outcome regression models, this approach induces bias in a number of scenarios [58, 59]; whether this is the case in the propensity score context has been questioned. [64]

Multiple imputation is a popular alternative, involving imputing (i.e. filling in) missing covariates with plausible values several times, by drawing from the predictive distribution of the missing covariates given observed data, to create a number of

'complete' imputed datasets. The full analysis (estimation of the propensity score then estimation of the treatment effect) is performed separately in each imputed dataset. The results are then combined using Rubin's rules to obtain an overall estimate of the treatment effect and standard error. [20, 44] Guidelines regarding optimal use of multiple imputation in conjunction with propensity score analysis have been proposed. [60] Standard implementations of multiple imputation require data to be missing at random. [44, 49]

5.6.1 The Missingness Pattern Approach (MPA)

The Missingness Pattern Approach (MPA) [46, 62] accounts for missing confounder data by separating patients into subgroups according to the possible combinations of confounders being observed or missing, i.e. the missingness patterns, and fitting a different propensity score model to each pattern.

Let R_{ij} be a missing indicator indicating whether the confounder j ($j = 1, \ldots, p$) for patient i ($i = 1, \ldots, n$) is observed ($R_{ij} = 1$) or not ($R_{ij} = 0$). This allows us to partition the values X_{ij} ($i = 1, \ldots, n; j = 1, \ldots, p$) into two sets: the set of values that are observed, X_{obs}, and the set of values that are missing, X_{mis}:

$$X = \{X_{obs}, X_{mis}\} \text{ where } X_{obs} = \{X_{ij}|R_{ij} = 1\} \text{ and } X_{mis} = \{X_{ij}|R_{ij} = 0\}. \quad (5.3)$$

We will use $R_i = (R_{i1}, ..., R_{ip})$ to refer to the vector of missing indicators for patient i, omitting the subscript i where unambiguous.

The generalized propensity score, $e^*(x)$, is defined as the probability of receiving treatment, conditional on both the observed confounder information and the missingness pattern: $e^*(x) = P(Z = 1|X_{obs}, R)$. This can be estimated by using a different propensity score model for each missingness pattern, including only the confounders observed in that pattern. For example, in a study with treatment and outcome both fully observed and a single partially observed confounder X, there are two missingness patterns: X is either observed or missing. For patients with X observed, the propensity score model would include X, whilst the propensity score model for patients with X missing would include only a constant term. The gener-

alised propensity score can then be used in the same way as the standard propensity score, [62] eg. by substituting in equation (5.2) to estimate the ATE.

5.6.1.1 Assumptions of the Missingness Pattern Approach

Three assumptions under which the MPA leads to valid inference are given by Mattei. [63] We present slightly weaker versions of these assumptions, under which the MPA still gives a consistent estimator of the ATE (proof in Supplementary Material: Section A). The first assumption is an extension of the SITA assumption (equation (5.1)), which we call the Missingness Strongly Ignorable Treatment Assignment (mSITA) assumption due to its similarities with the SITA assumption (equation (5.1)):

$$\text{mSITA:} \quad Z \perp Y(z)|X, R \qquad \text{for} \qquad z = 0, 1. \tag{5.4}$$

A key difference with equation (5.1) is the inclusion of information about the missingness pattern, represented by R, in the conditioning set. We assume that SITA holds in the full data, thus this assumption states that additionally conditioning on R does not introduce bias.

We call the two further assumptions [63]: the conditionally independent treatment (CIT) assumption and the conditionally independent outcomes (CIO) assumption.

$$\text{CIT:} \quad Z \perp X_{mis}|X_{obs}, R \tag{5.5a}$$

$$\text{CIO:} \quad Y(z) \perp X_{mis}|X_{obs}, R \qquad \text{for} \qquad z = 0, 1. \tag{5.5b}$$

If mSITA holds, and either CIT or CIO holds, then the MPA provides a consistent estimate of the treatment effect. We loosely term these the 'MPA's assumptions'.

We note that the assumptions underlying the MPA are different to Rubin's taxonomy of missing data [20,45] in the sense that classifying data according to Rubin's taxonomy does not provide any information as to whether the MPA's assumptions would hold. Rather, the MPA's assumptions require the associations between vari-

ables to differ across missingness patterns. It is possible for the MPA's assumptions to hold when data are missing not at random; conversely data being missing completely at random does not guarantee the MPA's assumptions will hold.

How to assess the plausibility of the MPA's assumptions — and thus the validity of the MPA itself — in a particular setting remains unclear.

5.6.1.2 Connections with the missing indicator approach

With a single partially observed confounder, the missing indicator approach can be shown to be equivalent to the MPA (Supplementary Material: Section B). Thus the missing indicator approach will provide a consistent estimator of the ATE if mSITA and either CIT or CIO holds.

In a more complex scenario with one partially observed and one fully observed confounder, the missing indicator approach is a simplified version of the MPA, additionally imposing the assumption that the association between the fully observed confounder and treatment is the same whether or not the other confounder is observed (Supplementary Material: Section B). Therefore, the missing indicator approach relies on the same assumptions as the MPA, and additionally requires no effect modification of the fully observed confounder(s) by the missingness patterns.

5.7 Plausibility of the CIT and CIO assumptions

The MPA provides valid inference if either CIT or CIO holds (equation (5.5)), in addition to the mSITA assumption. The plausibility of these assumptions in real-life settings will therefore determine how useful the MPA is as a missing data approach.

We have assumed that in the full data X is a confounder, and so is associated with both treatment and outcome. The CIT assumption requires that the confounder-treatment relationship is absent in the subset of patients with X unmeasured, whilst the CIO assumption requires that the confounder-outcome relationship is absent in patients with X unmeasured. Thus, if either the CIT or CIO assumption holds, X does not confound the relationship between treatment and outcome when it is missing (i.e. X is not associated with both treatment and outcome in the subset of

patients missing X). Informally, we refer to this property as X being a confounder only when it is observed.

The key point to consider is that the CIT and CIO assumptions are not about the missingness mechanisms that drive the missing data, as much as which relationships between variables exist in the subgroup of patients with missing confounder values.

5.7.1 The CIT assumption: an illustrative example

Consider a simplified version of our motivating example, investigating the effect of prescribing ACEI/ARBs on the risk of AKI.

Underlying kidney function prior to ACEI/ARB prescription is a likely confounder: kidney function is a known risk factor for AKI and is likely to influence whether ACEI/ARBs are prescribed. Kidney function is classified into the stage of chronic kidney disease (CKD), via a serum creatinine blood test. Where a clinician ordered a kidney function test prior to the prescribing decision, it is reasonable to assume that the information regarding CKD stage contributed to that decision. Where CKD stage was unavailable to the clinician, arguably it is unlikely to have influenced the prescribing decision.

In this simplified example, underlying CKD stage is always a risk factor for the outcome but is plausibly only associated with treatment allocation when it is measured. Thus, CIT holds; baseline CKD stage is only a confounder when it is observed.

5.7.2 The CIO assumption: an illustrative example

Suppose we were interested in estimating the effect of exposure to farming in early life on subsequent development of asthma. Childhood exposures to domestic allergens, e.g. dust mites, are potential confounders. Such domestic allergens may be measured by health visitors. Suppose that the relationship between dust mites and asthma has a threshold effect, i.e. an association is seen only once a certain concentration of dust mites is present.

Since health visitors do not collect information for the purposes of research, they

might plausibly record information more thoroughly for households where there were concerns about the child's environment. Missing data for dust mites would therefore be more likely to occur in households with little evidence of dust mites, and less likely in households with a high concentration.

In this example, concentration of dust mites may be associated with subsequent asthma only in households where dust mite concentration was recorded. In this case, CIO holds; dust mite concentration is a confounder only when measured.

5.8 Detecting and dealing with violations of the MPA's assumptions

The mSITA, CIT and CIO assumptions are statements of conditional independence. In this section, we describe how causal diagrams can be used to assess conditional independence statements. We demonstrate the use of causal diagrams in a simple scenario in order to draw some general conclusions about situations in which the MPA's assumptions are likely to be violated.

5.8.1 Causal diagrams

Causal diagrams, or directed acyclic graphs, are a useful tool for assessing conditional independencies under an assumed causal structure. Because the assumptions of the MPA involve the potential, rather than observed, outcomes we turn to a specific type of causal diagram: Single World Intervention Templates (SWITs). [67]

SWITs are standard directed acyclic graphs which have been adapted to show potential, instead of observed, outcomes. This involves 'splitting' the treatment node into two halves; the first represents the observed treatment Z, and the second represents an 'intervened-on value', z. Determinants of observed treatment affect the first half (i.e. incoming arrows go into the Z half), and effects of treatment are determined by the second (i.e. outcoming arrows leave from the z half). A consequence of this splitting is that variables affected by treatment now become potential rather than observed variables.

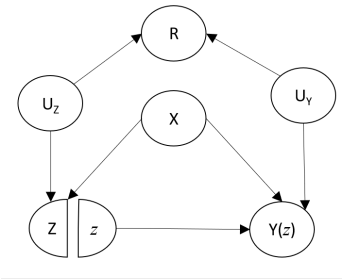


Figure 5.1: A single world intervention template showing a scenario in which the mSITA assumption is violated.
X: confounder. Z: treatment. $Y(z)$: potential outcome resulting from intervening to set Z equal to a particular value z. R: missing indicator (=1 if X observed, =0 if X is missing). U_Z: unobserved common cause between R and Z. U_Y: unobserved common cause between R and $Y(z)$.

Figure 5.1 shows a simple SWIT representing a typical confounding scenario where the confounder X has a causal effect on the treatment and the outcome. Additionally, this graph encodes the assumption that the missing indicator R (i.e. whether or not the confounder is missing) is associated with the treatment and the outcome, via shared common causes in both cases (denoted U_Z and U_Y respectively). In Figure 5.1, the outcome is affected by treatment so this SWIT includes the potential outcome $Y(z)$ rather than the observed outcome Y.

5.8.2 Assessing the MPA's assumptions using causal diagrams

5.8.2.1 Assessing the mSITA assumption

Suppose Figure 5.1 depicts the true underlying causal structure which gave rise to our study data. With a single partially observed confounder, the mSITA assumption states that $Z \perp Y(z)|X, R$. By applying d-separation to Figure 5.1 (as described in Supplementary Material: Section C), we find that the path from Z to $Y(z)$ through R is open after conditioning on X and R, thus Z is not conditionally independent of $Y(z)$ given X and R; mSITA is violated in this scenario.

For more complex causal diagrams, it may help to use software such as Dagitty to assess which conditional independencies hold. [77] R code which uses Dagitty to check the MPA's assumptions for the scenario shown in Figure 5.1 can be found in

Supplementary Material: Section F.

5.8.2.2 Assessing the CIT/CIO assumptions

The CIT and CIO assumptions state that $Z \perp X_{mis}|X_{obs}, R$, and $Y(z) \perp X_{mis}|X_{obs}, R$, respectively. With a single confounder X, these assumptions are trivially true in the subgroup of patients with X observed (because X_{mis} is empty given $R = 1$). In the subgroup of patients with X missing, the assumptions become: $Z \perp X|R = 0$, and $Y(z) \perp X|R = 0$, respectively.

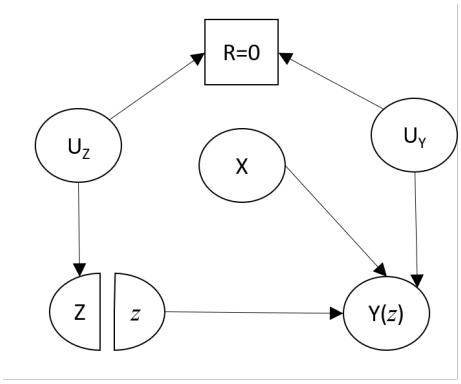


Figure 5.2: A single world intervention template modified (from Figure 5.1) to assess the CIT and CIO assumptions. The square box around R denotes the restriction of our attention to the subgroup $R = 0$.
X: confounder. Z: treatment. $Y(z)$: potential outcome resulting from intervening to set Z equal to a particular value z. R: missing indicator (=1 if X observed, =0 if X is missing). U_Z: unobserved common cause between R and Z. U_Y: unobserved common cause between R and $Y(z)$.

A minimum condition for CIT or CIO to be satisfied is that X cannot be a confounder when it is missing. Thus, for either of these assumptions to hold, there must be grounds for believing that the causal relationships that generate confounding bias in the full data are different in the subgroup with missing confounder values (compared to the subgroup with observed confounder values). For example, in Figure 5.1, if we believe that all the arrows shown exist in the subgroup with missing confounder values, then both CIT and CIO would be violated. In contrast, suppose we believe that this diagram depicted the correct situation with full data, but we believe that the arrow from the confounder to treatment did not exist when X was missing. In this case, Figure 5.2 would depict the underlying causal structure for the subgroup with X unmeasured.

In Figure 5.2, the only path connecting Z and X passes through $Y(z)$, a collider on the path; thus applying the d-separation rule shows that Z and X are conditionally independent (in the subgroup with $R = 0$). Here, CIT holds. Because there is a direct arrow from X to $Y(z)$, however, CIO does not hold.

5.8.3 Key violations of the MPA's assumptions

In this section, we use causal diagrams to explore when the MPA's assumptions are violated in a range of simple settings.

Scenarios considered: We consider scenarios where the outcome Y and treatment Z are fully observed. Initially, we focus on simple scenarios with a single partially observed confounder, X. Subsequently we extend this to consider scenarios with an additional, fully observed confounder, C. We consider all combinations of the scenarios discussed below, omitting those which give rise to cycles (i.e. we do not allow scenarios where a variable has a causal effect on itself).

Relationships between the confounder, treatment, and outcome: We consider causal structures where the relationships between the confounder X and the treatment and outcome are either a direct causal relationship (e.g. X causes treatment), or via shared unmeasured common causes (e.g. a third factor causes both X and treatment). The relationship between the confounder and the treatment is allowed to differ depending on whether the confounder is observed or missing; specifically, this relationship is allowed to be absent when $R = 0$. Similarly, the presence or absence of the relationship between the confounder and outcome is allowed to depend on R. This allows for X to be a confounder only when observed, as discussed in the previous section.

Missingness mechanisms: For each of the confounder, treatment and outcome, we considered: no relationship with the missing indicator, a causal effect on the missing indicator, the missing indicator has a causal effect on the variable, or an unobserved common cause with the missing indicator (allowing scenarios where one or more variables have both a direct causal relationship and a common cause with the missing indicator).

When a variable has a causal effect on the outcome, we assume that this effect operates on the potential outcome rather than the observed (e.g. X causes $Y(z)$ rather than X causing Y). Conversely, in the case where outcome is a cause of missingness, we have chosen to allow the observed outcome to cause missingness (Y causes R) rather than the potential outcome, since this is arguably more plausible in real data.

Assessment of assumptions: In each setting, we draw the appropriate causal diagram and assess the assumptions by applying d-separation to the causal diagram overall, and to the modified causal diagram restricted to the subgroup with X missing.

In some scenarios, a slightly more complex route must be taken to assess the conditional independencies involved in the MPA's assumptions. If the treatment or outcome is a cause of missingness then the relevant SWIT contains $R(z)$, the 'potential' missingness after intervening on treatment, rather than the observed pattern of missingness. Thus we can no longer use this graph to assess the relevant assumptions. In these cases we turn to twin networks [68, 69] (Supplementary Material: Section D).

5.8.3.1 Key violations of the mSITA assumption

In the scenarios we considered, most violations of mSITA occurred via collider bias on R. In order for this type of violation to occur, there needs to be a path from Z to R *and* a path from $Y(z)$ to R, each ending with arrows pointing towards R. These violations operate via a cause of R. We let U_X represent common causes of missingness and the confounder, U_Z represent common causes of missingness and the treatment, and U_Y represent common causes of missingness and the outcome.

The different 'Z-to-R' and 'R-to-Y' patterns that could occur are summarised in Figure 5.3. If one (or more) of each of these two patterns occurs then mSITA will be violated. For example, Figure 5.1 shows the violation which arises when both the indirect 'Z-to-R' pattern and the indirect 'R-to-Y' pattern occur (both patterns in the second row of Figure 5.3).

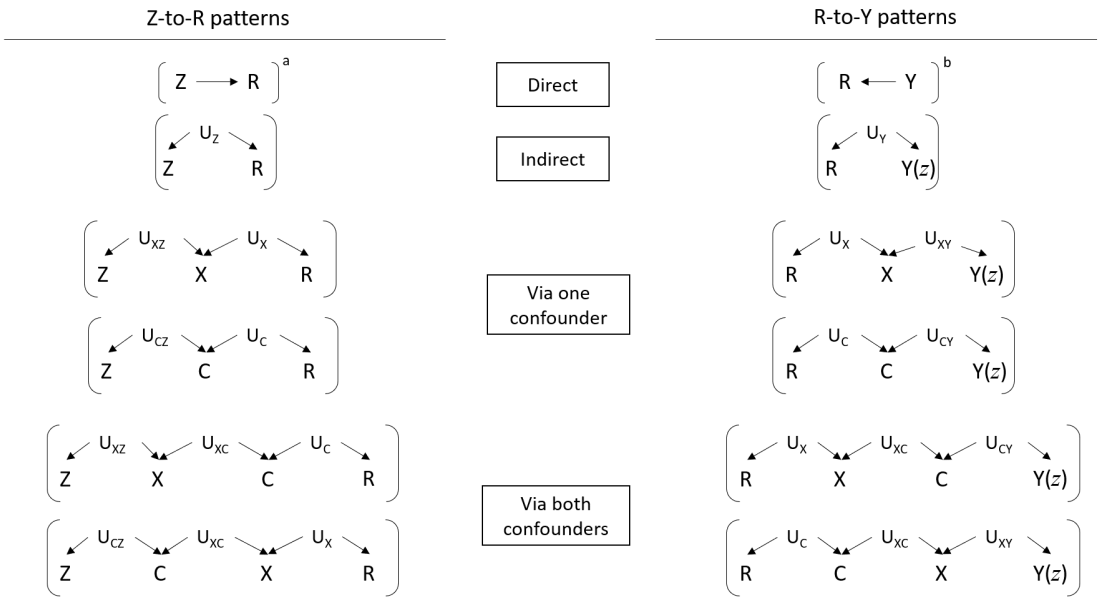


Figure 5.3: Summary of violations of the mSITA assumption. If one of the 'Z-to-R' patterns and one of the 'R-to-Y' patterns occurs in the causal diagram representing the study in question then the mSITA assumption will be violated.
[a] Also a violation if this occurs with additional 'R-to-Y patterns' shown in Supplementary Material: Section E; [b] Sufficient condition on its own, without a 'Z-to-R pattern'.
X: partially observed confounder. C: fully observed confounder. Z: treatment. $Y(z)$: potential outcome resulting from intervening to set Z equal to a particular value z. Y: observed outcome. R: missing indicator (=1 if X observed, =0 if X is missing). U_{st}: unobserved common cause between two variables s and t. U_s: unobserved common cause between R and another variable s.

A key result in Figure 5.3 is that when treatment and missingness are associated via shared common causes, and outcome and missingness are associated via (different) shared common causes, then mSITA is violated (as shown in Figure 5.1). So the MPA cannot be used in scenarios where there are unmeasured determinants of confounder missingness which are also associated with the treatment and the potential outcomes.

Another important result in Figure 5.3 is that if the outcome has a causal effect on confounder missingness, i.e. if $Y \rightarrow R$, then mSITA is violated without the need for any 'Z-to-R' patterns. So the MPA cannot be used in scenarios where outcome affects whether or not confounder values are missing. For instance, in our AKI example, suppose that more efforts were made to track down historical laboratory measures of eGFR for patients who were diagnosed with AKI, then this would immediately violate mSITA.

A third important result is that when treatment causes missingness, and miss-

ingness in turn has a causal effect on the potential outcomes, mSITA is violated (see footnote a in Figure 5.3), although whether this is likely to occur in practice is unclear.

5.8.3.2 Handling violations of the mSITA assumption

All violations of mSITA, other than those involving the treatment or the outcome causing missingness of the confounder, operate via a cause of R. Suppose it were possible to measure all such factors which determine whether or not the confounder is measured (although this may be difficult in practice). We could define a new set of confounders $\tilde{X} = \{X, U_X, U_Z, U_Y\}$ (or, where there is an additional fully observed confounder C, $\tilde{X} = \{X, C, U_X, U_C, U_Z, U_Y\}$). Including this new set of confounders in the propensity score model, and thus the conditioning set for mSITA, removes the violation of this assumption. In most cases, measuring a subset of these variables will suffice. For example, in Figure 5.1, if U_Z could be measured and included in the propensity score model, the mSITA assumption would become: $Z \perp Y(z)|X, R, U_Z$, which is satisfied in Figure 5.1.

5.8.3.3 Key violations of the CIT and CIO assumptions

Figure 5.4 summarises the possible violations of CIT and CIO, which fall into two broad groups: (A) violations related to X being a confounder when it is missing, and (B) violations due to collider bias via R.

Since mSITA is always violated if outcome causes missingness, some CIT/CIO violations involving $Y \rightarrow R$ are shown only in Supplementary Material: Section E, along with a few additional violations involving $Z \rightarrow R$.

Group (A) violations in Figure 5.4 relate to X being a confounder only when observed, in the sense that if one of the CIT group (A) violations or one of the CIO group (A) violations held, X would be a confounder when missing. For these violations, X has been replaced by X_{mis} to emphasise the fact that we need to focus on relationships that exist in the subgroup of patients with a missing confounder value when assessing this assumption.

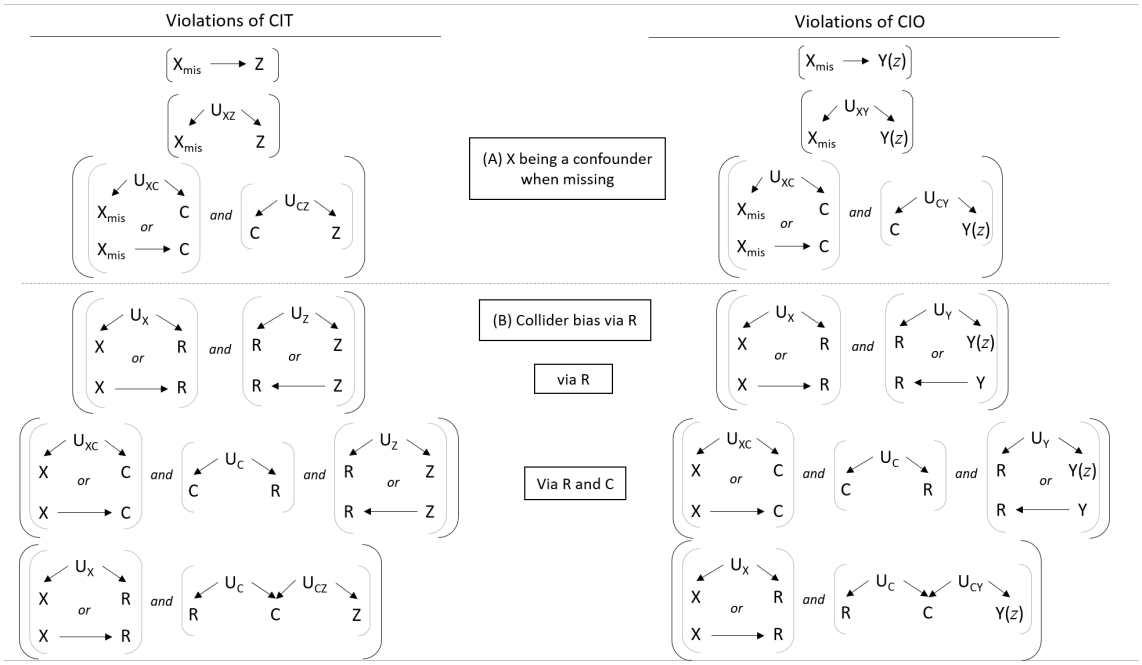


Figure 5.4: Summary of violations of the CIT and CIO assumptions. If one or more of the six sets of conditions on the left hand side appear in the relevant causal diagram (modified to reflect relationships in the subgroup with X unobserved i.e. restricted to $R = 0$), the CIT is violated. Similarly, if any of the six sets of conditions on the right hand side occur then CIO is violated. Additional violations involving $Y \to R$ and $Z \to R$ can be found in Supplementary Material: Section E.
X: partially observed confounder. C: fully observed confounder. Z: treatment. $Y(z)$: potential outcome resulting from intervening to set Z equal to a particular value z. Y: observed outcome. R: missing indicator (=1 if X observed, =0 if X is missing). U_{st}: unobserved common cause between two variables s and t. U_s: unobserved common cause between R and another variable s.

In contrast, Group (B) violations relate to collider bias induced by conditioning on R.

5.8.3.4 Handling violations of the CIT and CIO assumptions

As with violations of the mSITA assumption, many of the violations of the CIT and CIO assumptions — specifically those belonging to Group (B) — can be removed by measuring and conditioning on causes of R. However, if either (i) both the confounder and the treatment cause the missingness, or (ii) both the confounder and outcome cause the missingness, then CIT or CIO are violated, respectively; no conditioning can remove these violations.

5.9 Practical guide to assessing the mSITA, CIT and CIO assumptions

In order to decide if the MPA's assumptions hold in a particular clinical setting, the first, most important step, is to assess whether it is plausible for the partially observed confounder to be a confounder only when observed.

Second, key scenarios in which the MPA's assumptions do not hold, as identified in the previous section, should be carefully considered using substantive knowledge to ensure these do not apply in the setting at hand. These are: (I) outcome affects missingness of the confounder; (II) outcome and missingness have shared unmeasured common causes, and treatment and missingness have shared common causes; or (III) the confounder and treatment both affect missingness of the confounder and the confounder is associated with outcome in the subgroup with X missing.

Third, a causal diagram should be constructed, reflecting what is believed to be the underlying clinical structure. As with any causal diagram, any variable — measured or unmeasured — which may have a causal effect on two or more variables in the causal diagram must also be included. Missing indicators for the partially observed confounders should be included in the causal diagram at this stage. When there are multiple partially observed confounders, the causal diagram will include one missing indicator per partially observed confounder.

Fourth, the causal diagram should be converted into a SWIT or a twin network, as appropriate. Once the SWIT or twin network has been created, d-separation can be applied to determine whether mSITA holds.

To assess CIT and CIO, the SWIT or twin network should be modified to reflect the relationship thought to be absent in the subgroup of patients with missing confounder values (i.e. remove the arrows which reflect the assumption that the confounder is only a confounder when observed). In this modified diagram, the d-separation rule can be again applied to assess CIT and CIO.

Supplementary Material: Section F provides R code to assess the mSITA, CIT and CIO assumptions for Figure 5.1, and for the causal diagram associated with our

more complex motivating example.

When there are multiple partially observed confounders, we advise constructing modified diagrams for each missingness pattern with missing values and then applying the d-separation rule to each diagram to assess the CIT and CIO assumptions for that particular missingness pattern. An assumption holds only if it holds for each missingness pattern.

5.9.1 Assessing the validity of the assumptions in the motivating example

5.9.1.1 Confounders only when observed

For the MPA's assumptions to hold in the motivating example, we have to believe that the two partially-missing confounders — ethnicity and baseline chronic kidney disease (CKD) stage — act as confounders only when observed. If baseline CKD stage is not available, this unobserved information cannot be used to determine the General Practitioner's treatment decision whether or not to prescribe ACEI/ARBs. In practice, CKD stage may be recorded in a part of the patient record that the General Practitioner is aware of but researchers using CPRD data cannot access (e.g. letters from secondary care). However, this is likely to reflect advanced CKD for only a small proportion of the whole study population. So in general, it seems plausible that baseline CKD stage affects the clinician's prescribing decision only when recorded.

The National Institute for Health and Care Excellence antihypertensive prescribing guidelines (which include ACEI/ARBs) offer different recommendations depending on ethnicity. [78] So, it is plausible that, if a clinician chooses to prescribe or not prescribe an ACEI/ARB based on an individual's ethnicity, they would ensure that the individual's ethnicity was recorded.

Therefore, the CIT assumption may be reasonable for this scenario. Conversely, both baseline CKD stage and ethnicity are risk factors for AKI, whether measured or not. Thus the CIO assumption is not plausible here.

5.9.1.2 Checking plausibility of key violations

We also need grounds to believe that the three key scenarios listed above do not apply in this setting. Scenarios (I) and (III) rely on either outcome or treatment affecting missingness of the confounders. As CKD stage was defined at baseline, missingness of baseline CKD stage precedes treatment and, as a result, outcome. It also seems plausible that missingness of ethnicity occurs prior to treatment and outcome. Hence we believe that these scenarios do not apply here.

Scenario (II) is when outcome and missingness have shared unmeasured common causes, and treatment and missingness have shared common causes. Baseline CKD stage is more likely to be recorded for patients expected to have a higher risk of kidney disease due to age or chronic comorbidities (eg. hypertension, diabetes) or due to other signs that the patient has poor kidney function (i.e. CKD itself may affect the chance of the clinician measuring CKD stage). Whilst these risk factors are associated with missingness and treatment or outcome, they are already captured in the electronic health records.

With respect to ethnicity, patients who are hospitalised are more likely to have ethnicity recorded (due to linkage of primary and secondary care data). Missingness of ethnicity may be caused by service-level factors such as level of administrative support at the time patients are admitted to hospital. It seems unlikely that these factors are also determinants of treatments previously prescribed in primary care or whether patients develop acute illnesses that require admission to hospital. Since we believe that any relevant common causes are measured, scenario (II) does not apply in our setting.

After considering the three key scenarios mentioned above, in which the MPA's assumptions do not hold, we have found that these do not seem plausible in our motivating example. Having ruled out these violations, we proceed to the next step of our framework: to develop a causal diagram.

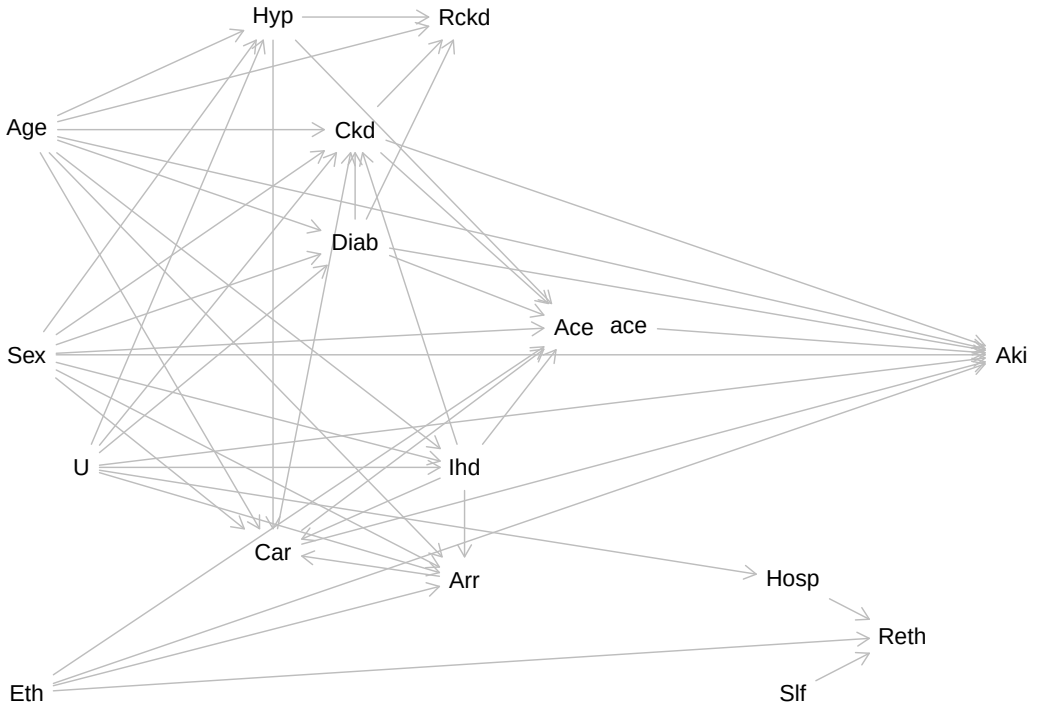


Figure 5.5: A single world intervention template for the motivating example.
Eth: Ethnicity. Ckd: Baseline chronic kidney disease. Hyp: Hypertension. Diab: Diabetes mellitus. Arr: Arrhythmia. Car: Cardiac failure. Ihd: Ischaemic heart disease. Ace: Prescription of ACE/ARBs (treatment). ace: intervened-on version of exposure. Aki: Acute kidney injury (outcome). Rckd: Missingness of Ckd. Reth: Missingness of Eth. Hosp: Hospitalisation. Slf: Service-level factors. U: unmeasured factor.

5.9.1.3 Developing a causal diagram

Figure 5.5 shows the single world intervention template (SWIT) developed for this example. This causal diagram encodes the investigators' assumptions that age, sex and ethnicity each affect both treatment and outcome. Age and sex affect the risk of developing diabetes, CKD, ischaemic heart disease, cardiac failure, arrhythmia and hypertension. Note that the treatment node, representing prescription of ACEI/ARBs, has been split into two: 'Ace' and 'ace', with the former representing the observed treatment and the latter representing the intervened-on treatment. Thus patient factors affect 'Ace' but not 'ace', and only 'ace' affects subsequent AKI.

5.9.1.4 Assessing the mSITA assumption

The mSITA assumption, for the motivating example, says that: $Z \perp Y(z)|R_{ckd}$, R_{eth}, Ckd, Eth, V, where Z represents ACEI/ARB prescription; $Y(z)$ the potential

outcome (AKI status that would be observed if the patient were prescribed level z of ACEI/ARB); V represents the confounders age, sex, diabetes, ischaemic heart disease, cardiac failure, arrhythmia and hypertension; R_{eth} and R_{ckd} are missing indicators for ethnicity and baseline CKD stage; and Ckd and Eth are the confounders CKD stage and ethnicity, respectively.

The d-separation rule can be applied to the SWIT in Figure 5.5, to assess whether this conditional independence holds under the causal assumptions encoded in the diagram (example code in Supplementary Material: Section F.2). In this case, the conditional independence statement is true; mSITA holds under the assumed causal diagram.

5.9.1.5 Assessing the CIT and CIO assumptions

We have already established that the CIO assumption does not hold in our motivating example. The CIT assumption states that:

$$\begin{aligned} Z &\perp (Ckd, Eth) \mid R_{ckd} = 0, R_{eth} = 0, \quad V, \\ Z &\perp Ckd \mid R_{ckd} = 0, R_{eth} = 1, Eth, V, \\ Z &\perp Eth \mid R_{ckd} = 1, R_{eth} = 0, Ckd, V. \end{aligned}$$

To assess the first of these, we create a modified version of Figure 5.5 which omits the arrows that we do not think exist when both ethnicity and baseline CKD stage are missing. So we remove the arrow from baseline CKD stage to ACEI/ARB prescription, and we remove the arrow from ethnicity to ACEI/ARB prescription.

We then assess whether, after conditioning on the two missing indicators and the fully observed confounders, the treatment is independent of both ethnicity and baseline CKD stage, by applying the d-separation rule for each partially observed confounder. In this case, the conditional independence holds (example code in Supplementary Material: Section F.2).

This process is repeated in the two subgroups where only one of ethnicity and baseline CKD stage is recorded, assessing the second and third independence statements above in the appropriately modified causal diagrams. In each case, the rel-

evant conditional independence holds. Thus, under the assumed causal diagram, CIT holds.

If our causal diagram correctly represents the causal structure giving rise to our study data, both mSITA and CIT hold. Under these two assumptions the MPA will provide consistent estimates of the ATE.

5.10 Motivating example: applying the MPA

5.10.1 Methods: ACEI/ARBs and AKI

We estimated the effect of prescription of ACEI/ARBs on the incidence of AKI within 5 years of follow-up as a risk difference, first with no adjustment for confounding, and then by using inverse probability of treatment weighting (IPTW). For IPTW, we estimated propensity scores using logistic regression to model ACEI/ARB prescription as a function of the covariates: age, sex, baseline CKD stage, ethnicity, diabetes mellitus, ischaemic heart disease, arrhythmia, cardiac failure and hypertension (including an interaction between age and ischaemic heart disease, and an interaction between age and hypertension). We applied non-parametric bootstrapping (500 replications of the combined process of propensity score estimation and treatment effect estimation) to obtain Normal approximation 95% confidence intervals.

To deal with missing data in baseline CKD stage and ethnicity, we applied complete records analysis, the MPA, the missing indicator approach and multiple imputation. For the MPA, the propensity scores were estimated separately in the four subgroups corresponding to whether or not baseline CKD stage and ethnicity were measured. For the missing indicator approach, we added 'missing' categories to each of baseline CKD stage and ethnicity. For multiple imputation, 10 imputed datasets were created using chained equations. The imputation model included AKI incidence within 5 years, ACEI/ARB prescription and all covariates and interactions included in the propensity score model. In each imputed dataset, propensity scores were estimated and IPTW was used to obtain treatment effect estimates, which

were then pooled using Rubin's rules. [60] To assess covariate balance, standardized differences [25] were calculated in the original sample and after IPTW with each analysis method used.

5.10.2 Results and discussion: ACEI/ARBs and AKI

The complete records analysis included 121,527 patients with full data. All other missing data methods included all 570,586 patients. Using any of the analysis methods with IPTW removes most of the imbalance present in the original dataset (Table S1 in Supplementary Material: Section G).

Table 5.2: Estimated effects of ACEI/ARBs on AKI using inverse-probability of treatment weighting (IPTW) to account for confounding.

Confounder adjustment	Missing data method	Risk difference (per 1000 people)	Normal-based bootstrap 95% CI
Crude	None	13.30	(12.52, 14.08)
IPTW	Complete Case Analysis	4.60	(2.76, 6.45)
IPTW	Missingness Pattern Approach	5.96	(5.10, 6.82)
IPTW	Missing Indicator Approach	5.93	(5.01, 6.85)
IPTW	Multiple Imputation	6.17	(5.27, 7.07)*

* Not bootstrapped; obtained by using Rubin's rules across 10 imputed datasets.

Estimates of the effect of ACEI/ARBs on AKI are shown in Table 5.2. All missing data methods greatly reduce the crude estimate of effect, with complete records analysis providing the smallest estimate and multiple imputation providing the estimate closest to the crude analysis. The MPA and missing indicator approach produce almost identical results, estimating that patients prescribed ACEI/ARBs had 6 additional cases of AKI within 5 years, per 1000 people, with a 95% confidence interval of (5,7), compared to patients who were not prescribed ACEI/ARBs.

We expect the MPA estimate to be consistent since — as discussed — the mSITA and CIT assumptions appear plausible here. Conversely, the missing at random assumption underlying our application of multiple imputation is questionable. Baseline CKD stage is more likely to be recorded for patients with a lower level of kidney function (e.g. if they are ill or have more risk factors for kidney disease that have led to testing) [79] and therefore baseline CKD stage may be MNAR. However, since factors related to a lower level of kidney function are likely already captured in the

observed data, the departure from the MAR assumption may be small. This may explain why multiple imputation and the MPA provide fairly similar estimates in this example, with multiple imputation giving an estimate closer to the crude estimate. Alternatively, having similar results may be due to misspecification of the parametric models or because ethnicity and baseline CKD stage may not be strong confounders.

In terms of precision, the complete records analysis has a very wide confidence interval, in contrast to the other missing data methods which all produce much narrower confidence intervals. This loss in precision, due to the exclusion of a large portion of the data, is recovered by the MPA, the missing indicator approach and multiple imputation.

5.11 Discussion

We have explored the three assumptions under which the missingness pattern approach to dealing with missing counfounders in propensity score analysis provides valid inference. We have described how d-separation can be applied to a causal diagram to assess the MPA's assumptions in a given setting and provided a framework and detailed example to allow researchers to ensure the appropriateness of this method in practice.

The key assumption required by the MPA is that the confounder acts as a confounder only when observed. Thus for the MPA to be an appropriate method to use, we must believe that the relationships between treatment, outcome, and confounder are different in the subgroup with the confounder unmeasured. While this assumption will be plausible only in specific scenarios, one setting where it may have broad applicability is in the area of electronic health record research. In such studies, missing confounder information reflects information that the clinician did not have when making prescribing decisions, thus the assumption that the missing values did not affect prescribing may well be reasonable.

If this key assumption is thought to be satisfied, careful consideration is required to ensure that the remaining assumptions of the MPA are satisfied. In particular, the

assumptions do not hold in the following scenarios: (i) where the outcome affects missingness of the confounder; (ii) where outcome and missingness have shared unmeasured common causes, and treatment and missingness have shared common causes; and (iii) where a partially-missing confounder and treatment both affect missingness of the confounder and the confounder is thought to be associated with outcome whether or not it is measured. We note that the scenario where the outcome affects the missingness of the confounder also gives biased estimates of the treatment effect when using complete records analysis [50]; multiple imputation can be used to deal with such scenarios if data are missing at random.

We also found that many violations of the MPA's assumptions can be dealt with by recording, and including in the analysis, auxiliary variables that are predictors of confounder missingness. Thus, although measuring such variables may be difficult in practice, careful consideration of the process by which data become missing is essential.

Our results demonstrate that classification of the missingness mechanism according to Rubin's taxonomy does not provide information as to whether the MPA's assumptions will hold. Unlike most missing data methods, for example, data being missing completely at random does not guarantee that the assumptions of the MPA are satisfied: the underlying relationships between the partially missing confounder and either the treatment or outcome (or both) would need to differ according to whether or not the confounder was missing. Also, if a confounder is missing not at random, but the confounder does not confound the treatment-outcome relationship when missing, the MPA's assumptions may hold.

The missing indicator approach is a popular and easy method to deal with missing confounder data. [57, 58] However, it is believed to be an 'ad hoc' method [57] that produces biased results. [58] Although the missing indicator approach is indeed biased under standard missing at random assumptions, [58] our results show that in the propensity score context, the missing indicator approach is a simplified version of the MPA, and hence requires the same assumptions for valid results, along with additional assumptions about interaction terms in the propensity score model.

Our work, therefore, allows researchers to use the missing indicator approach in a principled way.

There are several advantages to using the MPA, or the simpler missing indicator approach, when dealing with partially observed confounders in propensity score analysis. First, the method itself is simple to comprehend and easy to implement. Second, in contrast to complete records analysis, the MPA retains all patients in the analysis. Third, the MPA may be appropriate in some situations where multiple imputation is not, as the MPA does not require the missing at random assumption to hold.

A limitation of the MPA is that we require sufficient sample size in each missingness pattern in order to be able to estimate propensity scores. This is of particular concern when there are many missingness patterns, a scenario to which the MPA is not currently easily extendable. Qu and Lipkovich (2009) suggested a pattern pooling algorithm [70] to ensure sufficient sample size when estimating propensity scores when there are a large number of missingness patterns. Further work is needed to explore the performance of their algorithm in a range of scenarios. An extension to the MPA was proposed by D'Agostino et al. [64] They suggested that in each missingness pattern, propensity scores should be estimated in the wider group of all subjects with observed data for the relevant confounders, retaining estimated propensity scores only for those who actually observed that particular pattern. Further work is required to compare this extension with the original MPA, and to investigate how to account for the correlation induced by this method.

In scenarios with a large number of confounders, causal diagrams may be prohibitively complex to construct. An alternative strategy could be to perform sensitivity analyses to assess the extent of the violation that would be required to change the study's conclusions. However, further work is required to determine how best to implement such sensitivity analyses.

We have concentrated on scenarios where treatment and outcome are both fully observed. A hybrid method, combining the MPA and multiple imputation, was proposed by Qu and Lipkovich, [70] and studied by Seaman and White. [80]

The MPA is simple and easy to implement, and may be useful in settings where other missing confounder data methods are not appropriate. We believe that this approach will be particularly useful in areas using routinely collected data, particularly electronic health record research. We have produced practical guidance for researchers to decide whether the underlying assumptions of the MPA are plausibly satisfied in a particular clinical setting.

Acknowledgements

This work was supported by the Economic and Social Research Council [Grant Number ES/J5000/21/1]; the Medical Research Council [Project Grant MR/M013278/1]; and by Health Data Research UK [Grant Number EPNCZO90], which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and Wellcome. Ethics approval was given by the London School of Hygiene and Tropical Medicine Research Ethics Committee [Reference: 15880] and by the Clinical Practice Research Datalink Independent Scientific Advisory Committee [ISAC Protocol Number 14_208A2].

Supplementary material

In Section A, we prove that the missingness pattern approach (MPA) gives a consistent estimator of the average treatment effect under weaker versions of Mattei's assumptions, [63] referred to in the main text as the mSITA, CIT and CIO assumptions. In Section B, we explore the connection between the MPA and the missing indicator approach by comparing propensity score models for the two approaches. Section C describes the d-separation rule. Section D gives a brief overview of twin networks. In Section E, we present additional violations of the MPA's assumptions.

In Section F, we provide R code for assessing the MPA's assumptions in a simple example and in our motivating example. Section G gives standardized differences for our motivating example.

A Validity of the MPA

In this appendix, we demonstrate that $E\left[\frac{ZY}{e^*}\right] = E[Y(1)]$ under the weaker versions of the assumptions presented in the text.

First, using the consistency assumption and rearranging, we have that:

$$E\left[\frac{ZY}{e^*}\right] = E\left[\frac{ZY(1)}{e^*}\right] = E\left[E\left[\frac{ZY(1)}{e^*}|X_{obs}, R\right]\right]$$
$$= E\left[\frac{1}{e^*}E\big[ZY(1)|X_{obs}, R\big]\right], \tag{5.6}$$

where $e^* = E[Z|X_{obs}, R]$.

Switching briefly to summation notation:

$$\begin{aligned}
&E\big[ZY(1)|X_{obs}, R\big] \\
&= \sum\sum zyP(Z|X_{obs}, R)P(Y(1)|Z, X_{obs}, R) \\
&= \sum\sum zyP(Z|X_{obs}, R)\sum P(Y(1), X_{mis}|Z, X_{obs}, R) \\
&= \sum\sum\sum zyP(Z|X_{obs}, R)P(Y(1)|Z, X_{mis}, X_{obs}, R)P(X_{mis}|Z, X_{obs}, R)
\end{aligned}$$

Using mSITA ($Z \perp Y(z)|X, R$ for $z = 0, 1$) and CIT ($Z \perp X_{mis}|X_{obs}, R$), we have:

$$\begin{aligned}
&E\big[ZY(1)|X_{obs}, R\big] \\
&= \sum\sum\sum zyP(Z|X_{obs}, R)P(Y(1)|X_{mis}, X_{obs}, R)P(X_{mis}|X_{obs}, R) \\
&= \sum\sum zyP(Z|X_{obs}, R)\sum P(Y(1), X_{mis}|X_{obs}, R) \\
&= \sum\sum zyP(Z|X_{obs}, R)P(Y(1)|X_{obs}, R) \\
&= E\big[Z|X_{obs}, R\big]E\big[Y(1)|X_{obs}, R\big]
\end{aligned}$$

We can also show that $E\big[ZY(1)|X_{obs}, R\big] = E\big[Z|X_{obs}, R\big]E\big[Y(1)|X_{obs}, R\big]$ using mSITA with CIO ($Y(z) \perp X_{mis}|X_{obs}, R$ for $z = 0, 1$) in a similar manner. Thus, we can rewrite equation 5.6 as follows:

$$E\left[\frac{ZY}{e^*}\right] = E\left[\frac{1}{e^*}E\big[Z|X_{obs}, R\big]E\big[Y(1)|X_{obs}, R\big]\right].$$

Since $e^* = E[Z|X_{obs}, R]$:

$$E\left[\frac{ZY}{e^*}\right] = E\Bigg[E\big[Y(1)|X_{obs}, R\big]\Bigg] = E[Y(1)].$$

Similarly, we can show that $E[(1-Z)Y/(1-e^*)] = E[Y(0)]$. □

B The connection between the missingness pattern approach and the missing indicator approach

In this appendix, we consider the propensity score models for the MPA and the missing indicator approach to explore the connection between these approaches.

In a scenario with a single partially observed confounder X, the propensity score model for the MPA can be written as:

$$logit(P(Z=1)) = \begin{cases} \alpha_1 + \beta_1 X & \text{if } R=1 \\ \alpha_0 & \text{if } R=0 \end{cases}$$

with some parameters $\alpha_1, \beta_1, \alpha_0$.

Defining a new variable X^* which takes the value X if observed, and 0 otherwise, this can be rewritten as:

$$logit(P(Z=1)) = \alpha_0 + \beta_1 X^* R + (\alpha_1 - \alpha_0)R.$$

If X is binary, this is equivalent to creating a third category for X representing the missing values. If X is continuous, this sets missing values to 0 and adds an indicator variable for missing observations. This is exactly the missing indicator approach. If X were categorical, this could be extended to show that the MPA is similarly equivalent to adding a 'missing' category.

In a scenario with one partially observed confounder X, and one fully observed confounder C, the propensity score for the MPA can be written as:

$$\begin{aligned} logit(P(Z=1)) &= \begin{cases} \alpha_1 + \beta_1 X + & \gamma_1 C & \text{if } R=1 \\ \alpha_0 + & \gamma_0 C & \text{if } R=0 \end{cases} \\ &= \alpha_0 + \beta_1 X^* R + (\alpha_1 - \alpha_0)R + \gamma_0 C + (\gamma_1 - \gamma_0)CR \end{aligned}$$

with some parameters $\alpha_1, \beta_1, \gamma_1, \alpha_0, \gamma_0$.

In contrast, the propensity score model for the missing indicator approach is:

$$logit(P(Z = 1)) = \alpha + \beta X^* R + \eta R + \gamma C$$

with parameters $\alpha, \beta, \eta, \gamma$.

This is the MPA model, constraining γ_1 to be equal to γ_0, i.e. the missing indicator model additionally assumes there are no CR interactions in the true propensity score model.

C The d-separation rule

The d-separation rule, proposed within the context of directed acyclic graphs [66] and extended to SWITs, [67] determines whether a particular conditional dependency holds or not, under an assumed causal structure. Broadly speaking, association is transmitted through series of arrows — paths — in the assumed causal diagram. [4] A particular path will transmit association between the nodes at either end unless it contains a 'collider': a node which — in that path — has two incoming arrows. In Figure 5.1, the path $Z \leftarrow X \rightarrow Y(z)$ will transmit association between Z and $Y(z)$, but the path $Z \leftarrow U_Z \rightarrow R \leftarrow U_Y \rightarrow Y(z)$ will not because R is a collider in this path. Conditioning on a non-collider blocks associations through a specific path. Conversely, conditioning on a collider removes the blockage through that collider thereby allowing association to be transmitted. Introducing bias by conditioning on a collider is often termed collider bias. [81]

The d-separation rule states that two variables in the assumed causal diagram are conditionally independent given a set of variables V if for each path connecting the two variables: (i) the path contains two arrows which collide at a node in the path, and that node is neither in V, nor a cause of a variable in V; or (ii) the path has a non-collider which is in V. [4, 66]

In Figure 5.1, there are two paths between Z and $Y(z)$: $Z \leftarrow X \rightarrow Y(z)$, and $Z \leftarrow U_Z \rightarrow R \leftarrow U_Y \rightarrow Y(z)$. If the conditioning set is $V = \{X\}$, then Z and $Y(z)$ are conditionally independent given V. This is because the first path contains a non-collider (X) which is in V (condition (ii)) and the second contains a collider (R) which is not in V (condition (i)). In contrast, Z and $Y(z)$ are not conditionally independent given $V = \{X, R\}$, because the second path then contains a collider (i.e. R) which is in V, and neither X nor R is a non-collider in this path.

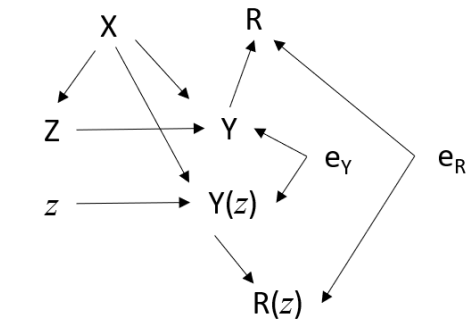


Figure 5.6: A simple example of a twin network.
X: partially observed confounder. Z: observed treatment allocation. Y: observed outcome. $Y(z)$: potential outcome resulting from intervening to set treatment to value z. R: observed missing indicator (=1 if X observed, =0 if X is missing). $R(z)$: potential missing indicator (=1 if X observed in counterfactual world, =0 if X is missing in counterfactual world). e_Y: unobserved error term between Y and $Y(z)$. e_R: unobserved error term between R and $R(z)$.

D Twin networks

When considering scenarios in which treatment, or the outcome, has a causal effect on missingness, by construction, the SWITs now include $R(z)$ instead of R. This means that the SWIT can no longer be used to test the MPA's assumptions. Instead, we can construct twin networks to check such scenarios, as described by Balke and Pearl, [68] and Shpitser and Pearl. [69]

Briefly, a twin network can be constructed from a directed acyclic graph, which involves real world variables and relationships, by adding counterparts of variables and relationships in the counterfactual world where treatment has been intervened upon to be set to some realisation of the random variable Z.

For example, Figure 5.6 shows a simple twin network of a scenario where the confounder X has a causal effect on both treatment and outcome, treatment has a causal effect on outcome, and outcome has a causal effect on missingness of the confounder. The 'real world' is shown by Z, Y, and R. The nodes z, $Y(z)$, and $R(z)$ show the counterfactual world – what would occur if we set treatment to value z. The observed outcome Y and potential outcome $Y(z)$ are connected by an unobserved error term e_Y. Similarly, the observed missing indicator R and potential missing indicator $R(z)$ are connected by an unobserved error term e_R. Because X has a causal effect on outcome, it also has a causal effect on the potential outcome.

It does not, however, affect the intervened-on value of treatment, z.

To assess mSITA in Figure 5.6, we need to assess whether $Z \perp Y(z)|R, X$. Conditioning on X blocks the confounding pathway between Z and $Y(z)$. There is a closed path $Z \rightarrow Y \leftarrow e_Y \rightarrow Y(z)$, blocked because Y is a collider on this path. However, conditioning on R opens this path, because conditioning on a descendant of a collider (i.e. something affected by the collider) has a similar, but weaker, effect as conditioning on the collider itself. Thus the path $Z \rightarrow Y \leftarrow e_Y \rightarrow Y(z)$ is open, after conditioning on R, so the mSITA assumption may not be appropriate here. [1]

Dagitty can be used to assess the assumptions in twin networks, just as for SWITs.

[1] As d-separation for twin networks is not complete, [67] caution should be used in considering the plausibility of results that suggest two variables are not d-separated.

E Additional violations of assumptions

Figure 5.7 summarises additional violations of the MPA's assumptions.

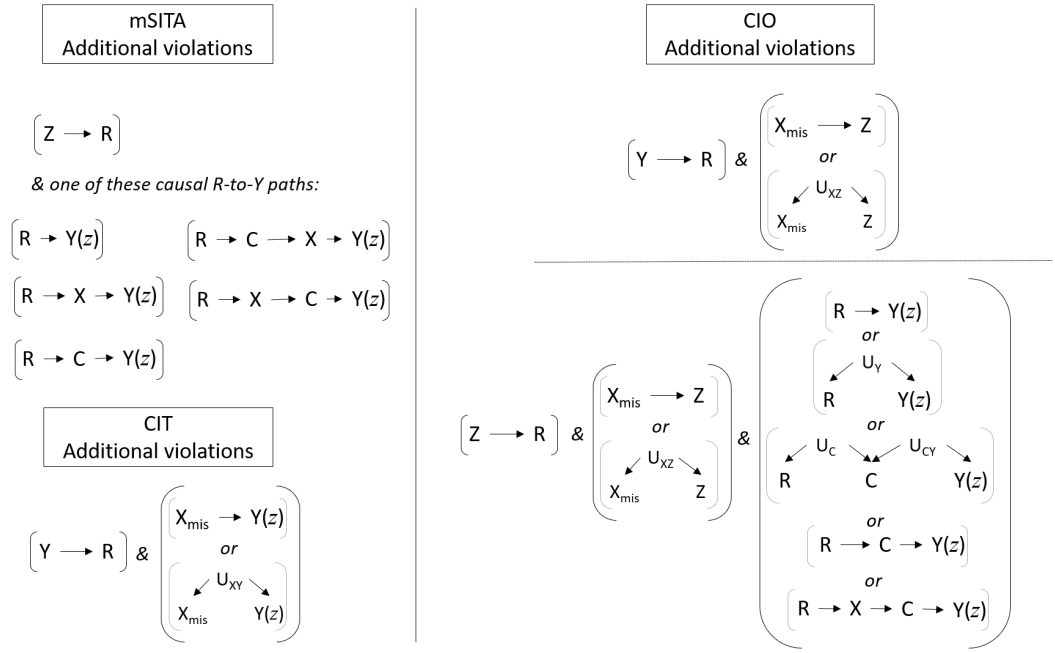


Figure 5.7: Summary of additional violations of the mSITA, CIT and CIO assumptions. X: partially observed confounder. X_{mis}: unobserved confounder values. C: fully observed confounder. Z: treatment. $Y(z)$: potential outcome resulting from intervening to set Z equal to a particular value z. Y: observed outcome. R: missing indicator (=1 if X observed, =0 if X is missing). U_{st}: unobserved common cause between two variables s and t. U_s: unobserved common cause between R and another variable s.

F Using Dagitty to assess the MPA's assumptions

F.1 Simple example: R code to use Dagitty to assess the MPA's assumptions

Run in R 3.4.0, [82] the R code below reads in our causal diagram for Figure 5.1 and uses d-separation to assess the mSITA, CIT and CIO assumptions. [77]

```
### R CODE TO USE DAGITTY: SIMPLE EXAMPLE


install.packages("dagitty")
library("dagitty")


############################
#   Load DAG into Dagitty  #
############################


# X       partially observed confounder
# R       observed covariate indicator: =1 if X observed, =0 otherwise.
# Z       treatment allocation (fully observed)
# Yz      potential outcome that would be observed when Z=z
# U_Y     unobserved common cause of R and Y
# U_Z     unobserved common cause of R and Z


g1 <- dagitty( 'dag {
z -> Yz
Z <- X -> Yz
R <- U_Z -> Z
R <- U_Y -> Yz
}')
coordinates( g1 ) <-
```

```
list( x=c(Z=1, z=1.2, X=2, Yz=3, R=2, U_Z=1.2, U_Y=2.8),
y=c(Z=3, z=3,   X=2, Yz=3, R=1, U_Z=1.7, U_Y=1.7) )
plot( g1 )

### Assess mSITA assumption:
### - Is Z indep of Yz given R, X, and z?
###      (note: we add z to the conditioning set because we are using
###  a SWIG)

# List all paths between Z and Yz
paths( g1, "Z", "Yz", c("R","X","z") )

# Check whether mSITA holds
dseparated( g1, "Z", "Yz", c("R","X", "z") )

# Check whether mSITA holds if U_Y were also measured and
# included in the confounder set
dseparated( g1, "Z", "Yz", c("R","X", "z", "U_Y") )



########################################
#   DAG for subgroup with X unmeasured  #
########################################

### Suppose we believe that X does not affect Z when unmeasured
### R is now written R0 as shorthand for "R=0"

g2 <- dagitty( 'dag {
z -> Yz
X -> Yz
```

```
R0 <- U_Z -> Z
R0 <- U_Y -> Yz
}')
coordinates( g2 ) <-
list( x=c(Z=1, z=1.2, X=2, Yz=3, R0=2, U_Z=1.2, U_Y=2.8),
y=c(Z=3, z=3,   X=2, Yz=3, R0=1, U_Z=1.7, U_Y=1.7) )
plot( g2 )

### Assess CIT assumption:
### - Is Z indep of X given R=0  (and z)?
###      (note: we add z to the conditioning set because we are using
###  a SWIG)

dseparated( g2, "Z", "X", c("R0", "z") )
paths( g2, "Z", "X", c("R0","z"))

### Assess CIO assumption:
### - Is Yz indep of X given R=0  (and z)?
###      (note: we add z to the conditioning set because we are using
###  a SWIG)

dseparated( g2, "Yz", "X", c("R0", "z") )
paths( g2, "Yz", "X", c("R0","z"))
```

F.2 Motivating example: R code to use Dagitty to assess the MPA's assumptions

Figure 5.5 shows the causal diagram which represents what the investigators believe to represent the underlying causal structure giving rise to the data. The R code below reads in our causal diagram for our motivating example and uses d-separation to assess the mSITA, CIT and CIO assumptions.

```
### R CODE TO USE DAGITTY: MOTIVATING EXAMPLE

#install.packages("dagitty")
library("dagitty")

############################
#   Load DAG into Dagitty  #
############################

# Outcome and treatment:
#   Aki   Acute Kidney Injury (outcome)
#   Ace   ACE/ARB (treatment)
#   ace   Intervened-on ACE/ARB (intervened-on treatment)

# Partially observed confounders and missing indicators:
#   Eth   Ethnicity (partially observed confounder)
#   Ckd   Baseline CKD (partially observed confounder)
#   Reth  Missingness of ethnicity
#   Rckd  Missingness of baseline CKD

# Determinants of missing data:
#   Slf   Service-level factors determining whether or not ethnicity
#         is measured
#   Hosp  Hospitalisation

# Fully observed confounders:
#   Age
#   Sex
#   Hyp   Hypertension
```

```
#   Diab  Diabetes
#   Arr   Arrhythmia
#   Car   Cardiac failure
#   Ihd   Ischaemic heart disease

# Unmeasured factors:
#   U     (e.g. frailty)

###   Draw DAG  ###
g1 <- dagitty( 'dag {
Age -> Hyp Age -> Diab Age -> Ckd Age -> Arr Age -> Car Age -> Ihd
Sex -> Hyp Sex -> Diab Sex -> Ckd Sex -> Arr Sex -> Car Sex -> Ihd
Reth <- Eth Reth <- Slf Reth <- Hosp
Rckd <- Hyp Rckd<- Ckd Rckd <- Diab Rckd <- Age
Diab -> Ckd Ihd -> Ckd Car -> Ckd
Eth -> Arr Ihd -> Arr Arr -> Car Hyp -> Car Ihd -> Car
U -> Ckd U -> Hyp U -> Diab U -> Hosp U -> Ihd U-> Arr
Hyp -> Ace Sex -> Ace Diab -> Ace Eth -> Ace Ckd -> Ace Car -> Ace
Ihd -> Ace
Age -> Aki Eth -> Aki Sex -> Aki Diab -> Aki Ckd -> Aki U -> Aki
Car -> Aki
ace -> Aki
}')
coordinates( g1 ) <-
list( x=c(Age=1, Sex=1, Eth=1, Ace=6,    ace=6.5,  Aki=10,
Arr=5, Car=3.25, Ihd=5,
Reth=9,   Slf=8, Hosp=8, Hyp=3.25, Ckd=4, Diab=4, Rckd=5, U=1.5),
y=c(Age=3, Sex=5, Eth=8, Ace=4.75, ace=4.75, Aki=5,
Arr=7, Car=6.75, Ihd=6,
Reth=7.5, Slf=8, Hosp=7, Hyp=2,    Ckd=3, Diab=4, Rckd=2, U=6) )
```

```
plot( g1 )


############################
#   Check mSITA assumption  #
############################

### mSITA assumption:
###    Is Z indep of Yz given R, X, and z?
### Here:  Is Ace indep of Aki given Rckd, Reth, Ckd, Eth, ...
### ...Age, Sex, Hyp, Diab, Arr, Car, Ihd and ace?
###

# List all paths between Z and Yz
paths( g1, "Ace", "Aki", c("Rckd", "Reth","Ckd", "Eth",
"Age", "Sex", "Hyp", "Diab",
"Arr", "Car", "Ihd", "ace") )

# Check whether mSITA holds
dseparated( g1, "Ace", "Aki", c("Rckd", "Reth","Ckd", "Eth",
"Age", "Sex", "Hyp", "Diab",
"Arr", "Car", "Ihd", "ace") )


#########################################################
#   DAG for subgroup with CKD and ethnicity unmeasured  #
#########################################################


### Suppose we believe that:
```

```
###     Ckd does not affect prescription of ACE when unmeasured
###     Eth does not affect prescription of ACE when unmeasured

###   Draw DAG  (group with neither ethnicity nor ckd measured) ###
g2 <- dagitty( 'dag {
Age -> Hyp Age -> Diab Age -> Ckd Age -> Arr Age -> Car Age -> Ihd
Sex -> Hyp Sex -> Diab Sex -> Ckd Sex -> Arr Sex -> Car Sex -> Ihd
Reth <- Eth Reth <- Slf Reth <- Hosp
Rckd <- Hyp Rckd<- Ckd Rckd <- Diab Rckd <- Age
Diab -> Ckd Ihd -> Ckd Car -> Ckd
Eth -> Arr Ihd -> Arr Arr -> Car Hyp -> Car Ihd -> Car
U -> Ckd U -> Hyp U -> Diab U -> Hosp U -> Ihd U-> Arr
Hyp -> Ace Sex -> Ace Diab -> Ace Car -> Ace Ihd -> Ace
Age -> Aki Eth -> Aki Sex -> Aki Diab -> Aki Ckd -> Aki U -> Aki
Car -> Aki
ace -> Aki
}')
coordinates( g2 ) <-
list( x=c(Age=1, Sex=1, Eth=1, Ace=6,    ace=6.5,  Aki=10,
Arr=5, Car=3.25, Ihd=5,
Reth=9,   Slf=8, Hosp=8, Hyp=3.25, Ckd=4, Diab=4, Rckd=5, U=1.5),
y=c(Age=3, Sex=5, Eth=8, Ace=4.75, ace=4.75, Aki=5,
Arr=7, Car=6.75, Ihd=6,
Reth=7.5, Slf=8, Hosp=7, Hyp=2,    Ckd=3, Diab=4, Rckd=2, U=6) )
plot( g2 )


################################
#   Check CIT/CIO assumption  #
################################
```

```
### CIT assumption:
###   Is Z indep of X given R=0  (and z)?
###   Here: is Ace indep of Ckd given Rckd=0 and Reth=0, conditional
###   on: Age, Sex, Hyp, Diab  (and ace)?
###   Here: is Ace indep of Eth given Rckd=0 and Reth=0, conditional
### on: Age, Sex, Hyp, Diab  (and ace)?

# Check whether CIT holds
dseparated( g2, "Ace", "Ckd", c("Rckd", "Reth",
"Age", "Sex", "Hyp", "Diab",
"Arr", "Car", "Ihd", "ace") )
dseparated( g2, "Ace", "Eth", c("Rckd", "Reth",
"Age", "Sex", "Hyp", "Diab",
"Arr", "Car", "Ihd", "ace") )

### CIO assumption:
###   Is Yz indep of X given R=0  (and z)?
###   Here: is Aki indep of Ckd given Rckd=0 and Reth=0, conditional
### on: Age, Sex, Hyp, Diab  (and ace)?
###   Here: is Aki indep of Eth given Rckd=0 and Reth=0, conditional
### on: Age, Sex, Hyp, Diab  (and ace)?

# Check whether CIO holds
dseparated( g2, "Aki", "Ckd", c("Rckd", "Reth",
"Age", "Sex", "Hyp", "Diab",
"Arr", "Car", "Ihd", "ace") )
dseparated( g2, "Aki", "Eth", c("Rckd", "Reth",
"Age", "Sex", "Hyp", "Diab",
```

```
"Arr", "Car", "Ihd", "ace") )



### Use similar steps to check CIT/CIO in other missingness pattern
### subgroups
```

G Balance of confounders in motivating example

In Table 5.3, we present standardized differences [25] calculated to assess the balance of confounders in our motivating example.

Table 5.3: Standardised mean differences of confounders, before and after inverse probability of treatment weighting for complete records analysis (CRA), missingness pattern approach (MPA), missing indicator approach (MIndA), and multiple imputation (MI). A standardized difference greater than 10% indicates imbalance for that variable. (* Standardized differences for multiple imputation were averaged over 10 imputed datasets.)

		Percentage standardized differences (absolute values)				
Covariate		In original sample	After CRA	After MPA	After MIndA	After MI*
Age (years)	18 to 42					
	43 to 53	14.64	1.33	0.49	0.36	0.26
	54 to 62	9.42	1.64	1.51	1.45	2.26
	63 to 71	2.48	2.17	2.11	1.98	2.93
	$\geq$ 72	4.07	2.25	3.70	3.60	4.81
Sex	Female	36.95	1.92	4.44	4.99	4.66
Chronic	$\leq$ Stage 2					
Kidney	Stage 3a	12.84	1.77	1.13	1.08	1.27
Disease	Stage 3b	8.62	0.07	0.35	0.38	4.56
	Stage 4	3.33	0.32	0.19	0.16	1.19
Ethnicity	White					
	South Asian	6.31	0.38	0.65	0.65	7.63
	Black	3.14	3.75	3.75	4.22	8.30
	Mixed	1.73	0.56	0.69	0.84	4.25
	Other	0.43	0.42	<0.01	0.01	1.12
Diabetes Mellitus		49.21	0.43	2.70	2.01	3.06
Ischaemic Heart Disease		19.25	2.52	2.83	2.28	6.00
Arrhythmia		4.84	0.68	2.16	3.07	2.04
Cardiac Failure		32.90	1.75	0.02	0.03	0.19
Hypertension		43.08	5.74	7.85	7.88	10.93

Chapter 6

Variance estimation for the missingness pattern approach

6.1 The theory of M-estimation

M-estimation provides a generalisable theory to obtain the large-sample variance for estimators that can be written as the solution to a set of estimating equations [83].

Suppose the observed data for individual i are $\mathbf{Y}_i$, and the data are independent and identically distributed according to distribution function F. A M-estimator is the solution, $\hat{\theta}$, to the estimating equations

$$\sum_{i=1}^{n} \psi(\mathbf{Y}_i, \theta) = 0.$$

So $\hat{\theta}$ is defined as the value that solves:

$$\sum_{i=1}^{n} \psi(\mathbf{Y}_i, \hat{\theta}) = 0.$$

The true value of the parameter, θ_0, is defined by

$$E_F[\psi(Y, \theta_0)] = \int \phi(y, \theta_0) dF(y) = 0.$$

Then the large-sample approximate distribution of the estimator $\hat{\theta}$ is:

$$\hat{\theta} \sim MVN\left(\theta_0, \frac{V(\theta_0)}{n}\right), \qquad \text{as} \qquad n \to \infty$$

where

$$V(\theta_0) = A(\theta_0)^{-1} B(\theta_0) \{A(\theta_0)^{-1}\}^T$$

with

$$A(\theta_0) = E\left[-\frac{\partial}{\partial \theta^T}\psi(Y_1, \theta_0)\right], \qquad B(\theta_0) = E[\psi(Y_1, \theta_0)\psi(Y_1, \theta_0)^T].$$

To obtain an estimate of the variance, the matrices $A(\theta_0)$ and $B(\theta_0)$ can be replaced by sample estimates of the relevant quantities, i.e..

$$\hat{A}(\theta_0) = -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \theta^T}\psi(Y_i, \hat{\theta}), \qquad \hat{B}(\theta_0) = \frac{1}{n}\sum_{i=1}^{n} \psi(Y_i, \hat{\theta})\psi(Y_i, \hat{\theta})^T.$$

6.2 Estimating the variance of the IPTW estimator with MPA for a partially missing confounder

This chapter considers a simplified setting with a single potential confounder, X, which is partially observed. The variable R indicates whether the confounder is observed ($R = 1$) or missing ($R = 0$).

6.2.1 The IPTW estimator with MPA as an M-estimator

If the propensity scores were given by π_i for individual i, then the two means which are contrasted to give the IPTW treatment effect estimator can be written as the solution to the estimating equations

$$\sum_{i=1}^{n} \mathbf{v}_i = 0$$

where

$$\mathbf{v}_i = \begin{pmatrix} (Y_i - \mu_1) Z_i \pi_i^{-1} \\ (Y_i - \mu_0)(1 - Z_i)(1 - \pi_i)^{-1} \end{pmatrix}$$

The treatment effect estimate is given by $\hat{\mu}_1 - \hat{\mu}_0$.

In fact, the propensity scores are themselves estimated, often via a logistic regression model for the treatment with the potential confounders as explanatory variables. Thus

$$\sum_{i=1}^{n} \mathbf{w}_i = 0$$

where, letting $\mathbf{x}_i = (1, X_i)^T$, and using $expit(.)$ to denote the function $expit(x) = exp(x)/(1 + exp(x))$, we have

$$\mathbf{w}_i = \begin{pmatrix} R_i \mathbf{x}_i (Z_i - expit(\lambda^T \mathbf{x}_i)) \\ (1 - R_i)(Z_i - expit(\zeta)) \end{pmatrix}$$

Putting these two estimation steps together, the two estimated means are obtained by solving the estimating equations

$$\sum_{i=1}^{n} \mathbf{u}_i = 0$$

where

$$\mathbf{u}_i = \begin{pmatrix} (Y_i - \mu_1) Z_i \pi_i(R_i, X_i, \lambda, \zeta)^{-1} \\ (Y_i - \mu_0)(1 - Z_i)(1 - \pi_i(R_i, X_i, \lambda, \zeta))^{-1} \\ R_i \mathbf{x}_i (Z_i - expit(\lambda^T \mathbf{x}_i)) \\ (1 - R_i)(Z_i - expit(\zeta)) \end{pmatrix}$$

with

$$\pi_i(R_i, X_i, \lambda, \zeta) = R_i \times expit(\lambda^T \mathbf{x}_i) + (1 - R_i) \times expit(\zeta)$$

The parameter being estimated by solving the set of estimating equations is $\theta = (\mu_1, \mu_0, \lambda^T, \zeta)^T$, and the parameter of interest — the estimated treatment effect — is given by $\delta = \mu_1 - \mu_0$.

6.2.2 Large sample variance

Partitioning the matrix B into four components, corresponding to the four components of $\mathbf{u}$, we can write

$$B(\theta_0) = \begin{pmatrix} b_{11} & 0 & b_{13} & b_{14} \\ 0 & b_{22} & b_{23} & b_{24} \\ b_{13}^T & b_{23}^T & b_{33} & 0 \\ b_{14}^T & b_{24}^T & 0 & b_{44} \end{pmatrix}$$

where the zeros arise from multiplying Z and $(1-Z)$ or R and $(1-R)$.

The matrix A can also be partitioned in a similar manner:

$$A(\theta_0) = \left(-E\left[\frac{\partial \mathbf{u}}{\partial \mu_1}\right], -E\left[\frac{\partial \mathbf{u}}{\partial \mu_0}\right], -E\left[\frac{\partial \mathbf{u}}{\partial \lambda^T}\right], -E\left[\frac{\partial \mathbf{u}}{\partial \zeta}\right] \right)$$

giving

$$A(\theta_0) = \begin{pmatrix} a_{11} & 0 & a_{13} & a_{14} \\ 0 & a_{22} & a_{23} & a_{24} \\ 0 & 0 & a_{33} & 0 \\ 0 & 0 & 0 & a_{44} \end{pmatrix}$$

where the zeros arise from differentiating with respect to a parameter not appearing in the relevant part of the estimating equation.

The inverse matrix is given by

$$A(\theta_0)^{-1} = \begin{pmatrix} a_{11}^{-1} & 0 & -a_{13}a_{11}^{-1}a_{33}^{-1} & -a_{14}a_{11}^{-1}a_{44}^{-1} \\ 0 & a_{22}^{-1} & -a_{23}a_{22}^{-1}a_{33}^{-1} & -a_{24}a_{22}^{-1}a_{44}^{-1} \\ 0 & 0 & a_{33}^{-1} & 0 \\ 0 & 0 & 0 & a_{44}^{-1} \end{pmatrix}$$

The large-sample approximate variance of the estimator $\hat{\theta} = (\hat{\mu}_1, \hat{\mu}_0, \hat{\lambda}^T, \hat{\zeta})^T$, is given by: $Var(\hat{\theta}) = A(\theta_0)^{-1}B(\theta_0)\{A(\theta_0)^{-1}\}^T$. The variance of the treatment effect

estimator $\hat{\delta} = \hat{\mu}_1 - \hat{\mu}_0$ is given by

$$Var(\hat{\delta}) = Var(\hat{\mu}_1) + Var(\hat{\mu}_0) - 2Cov(\hat{\mu}_1, \hat{\mu}_0)$$

Thus, to obtain an estimator for the variance, the following steps must be followed:

- Obtain sample estimates of the individual components of the matrices B and A
- Multiply out the matrices $\hat{A}(\hat{\theta})^{-1}\hat{B}(\hat{\theta})\{\hat{A}(\hat{\theta})^{-1}\}^T$
- Extract the variances $n\widehat{Var}(\hat{\mu}_1)$ and $n\widehat{Var}(\hat{\mu}_0)$, as the $(1,1)$ and $(2,2)$ entries of the matrix obtained. Similarly, extract $\widehat{Cov}(\hat{\mu}_1, \hat{\mu}_0)$ as the $(1,2)$ component of the matrix obtained.
- Substitute into the equation to obtain $\widehat{Var}(\hat{\delta})$

6.2.3 Estimating the matrix B

The components of this matrix can be estimated by:

$$\hat{b}_{11} = \frac{1}{n}\sum_{i=1}^{n}(Y - \hat{\mu}_1)^2 Z_i \hat{\pi}_i^{-2}$$

$$\hat{b}_{22} = \frac{1}{n}\sum_{i=1}^{n}(Y - \hat{\mu}_0)^2 (1 - Z_i)(1 - \hat{\pi}_i)^{-2}$$

$$\hat{b}_{13} = \frac{1}{n}\sum_{i=1}^{n}\{R_i \mathbf{x}_i (Y_i - \hat{\mu}_1) Z_i (1 - expit(\hat{\lambda}^T \mathbf{x}_i))\}\hat{\pi}_i^{-1}$$

$$\hat{b}_{14} = \frac{1}{n}\sum_{i=1}^{n}\{(1 - R_i)(Y_i - \hat{\mu}_1) Z_i (1 - expit(\hat{\zeta}))\}\hat{\pi}_i^{-1}$$

$$\hat{b}_{23} = -\frac{1}{n}\sum_{i=1}^{n}\{R_i \mathbf{x}_i (Y_i - \hat{\mu}_0)(1 - Z_i) expit(\hat{\lambda}^T \mathbf{x}_i)\}(1 - \hat{\pi}_i)^{-1}$$

$$\hat{b}_{24} = -\frac{1}{n}\sum_{i=1}^{n}\{(1 - R_i)\mathbf{x}_i (Y_i - \hat{\mu}_0)(1 - Z_i) expit(\hat{\zeta})\}(1 - \hat{\pi}_i)^{-1}$$

$$\hat{b}_{33} = \frac{1}{n}\sum_{i=1}^{n} R_i \mathbf{x}_i \mathbf{x}_i^T (Z_i - expit(\hat{\lambda}^T \mathbf{x}_i))$$

$$\hat{b}_{44} = \frac{1}{n}\sum_{i=1}^{n} (1 - R_i)(Z_i - expit(\hat{\zeta}))$$

with

$$\hat{\pi}_i = R_i \times expit(\hat{\lambda}^T \mathbf{x}_i) + (1 - R_i) \times expit(\hat{\zeta})$$

6.2.4 Estimating the matrix A

The components of the matrix A can be estimated as follows:

$$\hat{a}_{11} = -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \mu_1}\{(Y_i - \mu_1) Z_i \pi_i (R_i, X_i, \lambda, \zeta)^{-1}\} = \frac{1}{n}\sum_{i=1}^{n} \frac{Z_i}{\hat{\pi}_i}$$

Similarly,

$$\hat{a}_{22} = \frac{1}{n}\sum_{i=1}^{n} \frac{(1 - Z_i)}{(1 - \hat{\pi}_i)}$$

And

$$\begin{aligned}
\hat{a}_{33} &= -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \lambda^T}\{R_i \mathbf{x}_i (Z_i - expit(\lambda^T \mathbf{x}_i))\} \\
&= -\frac{1}{n}\sum_{i=1}^{n} R_i \mathbf{x}_i \mathbf{x}_i^T expit(\lambda^T \mathbf{x}_i)(1 - expit(\lambda^T \mathbf{x}_i))
\end{aligned}$$

Similarly,

$$\hat{a}_{44} = -\frac{1}{n}\sum_{i=1}^{n} (1 - R_i) expit(\zeta)(1 - expit(\zeta))$$

We also have

$$\hat{a}_{13} = \frac{1}{n}\sum_{i=1}^{n} R_i \mathbf{x}_i^T (Y_i - \hat{\mu}_1) Z_i \hat{\pi}_i^{-1} (1 - \hat{\pi}_i)$$

$$\hat{a}_{23} = -\frac{1}{n}\sum_{i=1}^{n} R_i \mathbf{x}_i^T (Y_i - \hat{\mu}_0)(1 - Z_i)\hat{\pi}_i (1 - \hat{\pi}_i)^{-1}$$

$$\hat{a}_{14} = \frac{1}{n}\sum_{i=1}^{n} R_i(Y_i - \hat{\mu}_1)Z_i(1 - R_i)\hat{\pi}_i^{-1}(1 - \hat{\pi}_i)$$

$$\hat{a}_{24} = \frac{1}{n}\sum_{i=1}^{n} R_i(Y_i - \hat{\mu}_0)(1 - Z_i)(1 - R_i)\hat{\pi}_i(1 - \hat{\pi}_i)^{-1}$$

6.3 Plans to evaluate and extend the variance formula

Further work will involve simulation studies, including a single partially observed confounder, in order to assess how well the large-sample variance formula performs in finite — particularly in small — sample sizes.

The variance formula above is immediately extendable to the case with multiple confounders, in the simplified setting where confounders are either all missing simultaneously or all observed. Future work will extend the variance formula to the case where multiple confounders are partially missing and others fully observed, with all possible combinations of the partially missing confounders being missing or observed.

In this latter scenario, sparsity of data patterns mean that some sort of 'pattern pooling' is likely to be required, whereby similar patterns of missingness are grouped together. This will enable estimation of the propensity score within each of the larger missingness patterns.

Chapter 7

The connection between the missingness pattern approach and the missing indicator approach

7.1 The MPA's connection to the missing indicator approach in propensity score analysis

The missing indicator approach (MIA) is a simple missing data method where all missing values are set to a particular value, say 0, and a missingness indicator is included in the analysis model. In our cohort study, this is equivalent to adding an 'absent' category to baseline CKD stage for patients with missing data for this confounder.

In a scenario with a single partially observed confounder, it can be seen that the MIA is equivalent to the MPA. With a single partially observed confounder, the propensity score model for the MPA can be written as:

$$logit(P(Z=1)) = \begin{cases} \alpha^1 + \beta^1 X & \text{if } R=1 \\ \alpha^0 & \text{if } R=0 \end{cases}$$

with coefficients $\alpha^0, \alpha^1, \beta^1$. This can be rewritten as:

$$logit(P(Z=1)) = (\alpha^1 + \beta^1 X)R + \alpha^0(1-R),$$

which is equivalent to the propensity score model for the MIA.

Since the MPA and MIA are equivalent in this simple scenario with a single partially observed confounder, it is clear that the MIA also requires the mSITA assumption and at least one of the CIT and CIO assumptions to hold.

In order to see if this equivalency can be extended to more complex scenarios, we now consider the scenario with one partially observed confounder and one fully observed confounder, C.

The propensity score model for the MIA is now:

$$logit(P(Z=1)) = \alpha + \beta XR + \eta R + \gamma C,$$

with coefficients $\alpha, \beta, \eta, \gamma$.

The propensity score model for the MPA can be written as follows:

$$logit(P(Z=1)) = \begin{cases} \alpha^1 + \beta^1 X + & \gamma^1 C & \text{if } R=1 \\ \alpha^0 + & \gamma^0 C & \text{if } R=0 \end{cases}$$

with coefficients $\alpha^0, \alpha^1, \beta^1, \gamma^0, \gamma^1$. This can be rewritten as:

$$\begin{aligned} logit(P(Z=1)) = &\ (\alpha^1 + \beta^1 X + \gamma^1 C)R + (\alpha^0 + \gamma^0 C)(1-R) \\ = &\ \alpha^0 + \beta^1 XR + (\alpha^1 - \alpha^0)R + \gamma^0 C + (\gamma^1 - \gamma^0)CR. \end{aligned}$$

In this more complex scenario, we find that the models for the MIA and the MPA are not equivalent, as there is an additional term for the interaction between C and R. Hence, the MIA can be seen to be a simplified version of the MPA, where the effect of the fully observed confounder on treatment is assumed to be the same for all missingness patterns (i.e. the coefficient for the interaction term is zero). Before using the MIA in practice, in addition to considering the plausibility of the MPA's assumptions, researchers also need to check the assumption that there are no

interactions between fully observed confounders and the missing indicator. Unlike the mSITA, CIT and CIO assumptions, this interaction assumption can be assessed in the data at hand, and the propensity score model adapted as necessary to ensure correct specification.

7.2 How MIA relates to MPA with multiple partially observed confounders

Let Z denote treatment allocation and X, W denote two partially observed confounders with corresponding missing indicators R_X and R_W. We define X^*, which takes the value X if X is observed and 0 otherwise, and similarly define W^*.

With two partially observed confounders, the propensity score for the MPA can be written as

$$
\begin{aligned}
logit(P(Z=1)) &= \begin{cases} \alpha_{00} & \text{if } R_X=0 \ \& \ R_W=0 \\ \alpha_{10}+\beta_{10}X & \text{if } R_X=1 \ \& \ R_W=0 \\ \alpha_{01}+\gamma_{01}W & \text{if } R_X=0 \ \& \ R_W=1 \\ \alpha_{11}+\beta_{11}X+\gamma_{11}W & \text{if } R_X=1 \ \& \ R_W=1 \end{cases} \\
&= \alpha_{00}+(\alpha_{10}-\alpha_{00})R_X+(\alpha_{01}-\alpha_{00})R_W \\
&\quad +(\alpha_{00}-\alpha_{10}-\alpha_{01}+\alpha_{11})R_XR_W+\beta_{10}X^*R_X\gamma_{01}W^*R_W \\
&\quad +(\beta_{11}-\beta_{10})X^*R_XR_W+(\gamma_{11}-\gamma_{01})W^*R_XR_W.
\end{aligned}
$$

In contrast, the propensity score model for the MIA is:

$$
logit(P(Z=1)) = \alpha+\beta X^*R_X+\gamma W^*R_W+\delta R_X+\eta R_W.
$$

This is the MPA model, where the coefficents for terms involving interactions between missing indicators are zero. So, in general, the missing indicator method is a simplification of the MPA, which makes additional assumptions about the absence of interactions between the missingness indicator(s) and other fully-observed confounders. These additional assumptions can be assessed in the data by testing for

interactions in the propensity score model.

7.3 Motivation for extending from propensity score analysis to outcome regression

When investigating the missingness pattern approach (MPA), I found that, with a single partially observed confounder, the MPA was equivalent to the missing indicator approach (MIA), and thus the MIA would provide unbiased estimates under the mSITA assumption and either the CIT or the CIO assumption. With an additional fully observed confounder, I showed that the MIA is a simplification of the MPA, with the additional assumption that the propensity score model is correctly specified. In particular, the MIA as typically applied implicitly assumes the absence of interactions in the true propensity score model between the missing indicator and the fully observed confounder, i.e. the effect of the fully observed confounder on treatment does not vary by missingness pattern.

So, in the propensity score context, the MIA can provide unbiased estimates under certain assumptions. Indeed, the use of missing indicators has been recommended by Stuart (2010) for use in propensity score analysis [55] and also by Hernán et al. (2009) and Kreif et al. (2018) in the context of non-systematic monitoring of covariates in settings with time-varying treatments [74, 75]. However, in the context of outcome regression, the MIA is often considered to be an "ad hoc" approach [56, 57] that gives biased results [58, 59]. The purpose of this chapter is to investigate whether our finding that the MIA can provide unbiased estimates extends to the context of outcome regression.

In order to investigate whether our work in the propensity score context can be extended to outcome regression, we must first consider how these contexts differ from each other. Although in both cases the aim is to remove confounding bias, in propensity score analysis, we wish to model the relationship between covariates and the *treatment*, whilst in outcome regression, we wish to model the relationship between covariates and the *outcome*. Thus the MIA (or other methods to handle

partially observed confounders) would be applied differently for each context: in the propensity score context, the MIA is applied when modelling the covariate-treatment relationship, whereas for outcome regression, the MIA would be used in the outcome model. Consequently, when extending the MIA to outcome regression, instead of assuming correct specification of the propensity score model, the analogous assumption would be that the outcome model is correctly specified. In particular, we might expect that that the effect of fully observed confounders on the outcome does not vary by missingness pattern.

7.3.1 Relating our findings to previous literature

I prove that the MIA gives unbiased treatment effect estimates in outcome regression when (i) the mSITA assumption holds, (ii) either the CIT or the CIO assumption holds, and (iii) the outcome model is correctly specified. Details are given in the MIA research paper pre-print (Chapter 8).

Jones assumed that the true model for the outcome Y was a linear regression model with two covariates X_1, X_2 and an independent normal error term ϵ, where Y and X_1 are assumed to be fully observed and X_2 may be partially observed [59]. Jones (1996) proved that the MIA for outcome regression gives biased least squares estimators and noted that the least squares estimators are unbiased when (i) the proportion of individuals with missing data is zero, or (ii) the sample covariance of X_1 and X_2 for individuals missing X_2 is zero.

As our interest lies in estimating the effect of treatment, we will henceforth replace Jones's X_1 with the treatment allocation variable Z and X_2 with our notation for a partially observed covariate: X. So, Jones's work suggests that the least square estimator for the treatment effect is unbiased when the sample covariance of Z and X for individuals missing X is zero. We prove below that if the CIT assumption holds, this sample covariance is indeed zero, and thus the least square estimator for the treatment effect is unbiased. Furthermore, if the true outcome model resembles a parametric model corresponding to the MIA, then the CIO assumption holds, and it is simple to show that the least squares estimator is unbiased.

Suppose that the true data generation model for outcome Y is:

$$Y_i = \beta_0 + \beta_1 Z_i + \beta_2 X_i + \epsilon_i,$$

for $i = 1, ..., n$ patients, where Z denotes treatment allocation, X is a single partially observed confounder, and ϵ is an independent normal error term. Suppose further that the true data generation model for Z is:

$$Z_i = \gamma_0 + \gamma_1 X_i R_i + \gamma_2 (1 - R_i), \ i = 1, ..., n \tag{7.1}$$

where R denotes the missing indicator. Under the model in equation (7.1), the CIT assumption holds ($Z \perp X | R = 0$).

The sample covariance of Z and X for patients missing X is defined as:

$$S = [n(1 - \bar{R})]^{-1} \sum (1 - R_i)(Z_i - \bar{Z}^m)(X_i - \bar{X}^m) \tag{7.2}$$

where $\bar{R} = n^{-1} \sum R_i$ and $\bar{V}^m = [n(1 - \bar{R})]^{-1} \sum (1 - R_i) V_i$ for $V = Z, X$.

Under the model in equation (7.1),

$$\bar{Z}^m = [n(1 - \bar{R})]^{-1} \sum (1 - R_i)\big(\gamma_0 + \gamma_1 X_i R_i + \gamma_2 (1 - R_i)\big).$$

Since the missing indicator is binary, we can rewrite this as:

$$\bar{Z}^m = [n(1 - \bar{R})]^{-1} \sum \big(\gamma_0 (1 - R_i) + \gamma_2 (1 - R_i)\big).$$

By cancelling, we get $\bar{Z}^m = \gamma_0 + \gamma_2$. Substituting this expression and equation (7.1) into equation (7.2):

$$\begin{aligned} S &= [n(1 - \bar{R})]^{-1} \sum (1 - R_i)\big(\gamma_0 + \gamma_1 X_i R_i + \gamma_2 (1 - R_i) - \gamma_0 - \gamma_2\big)(X_i - \bar{X}^m). \\ &= [n(1 - \bar{R})]^{-1} \sum \big(\gamma_1 X_i R_i (1 - R_i) - \gamma_2 R_i (1 - R_i)\big)(X_i - \bar{X}^m) \\ &= 0. \end{aligned}$$

Hence, if the CIT assumption holds, the sample covariance of Z and X for patients missing X is zero. Thus, the least square estimator for the treatment effect is unbiased [59].

Chapter 8

Research paper: Estimating treatment effects with partially observed covariates using outcome regression with missing indicators

Helen A. Blake[1,2], Clémence Leyrat[1,3], Kathryn E. Mansfield[3], Laurie A. Tomlinson[3], James Carpenter[1,4], and Elizabeth J. Williamson[1,5]

1. Department of Medical Statistics, London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT, UK
2. Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine, 15-17 Tavistock Place, London, WC1H 9SH, UK
3. Department of Non-communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT, UK
4. MRC Clinical Trials Unit at UCL, Institute of Clinical Trials and Methodology, 90 High Holborn, London, WC1V 6LJ, UK
5. Health Data Research UK, 215 Euston Road, London, NW1 2BE, UK

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT
www.lshtm.ac.uk

Registry
T: +44(0)20 7299 4646
F: +44(0)20 7299 4656
E: registry@lshtm.ac.uk

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

RESEARCH PAPER COVER SHEET

PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.

SECTION A – Student Details

Student	Helen A Blake
Principal Supervisor	Elizabeth Williamson
Thesis Title	Dealing with partially observed covariates in propensity score analysis of observational data

If the Research Paper has previously been published please complete Section B, if not please move to Section C

SECTION B – Paper already published

Where was the work published?			
When was the work published?			
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Choose an item.	Was the work subject to academic peer review?	Choose an item.

**If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.*

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	Biometrical Journal
Please list the paper's authors in the intended authorship order:	Helen Blake, Clemence Leyrat, Kathryn Mansfield, Laurie Tomlinson, James Carpenter, Elizabeth Williamson.
Stage of publication	**Undergoing revision**

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I undertook the novel research being presented in the manuscript, including the mathematical theory and performing the simulations. I wrote the first draft of the manuscript and critically revised it following comments from my supervisors and other co-authors.

Student Signature: [redacted] Date: 24/9/19

Supervisor Signature: [redacted] Date: 24/9/19

8.1 Overview of the research paper pre-print: Estimating treatment effects with partially observed covariates using outcome regression with missing indicators

In the following research paper pre-print, we extend the MIA from propensity score analysis to the context of outcome regression. After introducing the MIA and the method's underlying assumptions, we prove that the MIA gives unbiased treatment effect estimates when the mSITA assumption holds, either the CIT or the CIO assumption holds, and the outcome model is correctly specified. We show how this finding is compatible with previous work by Jones (1996) which found that the MIA generally gives biased estimates in outcome regression [59], and highlight additional interesting results. In addition, we use simulation studies to explore the extent of bias when the MIA's assumptions are violated. We then illustrate the MIA in outcome regression with a cohort study using electronic health records.

8.2 Abstract

Missing data is a common issue in research using observational studies to investigate the effect of treatments on health outcomes. When missingness occurs only in the covariates, a simple approach is to use missing indicators to handle the partially observed covariates. The missing indicator approach has been criticised for giving biased results in outcome regression. However, recent papers have suggested that the missing indicator approach can provide unbiased results in propensity score analysis under certain assumptions. We consider assumptions under which the missing indicator approach can provide valid inferences, namely: (1) no unmeasured confounding within missingness patterns; either (2a) covariate values of patients with missing data were conditionally independent of treatment; or (2b) these values were conditionally independent of outcome; and (3) the outcome model is correctly spec-

ified: specifically, the true outcome model does not include interactions between missing indicators and fully observed covariates. We prove that, under the assumptions above, the missing indicator approach with outcome regression can provide unbiased estimates of the average treatment effect. We use a simulation study to investigate the extent of bias in estimates of the treatment effect when the assumptions are violated and we illustrate our findings using data from electronic health records. In conclusion, the missing indicator approach can provide valid inferences for outcome regression, but the plausibility of its assumptions must first be considered carefully.

8.3 Introduction

Observational studies are a valuable source of information for research investigating the efficacy and safety of treatments in practice. We focus on scenarios where we want to estimate the effect of treatment on a health outcome. However, a common challenge when using observational data is how to deal with missing data. If not handled appropriately, missing data can lead to bias and a loss of efficiency [50]. When using observational data for research, missing data is often an issue in variables that may be considered as potential confounders, such as smoking status or ethnicity.

The simplest approach for handling partially observed covariates is complete record analysis (also called complete case analysis), where patients with missing data are excluded from analysis. Although complete record analysis can provide unbiased results [50], this approach will typically lead to a loss of efficiency due to the exclusion of information. Furthermore, if patients with complete records are not representative of the population of interest, results from a complete record analysis may not be generalizable to the population of interest [45, 84].

A popular alternative missing data method is multiple imputation, where missing values are imputed multiple times with plausible values in order to create multiple 'complete' imputed datasets. After analysing each dataset, the results are combined using Rubin's Rules to obtain an overall treatment effect estimate [20, 44]. Although multiple imputation is very powerful, it can be fairly complex and stan-

dard implementation requires the assumption that data are missing at random (i.e. the probability of being missing depends on observed data and, given these, does not depend on unobserved data) [44, 48]. The plausibility of the missing at random assumption should be considered when implementing multiple imputation [49]. In addition, imputing missing values in standard multiple imputation relies on parametric assumptions [45], the plausibility of which should also be considered [85].

Another simple way of dealing with partially observed covariates is to use missing indicators — variables which indicate whether the covariate is missing or observed. For a continuous covariate, missing observations are replaced with a fixed value, say 0, and a missing indicator is added to the analysis model, alongside the continuous variable. For a categorical covariate, the missing indicator approach is equivalent to adding a 'missing' category to the variable.

The use of missing indicators to handle missing covariates in outcome regression has been criticised in the literature for being "ad hoc" [56, 57], and for giving biased results [58, 59]. However, the missing indicator approach is often used to deal with missing covariates [52] and has been recommended as a missing data method for propensity score analysis [55]. Related methods, incorporating the last-observation-carried-forward approach, have been studied in the context of non-systematic monitoring of covariates in settings with time-varying treatments [74, 75]. Furthermore, our recent work in the propensity score context suggests that the missing indicator approach can provide unbiased estimates under certain assumptions [86]. In propensity score analysis, we want to model the relationship between the covariates and the treatment, whereas in outcome regression, we wish to model the relationship between the covariates and the outcome. So, we need to investigate whether the validity of our findings in the propensity score context also holds for outcome regression. Therefore, in this paper we consider whether our work can be extended to the context of outcome regression.

We begin in Section 8.4 by describing the basic principles of the missing indicator approach and the assumptions underlying its validity. In Section 8.5, we prove that the missing indicator approach can give unbiased estimates of the treatment effect in

outcome regression and show how our work fits in with the literature. In Section 8.6, we explore the extent of bias in the estimation of the treatment effect when these assumptions are violated. In Section 8.7, we apply the missing indicator approach in multivariable outcome regression to an illustrative example. We conclude with a discussion in Section 8.8.

8.4 Background

8.4.1 Notation and potential outcome framework

Let Z be a binary variable indicating treatment allocation (or exposure status, etc. depending on context) and let Y represent the observed outcome variable. In this paper, we will concentrate on missing data in covariates and assume that treatment Z and outcome Y are fully observed, as the missing indicator method does not accommodate missing data on the outcome or exposure.

To enable us to describe the assumptions underlying the missing indicator approach, we refer to the potential outcome framework, developed by Rubin (1974), for causal inference from observational data. We let $Y(z)$ represent the potential outcome that would be observed if Z was set equal to the value z ($z = 0, 1$).

We focus on a scenario with two confounders: a fully observed confounder C and a partially observed confounder X. The missing indicator R equals 1 if X is observed, and $R = 0$ if X is missing.

The confounder values can be partitioned as $\{X_{obs}, X_{mis}\}$, where X_{obs} is the set of X values that are observed and X_{mis} is the set of missing X values (i.e. X_{mis} contains the true unobserved X values). For each patient with $R = 1$, X_{obs} is equal to X and X_{mis} is empty. For each patient with $R = 0$, $X_{mis} = X$ and X_{obs} is empty. Additionally, we define $X^* = X$ when $R = 1$, and $X^* = 0$ when $R = 0$. Note that an alternative approach would be to define X_{obs} instead as RX (which is equivalent to X^*) and X_{mis} as $(1 - R)X$. However, for the purposes of this paper, we use the X_{obs} and X_{mis} notation, following the literature on which our theory builds [46,63].

Our estimand of interest is the average treatment effect (ATE): $\mathrm{E}[Y(1)] - \mathrm{E}[Y(0)]$.

To estimate the treatment effect, we make the following standard assumptions for causal inference with complete data: strongly ignorable treatment allocation (SITA), no interference, consistency, and positivity.

The SITA assumption — an important assumption in causal inference using observational data — is that there is no unmeasured confounding [33]. In a scenario with two confounders, C and X, the SITA assumption can be written as:

$$Z \perp Y(1), Y(0) | C, X. \tag{8.1}$$

Under the assumption of no interference, the treatment status of one patient does not affect the potential outcomes of another patient [31,34]. Assuming consistency, the observed outcome of a patient is equal to the potential outcome corresponding to the treatment they actually received, i.e. if $Z = z$ then $Y = Y(z)$ [31]. Finally, under positivity, all patients have a non-zero probability of being assigned to each value of treatment, given their characteristics [31,37]

8.4.2 The missing indicator approach

The missing indicator approach is a simple method of dealing with partially observed covariates. When using outcome regression, the missing indicator approach allows patients with missing data to be used for the estimation of the treatment effect on the outcome, given covariates.

For a continuous partially observed covariate, the missing indicator approach in outcome regression replaces missing covariate values with some fixed value: the same value (for example, 0) is used for all participants with that covariate missing. Both the modified covariate and the missing indicator R are then included in the analysis model. For a categorical partially observed covariate, the missing indicator approach is equivalent to adding a 'missing' category to the variable. The regression coefficient for treatment can then be used to obtain an estimate of the treatment effect using appropriate transformations (eg. the identity function for linear regression).

For example, using the missing indicator approach for linear regression, the analysis model is $E[Y] = \alpha_0 + \alpha_1 Z + \alpha_2 C + \alpha_3 X^* + \alpha_4 R$, where $X^* = X$ when $R = 1$,

and $X^* = 0$ when $R = 0$, and where α_1 is the regresssion coefficient corresponding to our estimate of the ATE.

We note that, in the propensity score context, the missing indicator approach allows patients with missing data to contribute to the estimation of the propensity score (i.e. the probability of receiving treatment, given patient characteristics). So, missing indicators are included in the propensity score model, rather than the outcome model (which only includes treatment allocation and the propensity score as covariates).

8.4.2.1 Assumptions underlying the missing indicator approach

Our recent work in the context of propensity score analysis has shown that the missing indicator approach relies on four assumptions [86]. In this paper, we extend this work by investigating whether these four assumptions also underlie the validity of the missing indicator approach in outcome regression, in order to understand when this approach is appropriate in practice.

The first assumption is that there is no unmeasured confounding within missingness patterns, i.e. within each subgroup of patients who have information recorded on the same variables [63]. We call this the missingness Strongly Ignorable Treatment Allocation (mSITA) assumption, due to the similarity to the SITA assumption (equation (8.1)). Mathematically:

$$\text{mSITA:} \quad Z \perp Y(z) | C, X, R \quad \text{for } z = 0, 1. \tag{8.2}$$

We call the second and third assumptions the Conditionally Independent Treatment (CIT) assumption and the Conditionally Independent Outcomes (CIO) assumption, respectively. The CIT assumption is that missing confounder values are conditionally independent of treatment, given the observed confounder values and the missing indicator, while the CIO assumption is that missing confounder values

are conditionally independent of the potential outcomes [63].

$$\text{CIT:} \qquad Z \perp X_{mis} | C, X_{obs}, R. \tag{8.3a}$$

$$\text{CIO:} \quad Y(z) \perp X_{mis} | C, X_{obs}, R \quad \text{for } z = 0, 1. \tag{8.3b}$$

Note that in scenarios with partially observed confounders, the mSITA, CIT and CIO assumptions replace the SITA assumption with respect to identification of the causal estimand.

The fourth assumption in the propensity score context is that the propensity score model is correctly specified; in particular, we assume that the true propensity score model does not include an interaction between the missing indicator R and the fully observed confounder C (CR interaction). In other words, the effect of the fully observed confounder on treatment allocation is assumed to be the same for all missingness patterns.

The analogue assumption for outcome regression is that the outcome model is correctly specified and, in particular, the true outcome model does not include a CR interaction. The plausibility of this correct specification assumption is context-dependent and can be assessed in the data at hand, allowing the possibility of adapting the model in order to ensure the outcome model is correctly specified.

We can obtain valid inferences from the missing indicator approach in propensity score analysis under the following sufficient assumptions: (i) the mSITA assumption holds; (ii) either the CIT or the CIO assumption holds; and (iii) the propensity score model is correctly specified [86]. In this paper, we extend this work to the outcome regression context, demonstrating in Section 8.5 that we can use the missing indicator approach with outcome regression to obtain valid inferences under the following assumptions: (i) the mSITA assumption holds; (ii) either the CIT or the CIO assumption holds; and (iii) the outcome model is correctly specified.

8.4.2.2 Plausibility of the assumptions underlying the missing indicator approach

In missing data methodology, when deciding if a particular method is appropriate, it is important to consider the way in which data becomes missing, i.e. the missingness mechanism. Rubin's taxonomy [20] is commonly used to classify data as being missing completely at random, missing at random, or missing not at random [44,45].

The plausibility of the assumptions in Section 8.4.2.1 rely instead on the underlying structure of the data (i.e. the causal associations between variables), rather than the missingness mechanisms [86]. For example, the CIT and CIO assumptions together mean that the partially observed confounder does not confound the relationship between treatment and outcome when it is missing [86]. So, either the confounder-treatment relationship is absent in individuals who have missing confounder values or the confounder-outcome relationship is absent in individuals who have missing confounder values. Hence, key violations of the CIT or CIO assumptions occur when the missing confounder values affect treatment allocation or the outcome, respectively.

If we believe that the SITA assumption (i.e. no unmeasured confounding) holds in full data, then the mSITA assumption says that additionally conditioning on missingness patterns does not introduce bias. One key way in which this can be violated is when there are: shared unmeasured common causes between outcome and missingness, and unmeasured common causes between treatment and missingness. This is an example of M-bias, which has been discussed extensively in the literature [4,81].

The correct specification assumption would be violated if the effects of fully observed confounders on the outcome varied by missingness pattern. Unlike this parametric assumption, which can be tested in the data, the mSITA, CIT and CIO assumptions are not testable. Instead, researchers should use substantive knowledge of the given clinical setting to determine the plausibility of the mSITA, CIT and CIO assumptions.

The first step to assess the plausibility of these assumptions would be to consider

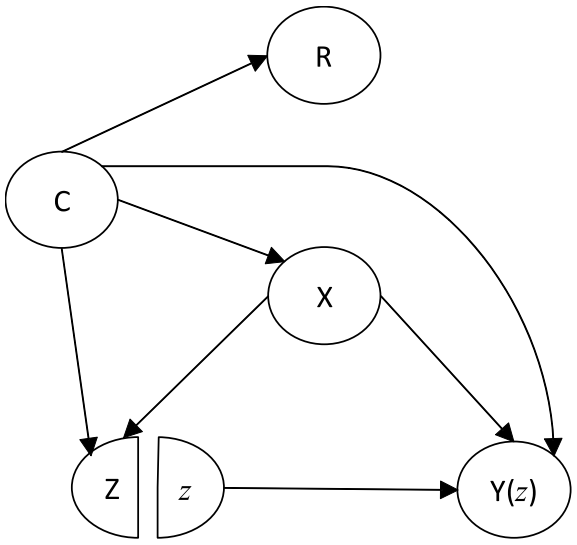


Figure 8.1: A causal diagram for a simple scenario with a partially observed confounder X and a fully observed confounder C, incorporating the missing indicator R. $Y(z)$ is the potential outcome resulting from intervening to set treatment Z to a particular value z.

whether it is clinically plausible that X is only a confounder when it is observed. If so, and if key violations of the assumptions can be ruled out, then researchers can construct a causal diagram to represent the underlying structural assumptions for the given clinical setting [86]. This causal diagram should include the missing indicator R. The next step is to convert this causal diagram to incorporate potential outcomes [67–69]. Then, the d-separation rule – which determines whether variables are conditionally independent given a set of other variables [66,67] – can be applied to the causal diagram to assess whether the mSITA assumption holds. In order to assess the CIT and CIO assumptions, the causal diagram should be restricted to patients with $R = 0$ and modified to reflect why it is plausible that X is only a confounder when it is observed [86]. The d-separation rule can then be applied to this final causal diagram to assess the CIT and CIO assumptions.

For example, consider a simple scenario with a partially observed confounder X and a fully observed confounder C, where C also has causal effects on both X and R. Further suppose that the X-Z relationship is absent in patients with missing X values. Hence, it is plausible that X is only a confounder when it is observed. Figure 8.1 shows a causal diagram representing this scenario, constructed in the form of a single world intervention graph in order to incorporate potential outcomes [67]. Applying the d-separation rule to Figure 8.1, as previously described [67, 86], we

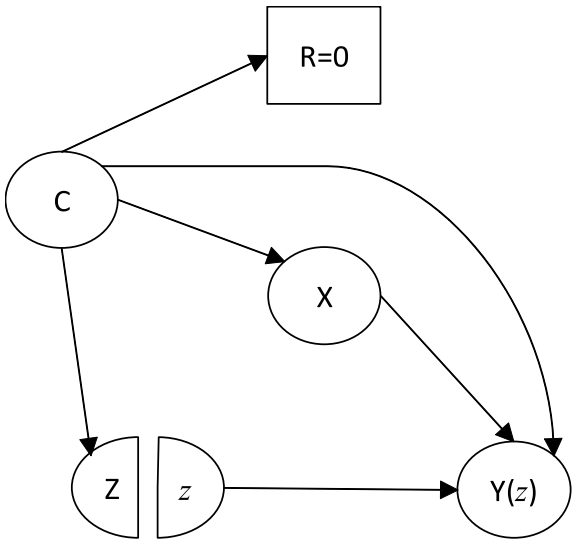


Figure 8.2: A causal diagram for a simple scenario with a partially observed confounder X and a fully observed confounder C, modified to assess the CIT and CIO assumptions. The square box around $R = 0$ indicates restriction to individuals with missing X values. $Y(z)$ is the potential outcome resulting from intervening to set treatment Z to a particular value z.

find that Z is conditionally independent of $Y(z)$ given C, X and R. Hence, the mSITA assumption holds in this example. In order to be able to assess the CIT and CIO assumptions, we modify Figure 8.1, by restricting to patients with $R = 0$ and removing the arrow from X to Z in order to encode the assumption that the X-Z relationship is absent in patients with $R = 0$. Figure 8.2 shows this modified causal diagram. Applying the d-separation rule to this diagram, we find that the CIT assumption holds and that the CIO assumption is violated. Hence, in this scenario, the mSITA and CIT assumptions hold and the missing indicator approach is considered appropriate.

When there are multiple partially observed confounders, R becomes a vector of the missing indicators, whilst X_{obs} now represents all of the sets of observed confounder values and X_{mis} represents all sets of missing confounder values. Assuming that the missingness of these confounders are not associated with each other or with the other confounders, we can assess the CIT and CIO assumptions for each confounder separately, but whilst conditioning on all sets of observed confounder values and all fully observed confounders. An assumption only holds if it holds for every confounder. Issues may arise if the missing indicator of one confounder changes the missing values of another confounder; however, this seems unlikely. For com-

plex scenarios, we recommend constructing a causal diagram that incorporates all relevant substantive knowledge and a missing indicator for each partially observed confounder, and then using software such as Dagitty [77] to assess the plausibility of the assumptions.

8.5 Unbiased estimation of the average treatment effect

In this section we prove that, under the four assumptions given in Section 8.4.2.1, the missing indicator approach in outcome regression can give an unbiased estimate of the average treatment effect (ATE). We also explore how this result relates to the findings in the literature that the missing indicator approach gives biased results [59], and how the assumptions relate to prior literature.

The target estimand is: ATE $= \mathrm{E}[Y(1)] - \mathrm{E}[Y(0)]$. We can rewrite this as:

$$\text{ATE} = \mathrm{E}\big[\mathrm{E}\big(Y(1)|C, X_{obs}, R\big) - \mathrm{E}\big(Y(0)|C, X_{obs}, R\big)\big],$$

which can then be written as:

$$\text{ATE} = \sum \big[\sum y\mathrm{P}\big(Y(1) = y|C, X_{obs}, R\big) - \sum y\mathrm{P}\big(Y(0) = y|C, X_{obs}, R\big)\big]. \quad (8.4)$$

Below in Section 8.5.1, we show that if the mSITA assumption holds and either the CIT assumption or the CIO assumption holds, then:

$$\mathrm{E}[Y(z)|C, X_{obs}, R] = \mathrm{E}[Y(z)|Z, C, X_{obs}, R] \quad (\text{for } z = 0, 1). \quad (8.5)$$

Hence, we can rewrite equation (8.4) as:

$$\begin{aligned}\text{ATE} &= \sum \big[\sum y\mathrm{P}\big(Y(1) = y|Z, C, X_{obs}, R\big) - \sum y\mathrm{P}\big(Y(0) = y|Z, C, X_{obs}, R\big)\big] \\ &= \mathrm{E}\big[\mathrm{E}\big(Y(1)|Z, C, X_{obs}, R\big) - \mathrm{E}\big(Y(0)|Z, C, X_{obs}, R\big)\big].\end{aligned}$$

Under the consistency assumption (Section 8.4.1), this is:

$$\text{ATE} = \mathrm{E}\big[\mathrm{E}\big(Y|Z=1, C, X_{obs}, R\big) - \mathrm{E}\big(Y|Z=0, C, X_{obs}, R\big)\big]. \tag{8.6}$$

So, we can model the relationship between the outcome and C, X_{obs}, R in each of the two treatment groups and — assuming that the outcome model is correctly specified — we can substitute estimates of the conditional expectations in equation (8.6) to obtain an unbiased estimate of the ATE. Thus, under the assumptions given in Section 8.4.2.1, we can get an unbiased estimate of the treatment effect by modelling the relationship between outcome and treatment, given confounders and the missing indicator.

The missing indicator approach suggests a particular parametric specification of the outcome model, at this stage. In particular, missing indicators are added as main effects only, thereby encoding the assumption that there are no interactions between the missing indicators and fully observed confounders. These parametric modelling assumptions can be assessed using the data at hand, although it is unclear whether such checks are common in practice.

8.5.1 Proof of equation (8.5)

We first suppose the mSITA and CIT assumptions hold (equations (8.2) and (8.3a), respectively). For $z = 0, 1$, we can write $\mathrm{E}[Y(z)|C, X_{obs}, R]$ (from equation (8.5)) in summation notation:

$$\begin{aligned}
&\sum y \mathrm{P}\big(Y(z) = y|C, X_{obs}, R\big) \\
&= \sum\sum y \mathrm{P}\big(Y(z) = y, X_{mis}|C, X_{obs}, R\big) \\
&= \sum\sum y \mathrm{P}\big(Y(z) = y|X_{mis}, C, X_{obs}, R\big)\mathrm{P}(X_{mis}|C, X_{obs}, R).
\end{aligned} \tag{8.7}$$

Under the mSITA assumption, the first probability in equation (8.7) can be written as $\mathrm{P}\big(Y(z) = y|Z, X_{mis}, C, X_{obs}, R\big)$, and under the CIT assumption, the second probability can be written as $\mathrm{P}(X_{mis}|Z, C, X_{obs}, R)$. So, for $z = 0, 1$, equation

(8.7) becomes:

$$\begin{aligned}
&\sum\sum y\mathrm{P}\big(Y(z)=y|Z,X_{mis},C,X_{obs},R\big)\mathrm{P}(X_{mis}|Z,C,X_{obs},R)\\
&\quad=\sum\sum y\mathrm{P}\big(Y(z)=y,X_{mis}|Z,C,X_{obs},R\big)\\
&\quad=\sum y\mathrm{P}\big(Y(z)=y|Z,C,X_{obs},R\big).\quad\square
\end{aligned}$$

Alternatively, if the mSITA and CIO assumptions (equations (8.2) and (8.3b)) hold, for $z=0,1$, we write:

$$\sum y\mathrm{P}\big(Y(z)=y|C,X_{obs},R\big)=\sum y\frac{\mathrm{P}\big(Y(z)=y|C,X_{obs},R\big)}{\mathrm{P}\big(Z|C,X_{obs},R\big)}\sum\mathrm{P}\big(Z,X_{mis}|C,X_{obs},R\big),\tag{8.8}$$

where the denominator is strictly positive under the positivity assumption.

Now, we can write:

$$\sum\mathrm{P}\big(Z,X_{mis}|C,X_{obs},R\big)=\sum\mathrm{P}\big(Z|X_{mis},C,X_{obs},R\big)\mathrm{P}\big(X_{mis}|C,X_{obs},R\big).\tag{8.9}$$

Under the mSITA assumption, the first probability in equation (8.9) can be written as $\mathrm{P}\big(Z|Y(z)=y,X_{mis},C,X_{obs},R\big)$, and under the CIO assumption, the second probability can be written as $\mathrm{P}(X_{mis}|Y(z)=y,C,X_{obs},R)$. So, for $z=0,1$, equation (8.9) becomes:

$$\begin{aligned}
\sum\mathrm{P}\big(Z,X_{mis}|C,X_{obs},R\big)&=\sum\mathrm{P}\big(Z|Y(z)=y,X_{mis},C,X_{obs},R\big)\\
&\quad\times\mathrm{P}\big(X_{mis}|Y(z)=y,C,X_{obs},R\big)\\
&=\sum\mathrm{P}\big(Z,X_{mis}|Y(z)=y,C,X_{obs},R\big)\\
&=\mathrm{P}\big(Z|Y(z)=y,C,X_{obs},R\big).
\end{aligned}$$

Hence, we can write equation (8.8) as:

$$\begin{aligned}
&\sum y \frac{\mathrm{P}\big(Y(z)=y|C,X_{obs},R\big)}{\mathrm{P}\big(Z|C,X_{obs},R\big)} \mathrm{P}\big(Z|Y(z)=y,C,X_{obs},R\big) \\
&= \sum y \frac{\mathrm{P}\big(Y(z)=y,Z|C,X_{obs},R\big)}{\mathrm{P}\big(Z|C,X_{obs},R\big)} \\
&= \sum y \mathrm{P}\big(Y(z)=y|Z,C,X_{obs},R\big). \quad \square
\end{aligned}$$

8.5.2 Connections to prior work on the missing indicator approach

Jones (1996) assumed that the true outcome model is a linear regression model with a fully observed covariate Z, a single partially observed covariate X and independent normal errors ϵ:

$$Y = \beta_0 + \beta_1 Z + \beta_2 X + \epsilon, \tag{8.10}$$

where ϵ is independent of (Z, X, R). Correspondingly, the missing indicator approach can be represented mathematically as:

$$E[Y] = \gamma_0 + \gamma_1 Z + \gamma_2 X^* + \gamma_3 R. \tag{8.11}$$

Jones (1996) showed that the least squares estimator of γ_1 is biased for β_1, noting that the least squares estimator is unbiased when the sample covariance of Z and X, for patients missing X, is zero. If the CIT assumption holds, this condition holds, since treatment allocation is independent of the confounder for those patients with missing confounder values.

The true outcome model assumed by Jones (1996) in equation (8.10) leads to the CIO assumption being violated as the outcome is dependent on the missing confounder values. However, if the CIO assumption does hold, then the true outcome model instead resembles the parametric model corresponding to the missing indicator approach in equation (8.11) (i.e. the true model is $Y = \beta_0+\beta_1 Z+\beta_2 X^*+\beta_3 R+\epsilon$), and it is simple to show that the least squares estimator is unbiased.

Hence, our findings are compatible with Jones's findings (1996). We have addi-

tionally shown that the missing indicator approach can give unbiased estimates when the mSITA and CIO assumptions hold (regardless of whether the CIT assumption additionally holds).

8.5.3 Connection to alternative statements of assumptions in the literature

The missing indicator method has been recommended for propensity score analysis [55], based on work in relation to the missingness pattern approach within propensity score analysis [46, 63]. This approach involves modelling the propensity score separately for each pattern of missing confounder data and can be thought of as a generalisation of the missing indicator method.

In Section 8.4.2.1, our statement of the mSITA, CIT and CIO assumptions follows Mattei (2009), who states assumptions sufficient for valid inference for the missingness pattern approach. Our assumptions differ from Mattei (2009) in that our version of the CIO assumption is slightly weaker, and requires the conditional independence statement to hold separately for each potential outcome, rather than jointly for the pair of potential outcomes as in the original presentation.

D'Agostino and Rubin (2000) instead provide the following assumption, sufficient for valid inference in the missingness pattern approach:

$$Z \perp (Y(0), Y(1), X_{mis}) | X_{obs}, R. \tag{8.12}$$

The mSITA and CIT assumptions imply that equation (8.12) holds. However, mSITA and CIO can hold while equation (8.12) is violated. Thus Mattei (2009) gives a wider set of assumptions under which the missingness pattern approach provides valid inference.

There are strong connections between the missingness pattern approach and other work exploring non-systematic monitoring of time-varying covariates [74, 75]. These papers suggest a version of the "no unmeasured confounding assumption"

which, in the single time-point exposure setting, can be written as:

$$Z \perp Y(z)|X_{obs}, R. \tag{8.13}$$

If the D'Agostino assumption holds, then assumption (8.13) holds. Further, if either the mSITA and CIT assumptions hold, or the mSITA and CIO assumptions hold, then assumption (8.13) holds. Compared to the D'Agostino assumption (8.12), therefore, the mSITA, CIT and CIO assumptions can be seen as a wider set of assumptions under which variants of missingness-pattern-type approaches can produce valid inference.

Kreif et al. (2018) focus on the scenario where the partially missing (non-systematically monitored) covariate is key to the treatment decision process and thus when the clinician does not have this covariate information, they must rely on the last measurement available. Therefore, in their setting – in contrast to our scenario – the covariate always contributes to the treatment decision, whether as an up-to-date measurement or as the last available measurement. However, both settings lead to a causal structure which satisfies a CIT-type assumption.

Here, we assume that full-data inference is the goal, i.e. if we were able to obtain full data then we would. Kreif et al. (2018), in contrast, treat the monitoring process (which induces the missing covariate data) as an intrinsic part of the setting, and as an attribute of interest in its own right. In particular, in time-varying settings the optimal treatment combination may depend on the intended monitoring process. This makes the inferential goals of Kreif et al. (2018) quite different to those laid out in the current paper. In particular, the set of assumptions we focus on (mSITA, CIT and CIO) require the investigator to consider the confounding structure in the full data setting and how missingness arises in that setting (mSITA), and then to subsequently explore how this structure may change when missing confounder values are present (CIT/CIO). We have found this two-step process useful in considering plausibility of assumptions in real-life settings.

All three sets of assumptions make it clear that they are likely to be satisfied in a setting where missing confounder values are unavailable to the individual making

the treatment decision and thus do not affect treatment. However, only the first version, with CIT and CIO as two separate sub-assumptions, makes it clear that there is another quite different set of scenarios in which missingness-pattern-type methods may provide valid inference.

8.6 Simulation Study

In this simulation study, we explored the extent of the bias introduced into the treatment effect estimation when each of the key assumptions is violated. Source code to reproduce the results is available as Supporting Information on the journal's web page (http://onlinelibrary.wiley.com/doi/xxx/suppinfo).

8.6.1 Data-generating mechanisms

We considered 81 data-generating mechanisms. For each, datasets of sample size $n = 500$ were generated. The data-generating mechanisms differed according to which of the assumptions hold. A factorial design was used to consider all possible combinations of each assumption having no violation, a weak violation or a strong violation.

We let U_Z represent a common cause between treatment and the missing indicator, U_Y represent a common cause between the outcome and the missing indicator, and e represent error in the outcome regression model. We generated U_Z, U_Y and e from independent standard Normal distributions.

Two binary confounders X and C were generated using Binomial distributions: $X \sim \text{Bin}(1, 0.67)$ and $C \sim \text{Bin}(1, 0.58)$. To create missing data in X, we generated a missing indicator $R \sim \text{Bin}(1, \text{P}(R = 0))$, where: $\text{logit}(\text{P}(R = 0)) = -0.5 + 1.48 \cdot U_Z + 1.36 \cdot U_Y$.

We also generated a binary treatment allocation variable $Z \sim \text{Bin}(1, \text{P}(Z = 1))$, where:

$$\text{logit}(\text{P}(Z = 1)) = -1.2 + \alpha U_Z + 1.38XR + \beta X(1 - R) + 2R + 1.69C.$$

The observed proportion of treated patients varied between 62.2% and 86.2%, depending on the data-generating mechanism. We generated a continuous outcome using the regression model:

$$Y = 1 - 2.35Z - 2.2\alpha U_Y - 1.55XR + \gamma X(1-R) + 1.8R - 1.7C + \delta CR + 3e.$$

where $\alpha \in \{0, 0.125, 1.25\}$, $\beta \in \{0, 0.138, 1.38\}$, $\gamma \in \{0, -0.155, -1.55\}$ and $\delta \in \{0, -0.42, -4.2\}$. If $\alpha = 0$, then the mSITA assumption holds. Similarly, if $\beta = 0$, $\gamma = 0$, or $\delta = 0$, then, respectively, the CIT assumption holds, the CIO assumption holds, or the outcome model is correctly specified. For each parameter, the smaller and larger non-zero values represent, respectively, a weak violation and a strong violation of the corresponding assumption.

Data were simulated using Stata 14.2 with 5000 simulation repetitions per data-generating mechanism.

8.6.2 Methods

Each simulated data set was analysed using the missing indicator approach with multivariable linear regression, by creating a new version of the partially observed binary covariate with a third 'missing' category. Our estimand is the average treatment effect, as estimated using the treatment coefficient from a linear regression model. Our performance measure of interest is absolute bias of the ATE: $\frac{1}{5000}\sum_{i=1}^{5000}\hat{\theta}_i - \theta$, where $\hat{\theta}_i$ is the estimated treatment effect from the ith repetition, and θ is the true treatment effect.

8.6.3 Results

In Figure 8.3, the left-hand panel presents the absolute bias in the estimated treatment effect for eight scenarios, depicting all possible combinations of the mSITA, CIT and CIO assumptions holding or not. The dark bars show scenarios where the mSITA assumption (required for valid inference) holds. The light bars show scenarios where it does not. As expected from our theory above, if the mSITA assumption

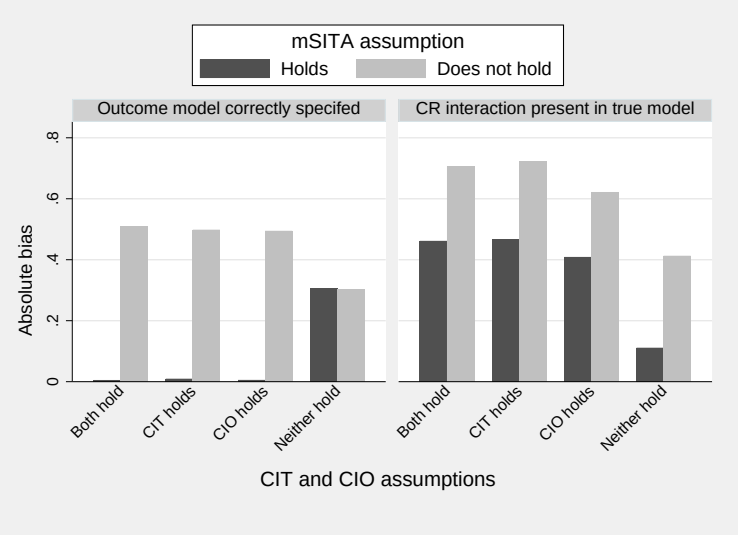


Figure 8.3: Results from a simulation study showing the absolute bias in the estimated treatment effect when using the missing indicator approach for multiple linear regression under different data-generating mechanisms, which vary according to: (i) whether the mSITA assumption holds; (ii) whether the CIT assumption holds; (iii) whether the CIO assumption holds; and (iv) whether there is an interaction between the fully observed confounder C and the missing indicator R in the true outcome model. True treatment effect: -2.35. Sample size: $n = 500$. Number of replications: 5000.

is violated (light bars), bias is present. The four sets of bars show combinations of the CIT and CIO assumptions holding or not, for scenarios where the mSITA assumption holds (dark bars); bias is present only when both CIT and CIO are violated. The right-hand panel of Figure 8.3 shows the same eight scenarios, but with a violation of the parametric assumption: the outcome model fitted assumes no interaction between the missingness indicator and the fully observed confounder C, but in truth this interaction does exist. Violation of this parametric assumption leads to bias in all eight scenarios.

Figure 8.4 shows a number of scenarios in which the outcome model is correctly specified but the other three assumptions (mSITA, CIT and CIO) may be violated. The three panels show – from left to right – increasing levels of violation of the CIO assumption. Within the three panels, the three sets of bars show – from left to right – increasing levels of violation of the CIT assumption. Within each set of bars, the bars show – from left to right – increasing levels of violation of the mSITA assumption. Large bias is seen when either a strong violation of the mSITA assumption is present, or when strong violations of both the CIT and CIO assumptions are present.

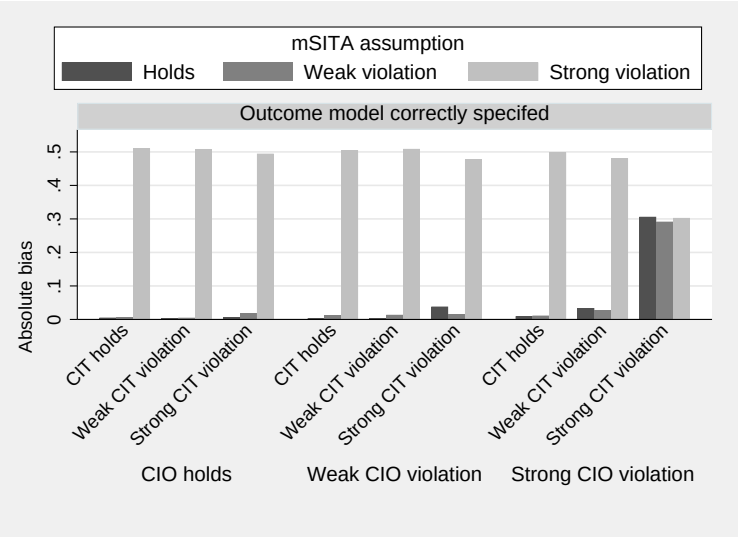


Figure 8.4: Results from a simulation study showing the absolute bias in the estimated treatment effect when using the missing indicator approach for multiple linear regression under different data-generating mechanisms, which vary according to whether there is no violation, a weak violation or a strong violation of: (i) the mSITA assumption, (ii) the CIT assumption, and (iii) the CIO assumption. For all data-generating mechanisms, the outcome model is correctly specified. True treatment effect: -2.35. Sample size: $n = 500$. Number of replications: 5000.

Figure 8.5 shows the same scenarios as Figure 8.4, but with weak violations of the parametric assumptions (weak CR interactions present but not included in the fitted model) shown in the top panel, and strong violations of the parametric assumptions shown in the bottom. Weak violations of the parametric assumptions induced additional small amounts of bias compared to Figure 8.4. Strong violations of the parametric assumptions induced large amounts of bias under most settings.

The missing indicator approach gives unbiased estimates of the treatment effect when the mSITA assumption holds, the outcome model is correctly specified and either one, or both, of the CIT and CIO assumptions hold. When both the CIT and CIO assumptions are violated, the missing indicator approach gives biased results, whether or not the other two assumptions hold. The worst bias occurs when both the mSITA assumption and the correct specification assumption is violated and the CIT assumption holds, whether or not the CIO assumption holds. In general, having the mSITA assumption violated results in larger biases for the settings explored in the simulation study. In addition, incorrectly specifying the outcome model, i.e. failing to include an interaction between the fully observed confounder C and the missing indicator R in the true outcome model, generally results in larger biases

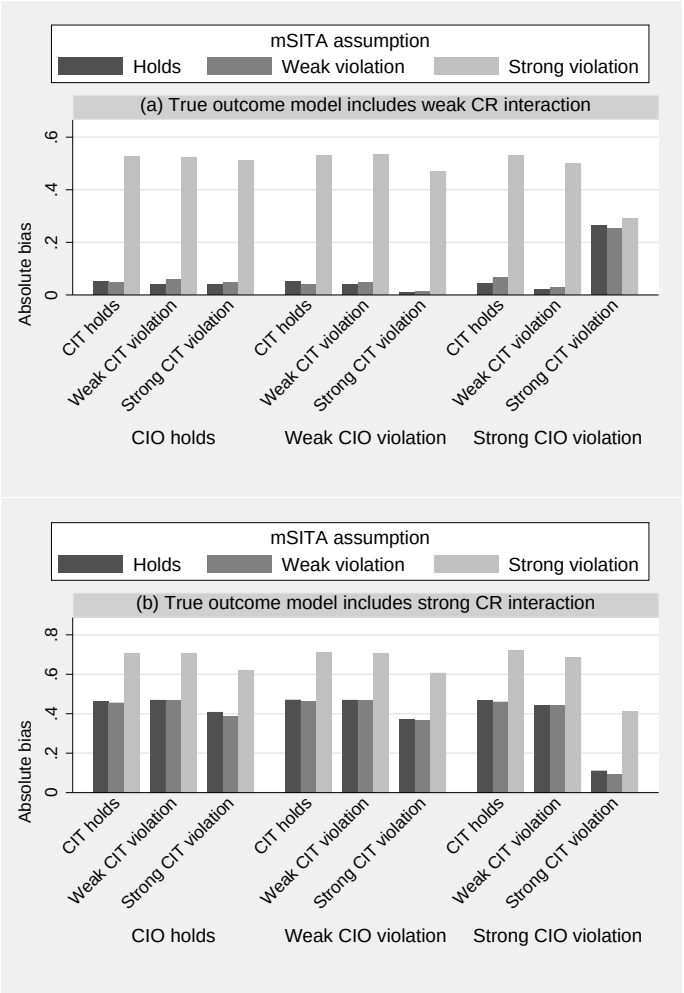


Figure 8.5: Results from a simulation study showing the absolute bias in the estimated treatment effect when using the missing indicator approach for multiple linear regression under different data-generating mechanisms, which vary according to whether there is no violation, a weak violation or a strong violation of: (i) the mSITA assumption, (ii) the CIT assumption, and (iii) the CIO assumption. For all data-generating mechanisms, the true outcome model contains either a weak interaction (8.5a) or a strong interaction (8.5b) between the fully observed confounder C and the missing indicator R. True treatment effect: -2.35. Sample size: $n = 500$. Number of replications: 5000.

than when the outcome model is correctly specified.

When the outcome model is correctly specified, weak violations of the other assumptions results in similar biases compared to when the assumptions hold (Figure 8.4). Similar results were found when considering data-generating mechanisms where the true outcome model includes a weak CR interaction and when considering data-generating mechanisms with a strong CR interaction (Figures 8.5a and 8.5b respectively). In general, having a weak CR interaction resulted in similar or larger biases compared to scenarios where the outcome model is correctly specified.

8.7 Application to illustrative example

8.7.1 Study description

Our illustrative example is a cohort study using electronic health records data from the UK Clinical Practice Research Datalink and the Hospital Episode Statistics [11]. The cohort study aimed to investigate the association between risk of acute kidney injury (AKI) and use of angiotensin-converting enzyme inhibitors or angiotensin receptor blockers (ACEI/ARBs), compared to other antihypertensive drugs. An important covariate in the study was chronic kidney disease, which was categorised into stages based on a continuous measure of kidney function called the estimated glomerular filtration rate (eGFR). Lower values of eGFR indicate worse kidney function.

Data were obtained for 570 586 new adult users of antihypertensive drugs between 1997 and 2014. Follow-up began at first prescription of ACEI/ARBs, beta blockers, calcium channel blockers, or diuretics. The treatment of interest was prescription of ACEI/ARBs. Our outcome of interest was kidney function within 2 months of first prescription of an antihypertensive drug, as measured using eGFR [87]. Due to conditions of the data use agreement, we can no longer access the eGFR data after treatment initiation, so we have simulated this variable, based on observed relationships in prior studies (see Appendix for details). As a result, the 'true' treatment effect is known.

In this study there were a number of fully observed potential confounders: age, sex, chronic comorbidities, other antihypertensive or diuretic drugs, and calendar period. In addition, two potential confounders were partially observed: ethnicity, which had 59.0% missing data; and baseline eGFR category, which had 52.9% missing data.

In this example, only 21% of patients had complete data for both ethnicity and baseline eGFR category; the majority of patients records would be discarded, leading to a loss of efficiency, if complete record analysis was used for this example. Furthermore, standard multiple imputation may not be appropriate since the miss-

ing at random assumption is questionable: baseline eGFR category is more likely to be measured for patients with worse kidney function. The assumptions underlying the missing indicator approach seem reasonable in this context. First, the mSITA assumption would be violated if there are any unobserved common causes between missingness of baseline eGFR category and treatment allocation or the outcome. In this example, it seems plausible that any such common causes, such as age or chronic comorbidities, are measured and able to be included in the analysis model. In addition, predictors of missingness in ethnicity seem unlikely to also be predictors of prescription decisions. Thus the mSITA assumption seems plausible here.

Second, it is plausible to assume that information about a patient's baseline eGFR category is unlikely to influence the clinician's decision to prescribe if this information is not available to the clinician (eg. if a kidney function test had not been ordered beforehand). In practice, proxy information about a patient's baseline eGFR category may be available to the clinician (but not to researchers using electronic health records). However, this is likely to reflect poor kidney function for only a small proportion of the whole study population. In addition, it is plausible that a clinician would ensure information on patient's ethnicity is recorded if they believe that this information is an important factor in their decision whether or not to prescribe ACEI/ARBs. Thus we believe that the CIT assumption is plausible.

Third, it seems plausible that the effect of the other fully observed risk factors on AKI would not vary according to whether or not ethnicity and baseline eGFR category were measured. Furthermore, this assumption can be tested in the data.

Fourth, the CIO assumption does not seem plausible in this context — since baseline kidney function remains a risk factor for change in eGFR, whether or not baseline eGFR category is measured. Since we can obtain valid inferences from the missing indicator approach when just one of the CIT and CIO assumptions hold (in addition to the mSITA and correct specification assumptions holding), the CIO assumption being violated is not an issue here; the mSITA, CIT and correct specification assumptions seem plausible and thus the missing indicator approach is considered appropriate.

8.7.2 Method

We applied linear regression, adjusted for ethnicity, baseline eGFR category and fully observed confounders (age category, sex, chronic comorbidities, and calendar period), to obtain estimates of the treatment effect comparing patients prescribed ACEI/ARBs at start of follow-up time exposed to ACEI/ARBs versus patients not prescribed ACEI/ARBs at baseline. To handle missing data in ethnicity and baseline eGFR category, we applied complete record analysis and the missing indicator approach. Analysis was conducted in Stata 14.2.

8.7.3 Results

Our results are given in Table 8.1. The missing indicator approach uses all missingness patterns; in addition to the 121 527 patients with complete data, 112 142 patients had missing data for baseline eGFR category, 147 011 have ethnicity missing, and 189 906 had missing data for both. Using the missing indicator approach, the estimated treatment effect was closer to the true treatment effect than the estimate from complete record analysis. In addition, the complete record analysis estimate has a wider confidence interval due to the exclusion of over 75% of the patient records. When interactions between the missing indicators and the fully observed confounders are added into the regression model, the results do not change much compared to the missing indicator approach (-0.6575, 95% CI: [-0.7509, -0.5640]), and so there is no evidence of a violation of the parametric assumption.

Table 8.1: Estimated treatment effects (mean differences) and 95% confidence intervals (CIs) using linear regression to compare the effect on (simulated) kidney function of being prescribed ACEI/ARBs at start of follow-up versus not being prescribed ACEI/ARBs at baseline. True treatment effect: -0.6831

Missing data method	Treatment effect (95% CI)	Number of patients analysed
Complete record analysis	-0.6150 $(-0.7977, -0.4324)$	121 527
Missing indicator approach	-0.6496 $(-0.7424, -0.5567)$	570 586

8.8 Discussion

In this paper, we have shown that the missing indicator approach in outcome regression is unbiased when (i) there is no unmeasured confounding within missingness patterns; (ii) either confounder values of patients with missing data are conditionally independent of treatment assignment, or these missing confounder values are conditionally independent of the outcome; and (iii) the effect of fully observed confounders on the outcome is the same for all missingness patterns. We have applied the missing indicator approach to an illustrative example using routinely collected data, a key area in which the method's underlying assumptions may be plausible [86].

An advantage of the missing indicator approach for outcome regression is that it is easy to implement and, unlike complete record analysis, avoids discarding much information when the proportion of missing data is large. In addition, the missing indicator approach may be appropriate in situations where multiple imputation is not, as the missing indicator approach does not rely on the conventional classification of missingness mechanisms. Whereas standard implementation of multiple imputation is guaranteed to be valid when data are missing at random, the CIT and CIO assumptions are not about the missingness mechanism, but are rather about whether the partially observed covariate confounds the relationship between treatment and outcome when it is missing. When either the CIT or the CIO assumption holds, the relationships between variables among patients with observed data are not the same as those among patients with missing data, and so multiple imputation may not be appropriate. In contrast, the missing indicator approach may be unbiased under missing not at random mechanisms, and biased under some missing completely at random mechanisms.

The missing indicator approach has been criticised in the missing data methodology literature as being 'ad hoc' [57] and biased [58, 59]. We have shown that the missing indicator approach can give unbiased results under certain assumptions. Researchers seeking to use the missing indicator approach should first consider whether these assumptions seem plausible within the context of a given clinical setting, with

the help of causal diagrams. In our simulation study, we considered scenarios with a single partially observed variable. Our suggested approach to handling multiple partially observed confounders within the missing indicator framework requires the assumption that the missingness of one confounder does not affect the missing values of another confounder. In practice, researchers should carefully consider the plausibility of such an assumption, in addition to considering the plausibility of the mSITA, CIT, CIO, and correct specification assumptions. If the assumptions underlying the missing indicator approach are found to not be appropriate, then researchers should consider whether the assumptions underlying complete record analysis or multiple imputation are more appropriate in the given scenario.

The missing indicator approach is a method for handling missing covariate data, but cannot handle missing data on the outcome or treatment allocation. Further work is required to extend the approach to handle other missing data, perhaps by combining with other methods such as multiple imputation. Another limitation of the missing indicator approach, in the context of propensity score analysis, is that estimation issues may arise if there are many missingness patterns and some of these patterns have low sample size. Qu and Lipkovich (2009) proposed a pattern-pooling algorithm to ensure sufficient sample size for estimation in propensity score analysis [70]. Further work is needed to explore the impact of low sample size in missingness patterns in the context of outcome regression and whether this impact can be alleviated by using pattern-pooling algorithms. A limitation of our simulation study is that we did not assess the impact of changing the proportion of missing data. However, when the assumptions do not hold, bias is expected to increase with the proportion of missing data. Furthermore, in this paper, we have focused on linear regression. We believe that our theoretical results can be extended to risk difference estimation and Poisson regression; further work is required to confirm this. Careful consideration would be required to translate these results to the odds ratio setting due to non-collapsibility issues.

In conclusion, the missing indicator approach for outcome regression can be applied in a principled way and can give valid results under a particular set of assump-

tions, but researchers must first consider whether these assumptions seem plausible in the clinical setting of interest. We end by noting that standard application of the missing indicator approach makes rather strong parametric assumptions about absence of interactions between missing indicators and fully observed confounders; we recommend that checking these assumptions in the data at hand should form part of routine practice when applying this approach.

Acknowledgements

HAB was supported by the Economic and Social Research Council [Grant Number ES/J5000/21/1]. CL was supported by the Medical Research Council [Project Grant MR/M013278/1]. LAT was supported by a Wellcome Trust intermediate clinical fellowship [Grant Number 101143/Z/13/Z]. EJW was supported by Health Data Research UK [Grant Number EPNCZO90], which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and Wellcome. Ethics approval was given by the London School of Hygiene and Tropical Medicine Research Ethics Committee [Reference: 15880] and by the Clinical Practice Research Datalink Independent Scientific Advisory Committee [ISAC Protocol Number 14_208A2].

Conflict of interest statement

The authors have declared no conflict of interest.

Appendix

A. Simulating kidney function for the illustrative example

We simulated a continuous outcome $Y = \mathbf{X}\beta + e$ where $\mathbf{X}$ denotes the design matrix, β represents the vector of regression coefficients and e denotes the vector of error terms, where $e \sim (0, 14.65)$. The design matrix contains the vector with all entries equal to 1 and the following variables: prescription of ACEI/ARBs at baseline; diabetes mellitus status at baseline; hypertension status at baseline; cardiac failure status at baseline; arrhythmia status at baseline; ischaemic heart disease status at baseline; sex; ageband at baseline; calendar period at baseline; ethnicity; and baseline eGFR category. The regression coefficients are given in Table 8.2.

Table 8.2: Regression coefficients for using baseline characteristics to simulate an outcome variable measuring kidney function within two months of prescription of antihypertensive drugs. ACEI/ARBs: angiotensin-converting enzyme inhibitors or angiotensin receptor blockers. eGFR: estimated glomerular filtration rate.

Coefficient	Variable	Coefficient	Variable
-0.6831	ACEI/ARBs prescription	1.3974	calendar period $2001-2004$
0.4847	diabetes mellitus	2.7825	calendar period $2005-2008$
-5.5041	hypertension	4.2181	calendar period $2009-2011$
-1.9321	cardiac failure	4.9409	calendar period $2012-2014$
-1.6349	arrhythmia	4.1883	ethnicity recorded as south asian
-3.4547	ischaemic heart disease	-2.6490	ethnicity recorded as black
-1.6717	female	2.7238	ethnicity recorded as other
-12.8473	$45 \leq$ age < 55	3.3971	ethnicity recorded as mixed
-17.6097	$55 \leq$ age < 60	0.1647	ethnicity missing
-20.0686	$60 \leq$ age < 65	-36.7126	baseline eGFR < 30
-22.1784	$65 \leq$ age < 70	-25.1941	$30 \leq$ baseline eGFR < 45
-24.1881	$70 \leq$ age < 75	-16.3931	$45 \leq$ baseline eGFR < 60
-26.5288	$75 \leq$ age < 85	-4.4043	baseline eGFR missing
-25.6283	age ≥ 85	94.0335	constant

Chapter 9

Discussion

In this thesis, I have investigated missing data methods incorporating missingness information to deal with partially observed confounder data when using causal inference methods in observational studies. In particular, I have focused on the missingness pattern approach (MPA) for propensity score analysis, as well as the related missing indicator approach for propensity score analysis and for outcome regression.

9.1 Objective 1: Exploring the assumptions of the missingness pattern approach

The first objective of my thesis was to explore the assumptions under which the MPA would provide valid inference, by: (i) investigating the connection between the MPA and the conventional classification of missing data proposed by Rubin (1976) [20], (ii) identifying settings where the assumptions are likely to be plausible, and (iii) developing ways of assessing the assumptions.

Two methodological articles in the literature, D'Agostino and Rubin (2000) and Mattei (2009), have proposed assumptions underlying the validity of the MPA [46, 63]. However, the connection between these two sets of assumptions is unclear. Nor is it clear how the assumptions relate to the conventional classification of missingness mechanisms proposed by Rubin [20, 44] or how to assess whether or not the assumptions hold.

In this thesis, I have clarified the connection between the two sets of assumptions given in the literature, finding that the set of assumptions proposed by Mattei (2009) [63] is a wider statement of the assumptions underlying the MPA than the assumption stated by D'Agostino and Rubin (2000) [46]. Following Mattei's statement of the assumptions sufficient for valid inference using the MPA [63], I have stated weaker assumptions that hold separately for each potential outcome, rather than holding jointly. I have proved that the MPA can obtain unbiased estimates of the treatment effect under these weaker assumptions.

In addition, I have found that the assumptions underlying the MPA are separate from the conventional classification of missingness mechanisms, as classifying missing data according to Rubin's taxonomy is not informative with respect to assessing plausibility of the MPA's assumptions.

In order to be able to assess the plausibility of the MPA's assumptions, I used single world intervention graph templates (SWITs) incorporating the missing indicator. I then adapted these causal diagrams in order to be able to assess the CIT and CIO assumptions. My initial strategy was to construct SWITs where the confounder node was split into two component parts: the observed confounder values and the missing confounder values. However, this strategy required the use of deterministic arrows which meant that applying d-separation may not identify all conditional independencies. Thus, I used an alternative strategy to modify SWITs for use in assessing the CIT and CIO assumptions, where SWITs are constructed by conditioning on missingness patterns, and d-separation is applied to the SWIT(s) representing the pattern(s) with missing data.

9.2 Objective 2: Guidance for assessing the assumptions underlying the missingness pattern approach

The second objective of my thesis was to develop practical guidance for assessing the MPA's assumptions in a given setting. Prior to this work, no guidance existed

for how to assess the MPA's assumptions; this may be a major factor as to why the MPA is not used much in practice.

In this thesis, I have used causal diagrams to explore when the MPA's assumptions are violated in a range of simple settings, by varying: the causal relationships between confounders and treatment allocation, the causal relationships between confounders and outcome, and causal relationships with the missing indicator. I considered all combinations of the factors considered, applying d-separation to causal diagrams representing each scenario and employing the use of twin networks when treatment allocation or potential outcomes has a causal effect on the missing indicator.

On the basis of the results, I have developed guidance for assessing the MPA's assumptions. Initially, I developed the guidance in the form of a decision tree and provided a worked example of how it could be used to assess the assumptions in settings with a single partially observed confounder and restricted to certain temporal assumptions.

I then developed more comprehensive guidance in a step-by-step format that focuses on assessing the plausibility of possible violations and recommends the use of causal diagrams to help assess the plausibility of the MPA's assumptions.

The step-by-step guidance for assessing the MPA's assumptions in practice begins by considering whether or not it is plausible that the confounder with missing data is only a confounder when observed. In other words, given that the observed confounder values do indeed confound the relationship between treatment allocation and the potential outcomes, either the missing confounder values have no effect on treatment allocation or the missing confounder values have no effect on the potential outcomes (or both). If it is decided that it is indeed plausible that the confounder is only a confounder when it is observed, then the next step is to assess the plausibility of key violations in the setting of interest. If it is considered plausible that these violations are not present, then I recommend constructing a causal diagram where possible and using the d-separation rule to assess the MPA's assumptions.

I have demonstrated the step-by-step guidance in a real-data example, discussing

the clinical context of the example setting and showing how this information is utilised when assessing the plausibility of the MPA's assumptions. In addition, I have provided code for constructing causal diagrams and applying d-separation using the DAGitty package in R, both for a general hypothetical example as well as for the real world example.

9.3 Objective 3: The missing indicator approach for propensity score analysis

My third thesis objective was to explore the relationship between the MPA and the missing indicator approach in the context of propensity score analysis, in particular investigating the implications of this relationship on the assumptions under which the missing indicator approach can provide valid inference.

The missing indicator approach has been criticised in the missing data methodology literature for being an 'ad hoc' missing data method [56,57] that leads to biased results in general [58, 59]. Whilst Stuart (2010) [55] and Williamson and Forbes (2014) [25] have recommended the use of the missing indicator approach to handle partially observed confounder data in propensity score analysis, no formal evidence has been yet provided to support their recommendations. Despite this, incorporating missing indicators into regression models is a popular method in epidemiological studies [52]. Indeed, my motivating example is from a pharmacoepidemiological study using electronic health records that includes an 'absent' category in the baseline chronic kidney disease (CKD) stage variable. Baseline CKD stage is more likely to be recorded for patients with poorer kidney function and so excluding patients missing baseline CKD stage would induce selection bias. Hence my third thesis objective is to consider whether using missing indicators was an appropriate approach to handing missing confounder data.

In this thesis, I have shown a clear mathematical connection between the between the MPA and the missing indicator approach, finding that the missing indicator approach is a simplification of the MPA. I also demonstrated that the missing

indicator approach can provide valid inference under the MPA's assumptions and an additional parametric assumption that there are no interactions between fully observed confounders and the missing indicator in the propensity score model. Hence, I have shown that the missing indicator approach for dealing with partially observed confounder data can provide valid inference and so enable researchers to use this approach in a principled way.

9.4 Objective 4: The missing indicator approach for outcome regression

My fourth thesis objective was to investigate the use of the missing indicator approach in the context of outcome regression. The general missing data methodology literature consider the missing indicator approach to be 'ad hoc' and unprincipled, and recommend avoiding the approach, with some theoretical work [59] showing that the approach is generally biased. Indeed, our result from Objective 3 showed that the missing indicator approach can provide valid inference in the propensity score context. The continued use of the missing indicator approach in outcome regression settings, such as our motivating example encourages a further consideration of this approach.

In this thesis, I have proved that the missing indicator approach can obtain valid inference in the context of outcome regression under the MPA's assumptions and further parametric assumptions. Furthermore, I have shown how this finding relates to the results published by Jones (1996) [59].

In particular, although Jones ultimately concludes that the missing indicator approach for outcome regression is generally biased, Jones notes situations where the approach is unbiased, including scenarios where data are fully observed. I have shown that the other situation highlighted by Jones corresponds to settings where the MPA's assumptions — in particular the CIT assumption — hold. I have also shown that the assumed true outcome model in Jones (1996) [59] precludes the CIO assumption from being satisfied. If instead, the CIO assumption does hold, then

the true outcome model incorporates the missing indicator as a covariate, and the missing indicator approach can thus provide valid inference.

In addition, I have implemented a simulation study to explore the extent of bias introduced in outcome regression when the assumptions underling the missing indicator approach are violated, finding that weak violations of the assumptions may not introduce substantial amounts of bias compared to strong violations. A key finding from the simulation studies conducted during the PhD was that the conventional approach of generating missing data after the main data generation of confounders, treatment and outcome leads to violations of both the CIT and CIO assumptions, and thus the MPA would generally yield biased results. Instead, I discovered that for the CIT or CIO assumptions to be able to hold, the missingness generation must be incorporated into the main data generation process: the missingness pattern must be able to influence the treatment allocation and/or the potential outcomes.

9.5 Dissemination of my research so far

In this thesis, I have included two papers that have been developed over the course of my PhD. The first paper, providing step-by-step guidance for assessing the MPA's assumptions in propensity score analysis, has been resubmitted to Statistics in Medicine after undergoing review and revisions based on the reviewers' comments. The second paper, exploring the use of the missing indicator approach in the context of outcome regression, has undergone revisions based on reviewer comments and has been accepted by the Biometrical Journal pending further minor revisions. Other dissemination efforts include presentations at two international conferences and internal meetings throughout the PhD.

In March 2017, I presented a poster on the decision-tree guidance at a research degree poster day at the London School of Hygiene and Tropical Medicine for a general audience. I also presented this work in more detail at a meeting in May 2017 for researchers involved in missing data methodological research, and in July 2017 at the 38th Annual Conference of the International Society for Clinical Biostatistics in Vigo, Spain.

In August 2018 I presented preliminary work, on extending the work on the missing indicator approach being a principled approach for handling missing confounder data in propensity score analysis to the context of outcome regression, at the Joint International Society for Clinical Biostatistics and Australian Statistical Conference 2018 in Melbourne, Australia.

9.6 Further areas for research

During the course of my PhD, I have identified areas for further work. First, whilst the MPA is a relatively simple method to implement compared to multiple imputation, as the number of missingness patterns increases (and thus the number of propensity models increases), implementation of the MPA in standard software may become more complex. Further work would be required to write functions for implementing the MPA in standard software such as Stata and R, including calculating appropriate standard errors.

Another issue when dealing with a large number of missingness patterns is sparsity of data: some missingness patterns may be rare and thus lead to model estimation issues. Approaches for dealing with sparsity have been suggested in the literature. D'Agostino et al. (2001) suggested a variation of the MPA where, in each missingness pattern, propensity scores are estimated in the wider group of all patients with observed data for the relevant confounders but retaining the estimated propensity scores only for those who actually had that missingness pattern [64]. However, this approach may lead to correlation issues as patients will be used in multiple patterns. How to deal with this correlation in order to calculate appropriate standard errors requires further consideration.

An alternative approach to dealing with sparse missingness patterns was suggested by Qu and Lipkovich (2009). Their suggested approach was a pattern pooling algorithm that groups 'similar' missingness patterns using a distance metric in order to obtain a set of pooled patterns that have a minimum sample size [70]. Further work is required to investigate the performance of this algorithm in practice and to compare pattern pooling to the D'Agostino modification of the MPA. Writing func-

tions in standard software for implementing the pattern pooling algorithm would also be a valuable contribution.

Qu and Lipkovich (2009) also proposed an approach that combines multiple imputation and the MPA [70]. Seaman and White (2014) explored this approach [80], but further consideration is required to investigate assumptions underlying the method and their connections with the MPA's assumptions and Rubin's taxonomy. It would be also interesting to explore the use of multiple imputation for dealing with missing outcome or treatment allocation values with missingness pattern-type approaches to deal with partially observed confounders, and other combinations of missing data methods.

It is well known in the propensity score literature that propensity score analysis is a two-stage process: the first stage estimates the propensity score and the second stage uses this estimated propensity score in the estimation of the treatment effect [25, 88]. Thus, when estimating the variance of the treatment effect estimate, if the propensity score estimation stage is not taken into account, then the standard error is likely to be conservative, with this loss of precision being a greater issue in scenarios with a continuous outcome [25, 88]. Williamson et al. (2014) [88] have derived a variance estimator for propensity score analysis in complete data. This has not yet been extended to settings where the MPA is used to deal with multiple missing confounders, and is a key area for further work.

The simulation studies in this thesis are relatively simple. Future work could explore magnitudes of bias in realistic settings by using plasmode simulation studies.

Another opportunity for further work would be to extend the MPA to settings with more than two treatment arms. In addition, it is not currently clear how to assess balance after use of the MPA or missing indicator approach.

Future work could be to consider the use of the MPA when using alternative methods of estimating the propensity score, such as classification trees, random forests and generalised boosted modelling [40]. Using non-parametric approaches to estimating the propensity score, would enable development of a parsimonious version of the MPA model, including only necesssary interactions (as opposed to

using the MPA-equivalent approach of including covariates, missing indicators and all interactions between missing indicators and covariates).

Another key area for further research is how to perform sensitivity analyses for violation of the MPA's assumptions in real-world data examples.

9.7 Implications for research

Researchers must carefully consider the assumptions underlying approaches for handling missing confounder data in causal inference. If considering employing missingness pattern-type approaches, it is not enough to consider which missingness mechanism seems most plausible. Researchers should carefully consider the plausibility of the MPA's assumptions in the setting of interest prior to analysis. This can be achieved by following the practical guidance developed over the course of my PhD, in particular by: (i) considering the clinical context to assess whether it is plausible that the confounder(s) with missing data is only a confounder when observed, (ii) considering the key violations identified in the guidance, and (iii) constructing a causal diagram to represent the scenario of interest. The plausibility of the MPA's assumptions should be considered prior to analysis to enable capture of predictors of missingness that might have otherwise been completely unobserved.

Advantages of the MPA and the missing indicator approach are that they are simple to understand and they retain all patients in the analysis whether or not they have missing data. In addition, these approaches may be appropriate where multiple imputation is not as they do not rely upon data being missing at random. However, there are still some key areas of research that could be explored, including: dealing with sparseness and settings with many missingness patterns; developing ways to assess covariate balance after using missingness pattern-type approaches in propensity score analysis; and investigating how to perform sensitivity analyses for violations of the MPA's assumptions. Settings in which the missingness pattern approach and the missing indicator approach are likely to be useful are studies using routinely collected data such as electronic health records. In these settings, whilst the CIO assumption may be less plausible, the CIT assumption is likely to be

plausible as making decisions about treatment allocation can only take into account the information available.

The information available to the researcher using routinely collected data is in general the same as the information that was available to the clinician making the treatment decision. Thus, we believe it is plausible to assume that treatment allocation is not associated with confounder information that is unavailable to the researcher. Hence, research using electronic health records is a key area in which the CIT assumption is likely to be plausible.

Furthermore, as studies using large health datasets become ever more popular with the increasing popularity of big data, approaches using missingness patterns to handle partially observed confounder data may be less computationally intensive than multiple imputation.

9.8 Conclusion

Using missingness patterns to deal with missing confounder data is a simple alternative to conventional missing data methods which can provide valid inference under certain assumptions. In this thesis, I have clarified the connection between the sets of assumptions given in the literature and I have found that classifying missing data according to Rubin's taxonomy is not informative with respect to assessing plausibility of the assumptions underlying the missingness pattern approach. I have provided guidance for assessing the plausibility of these assumptions in practice. I have also shown that using missing indicators to deal with missing confounder data is a simplified version of the missingness pattern approach, and thus is a principled approach in propensity score analysis and outcome regression contexts.

Appendix A

Ethics approval

In this appendix, I include:

- the cover letter and application for ethics approval from CPRD by an Independent Scientific Advisory Committee
- the favourable ethics approval letter from CPRD
- the (redacted) application form for ethics approval from LSHTM
- the favourable ethics approval letter from LSHTM

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT
United Kingdom
Switchboard: +44 (0)20 7636 8636

www.lshtm.ac.uk

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

The Chair
ISAC
The Medicines and Healthcare products Regulatory Agency
CPRD Division
151 Buckingham Palace Road
Victoria
London SW1W 9SZ

Wednesday 30th August 2017

Dear Sir or Madam,

Re: Amendment to ISAC protocol number 14_208 - The incidence and mortality of acute kidney injury associated with prescribing of angiotensin converting inhibitors and angiotensin-receptor blockers

We have undertaken some mathematical work investigating various approaches to dealing with missing data in collaboration with two of the investigators named on the ISAC above (Kathryn Mansfield and Laurie Tomlinson). Our theoretical results suggest that methods based on a simple approach of creating an extra variable indicating which values of a confounder are missing, and incorporating this into the analysis, may be valid in a wide range of settings.

We believe our results will be particularly helpful in the context of studies using electronic health record data. Therefore, we would like to illustrate our theoretical results by applying a number of approaches for dealing with missing data to an example using CPRD data (specifically, the study described in the ISAC above). This would involve presenting the results from a number of sensitivity analyses to assess the impact of missing data. No additional data would be required.

As requested in correspondence with CPRD (reference: CPRD00002581), we are submitting an amendment to ISAC protocol number 14_208.

I look forward to hearing from you.

Yours sincerely,

[redacted]

Helen Blake
PhD student
Department of Medical Statistics
Faculty of Epidemiology and Population Health
E: Helen.Blake@lshtm.ac.uk

ISAC APPLICATION FORM

PROTOCOLS FOR RESEARCH USING THE CLINICAL PRACTICE RESEARCH DATALINK (CPRD)

ISAC use only: Protocol Number Date submitted		**IMPORTANT** **If you have any queries, please contact ISAC Secretariat:** ISAC@cprd.com

1. Study Title

The incidence and mortality of acute kidney injury (AKI) associated with prescribing of angiotensin converting inhibitors and angiotensin-receptor blockers

2. Principal Investigator (full name, job title, organisation & e-mail address for correspondence regarding this protocol)

Dr Laurie Tomlinson, Lecturer, London School of Hygiene and Tropical Medicine and Honorary Consultant Nephrologist at Brighton and Sussex University Hospitals NHS Trust. laurie.tomlinson@lshtm.ac.uk

3. Affiliation (full address)

Department of Non-communicable Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT

4. Protocol's Author (if different from the principal investigator)

Kathryn Mansfield

5. List of all investigators/collaborators (*please list the names, affiliations and e-mail addresses* of all collaborators*, other than the principal investigator)

Kathryn Mansfield, LSHTM, kathryn.mansfield@lshtm.ac.uk
Krishnan Bhaskaran, LSHTM, krishnan.bhaskaran@lshtm.ac.uk
Dorothea Nitsch, LSHTM, dorothea.nitsch@lshtm.ac.uk
Liam Smeeth, LSHTM, liam.smeeth@lshtm.ac.uk
Helen Blake, LSHTM, helen.blake@lshtm.ac.uk
Elizabeth Williamson, LSHTM and Farr Institute of Health Informatics, elizabeth.williamson@lshtm.ac.uk
Clémence Leyrat, LSHTM, clemence.leyrat@lshtm.ac.uk
Ian White, MRC Clinical Trials Unit at UCL, ian.white@ucl.ac.uk
Shaun Seaman, MRC Biostatistics Unit at Cambridge Institute of Public Health, shaun.seaman@mrc-bsu.cam.ac.uk
James Carpenter, LSHTM and MRC Clinical Trials Unit at UCL, james.carpenter@lshtm.ac.uk

Please note that your ISAC application form and protocol **must be copied to all e-mail addresses listed above at the time of submission of your application to the ISAC mailbox. Failure to do so will result in delays in the processing of your application.*

6. Type of Institution (please tick one box below)

Academia	☒	Research Service Provider	☐	Pharmaceutical Industry	☐
NHS	☐	Government Departments	☐	Others	☐

7. Financial Sponsor of study

Pharmaceutical Industry *(please specify)*	☐	Academia*(please specify)*	☒ Wellcome Trust
Government / NHS *(please specify)*	☐	None	☐
Other *(please specify)*	☐		

8. Data source *(please tick one box below)*

Sponsor has on-line access	☒	Purchase of ad hoc dataset	☐
Commissioned study	☐		
Other	☐ *(please specify)*		

9. Has this protocol been peer reviewed by another Committee?

Yes* ☐ No ☒

** Please state in your protocol the name of the reviewing Committee(s) and provide an outline of the review process and outcome.*

10. Type of Study *(please tick all the relevant boxes which apply)*

Adverse Drug Reaction/Drug Safety	☒	Drug Use	☐	Disease Epidemiology	☒
Drug Effectiveness	☐	Pharmacoeconomic	☐	Other	☐

11. This study is intended for:

Publication in peer reviewed journals	☒	Presentation at scientific conference	☒
Presentation at company/institutional meetings	☒	Other	

12. Does this protocol also seek access to data held under the CPRD Data Linkage Scheme?

Yes ☒ No ☐

13. If you are seeking access to data held under the CPRD Data Linkage Scheme*, please select the source(s) of linked data being requested.

☒ Hospital Episode Statistics ☐ Cancer Registry Data**
☐ MINAP ☐ ONS Mortality Data
☒ Index of Multiple Deprivation/ ~~Townsend Score~~
☐ Mother Baby Link ☐ Other: (please specify)

** As part of the ISAC review of linkages, the protocol may be shared - in confidence - with a representative of the requested linked data set(s) and summary details may be shared - in confidence - with the Confidentiality Advisory Group of the Health Research Authority.*

***Please note that applicants seeking access to cancer registry data must provide consent for publication of their study title and study institution on the UK Cancer Registry website. Please contact the CPRD Research Team on +44 (20) 3080 6383 or email kc@cprd.com to discuss this requirement further.*

14. If you are seeking access to data held under the CPRD Data Linkage Scheme, have you already discussed your request with a member of the Research team?

Yes ☒ No* ☐

**Please contact the CPRD Research Team on +44 (20) 3080 6383 or email kc@cprd.com to discuss your requirements before submitting your application.*

Please list below the name of the person/s at the CPRD with whom you have discussed your request.
Discussed with Kendal Chidwick via email (MHRA/CPRD enquiry reference: OCR2987).

15. If you are seeking access to data held under the CPRD Data Linkage Scheme, please provide the following information:

The number of linked datasets requested: 2

A synopsis of the purpose(s) for which the linkages are required:
Hospital Episode Statistics

1. Primary outcome measure: To identify additional cases of the primary outcome, acute kidney injury, based on hospital morbidity coding.
2. Secondary outcome measures: We also intend to use end stage renal disease (ESRD) as a secondary outcome measure and hope to identify cases based on both primary care and hospital morbidity coding.
3. Time at risk: We will undertake a secondary analysis excluding hospital admission time from time at risk.

Index of Multiple Deprivation
We intend to use IMD as a covariate in analyses.

Is linkage to a local dataset with <1 million patients being requested?

Yes* ☐ No ☒

** If yes, please provide further details:* N/A

16. If you have requested linked data sets, please indicate whether the Principal Investigator or any of the collaborators listed in response to question 5 above, have access to any of the linked datasets in a patient identifiable form, or associated with a patient index.

Yes* ☐ No ☒

** If yes, please provide further details:* N/A

17. Does this protocol involve requesting any additional information from GPs?

Yes* ☐ No ☒

** Please indicate what will be required:*

Completion of questionnaires by the GP[v] Yes ☐ No ☒
Provision of anonymised records (e.g. hospital discharge summaries) Yes ☐ No ☒
Other (please describe): N/A

[v] *Any questionnaire for completion by GPs or other health care professional must be approved by ISAC before circulation for completion.*

18. Does this protocol describe a purely observational study using CPRD data (this may include the review of anonymised free text)?

Yes* ☒ No** ☐

** Yes: If you will be using data obtained from the CPRD Group, this study does not require separate ethics approval from an NHS Research Ethics Committee.*
*** No: You may need to seek separate ethics approval from an NHS Research Ethics Committee for this study. The ISAC will provide advice on whether this may be needed.*

19. Does this study involve linking to patient *identifiable* data from other sources?

Yes ☐ No ☒

20. Does this study require contact with patients in order for them to complete a questionnaire?

Yes ☐ No ☒

N.B. Any questionnaire for completion by patients must be approved by ISAC before circulation for completion.

21. Does this study require contact with patients in order to collect a sample?

Yes* ☐ No ☒

** Please state what will be collected* N/A

22. Experience/expertise available

Please complete the following questions to indicate the experience/expertise available within the team of researchers actively involved in the proposed research, including analysis of data and interpretation of results

	Previous GPRD/CPRD Studies	Publications using GPRD/CPRD data
None	☐	☐
1-3	☐	☐
> 3	☒	☒

	Yes	No
Is statistical expertise available within the research team?	☒	☐
If yes, please outline level of experience	*very experienced*	
Is experience of handling large data sets (>1 million records) available within the research team?	☒	☐
If yes, please outline level of experience	*very experienced*	
Is UK primary care experience available within the research team?	☒	☐
If yes, please outline level of experience	*very experienced*	

23. References relating to your study

Please list up to 3 references (most relevant) relating to your proposed study.

Tomlinson LA, Abel GA, Chaudhry AN, Tomson CR, Wilkinson IB, Roland MO, et al. ACE inhibitor and angiotensin receptor-II antagonist prescribing and hospital admissions with acute kidney injury: a longitudinal ecological study. PLoS One. 2013;8(11):e78465.

Lapi F, Azoulay L, Yin H, Nessim SJ, Suissa S. Concurrent use of diuretics, angiotensin converting enzyme inhibitors, and angiotensin receptor blockers with non-steroidal anti-inflammatory drugs and risk of acute kidney injury: nested case-control study. BMJ. 2013;346–8525.

PROTOCOL CONTENT CHECKLIST

In order to help ensure that protocols submitted for review contain adequate information for protocol evaluation, ISAC have produced instructions on the content of protocols for research using CPRD data. These instructions are available on the CPRD website (www.cprd.com/ISAC). All protocols using CPRD data which are submitted for review by ISAC must contain information on the areas detailed in the instructions. IF you do not feel that a specific area required by ISAC is relevant for your protocol, you will need to justify this decision to ISAC.

Applicants must complete the checklist below to confirm that the protocol being submitted includes all the areas required by ISAC, or to provide justification where a required area is not considered to be relevant for a specific protocol. Protocols will not be circulated to ISAC for review until the checklist has been completed by the applicant.

Please note, your protocol will be returned to you if you do not complete this checklist, or if you answer 'no' and fail to include justification for the omission of any required area.

	Included in protocol?		
Required area	Yes	No	If no, reason for omission
Lay Summary (max.200 words)	☒	☐	
Background	☒	☐	
Objective, specific aims and rationale	☒	☐	
Study Type *Descriptive* *Hypothesis Generating* *Hypothesis Testing*	 ☐ ☐ ☒	 ☐ ☐ ☐	
Study Design	☒	☐	
Sample size/power calculation (Please provide justification of sample size in the protocol)	☒	☐	
Study population (including estimate of expected number of relevant patients in the CPRD)	☒	☐	
Selection of comparison group(s) or controls	☒	☐	
Exposures, outcomes and covariates *Exposures are clearly described* *Outcomes are clearly described*	☒ ☒ ☒	☐ ☐ ☐	
Use of linked data (if applicable)	☒	☐	
Data/ Statistical Analysis Plan *There is plan for addressing confounding* *There is a plan for addressing missing data*	☒ ☒ ☒	☐ ☐ ☐	
Patient/ user group involvement †	☒	☐	
Limitations of the study design, data sources and analytic methods	☒	☐	
Plans for disseminating and communicating study results	☒	☐	

† ***It is expected that many studies will benefit from the involvement of patient or user groups in their planning and refinement, and/or in the interpretation of the results and plans for further work. This is particularly, but not exclusively true of studies with interests in the impact on quality of life. Please indicate whether or not you intend to engage patients in any of the ways mentioned above.***

Voluntary registration of ISAC approved studies:

Epidemiological studies are increasingly being included in registries of research around the world, including those primarily set up for clinical trials. To increase awareness amongst researchers of ongoing research, ISAC encourages voluntary registration of epidemiological research conducted using MHRA databases. This will not replace information on ISAC approved protocols that may be published in its summary minutes or annual report. It is for the applicant to determine the most appropriate registry for their study. Please inform the ISAC secretariat that you have registered a protocol and provide the location.

Study Protocol

1. Title

The incidence and mortality of acute kidney injury (AKI) associated with prescribing of ACE inhibitors and angiotensin-receptor blockers

2. Lay Summary

This study uses very large numbers of linked electronic health records to answer important questions about episodes of sudden decline in kidney function with the aim of preventing serious illness and death, and creating substantial savings for the NHS. A sudden decrease in kidney function or acute kidney injury (AKI) is common and associated with an increased risk of death, prolonged hospital stay and risk of permanent kidney failure. Rates of AKI are increasing and causing significant cost to the NHS. Some limited evidence suggests that AKI can occur as a side effect of angiotensin converting enzyme inhibitors (ACEI) and angiotensin-receptor blockers (ARBs), particularly when prescribed with water tablets (diuretics) and anti-inflammatory painkillers. ACEI and ARBs are commonly prescribed for conditions such as high blood pressure and heart problems. At present it is not known how common AKI is among people taking these drugs or whether there are any conditions (e.g. diabetes, existing kidney problems) that modify the risk. This is important because if it was understood who is likely to develop AKI and in what circumstances, strategies could be developed to prevent AKI.

3. Objectives

The overall aims of the project are to investigate the incidence and mortality of acute kidney injury (AKI) associated with prescribing of ACE inhibitors and angiotensin-receptor blockers, and other commonly prescribed antihypertensive drugs (calcium channel blockers, beta-blockers, and diuretics), and to investigate what chronic comorbidities are associated with the development of drug-associated AKI.

Specifically:

1. To evaluate the validity of an operational case definition for AKI based on morbidity coding (from Read-coded primary care data and ICD-10 coded hospital data) and biochemical test results.
2. To obtain estimates for the incidence of antihypertensive-associated AKI in the UK general population and investigate the variation in incidence over time.
3. To explore differences in the rate of antihypertensive-associated AKI in different classes of antihypertensive drugs (ACEI/ARBs, beta-blockers, calcium channel blockers and diuretics).
4. To determine which chronic comorbidities are associated with increased risk of drug-associated AKI.
5. To determine rates of mortality, hyperkalaemia and dialysis following drug-associated AKI.

4. Background

Acute kidney injury (AKI) is a sudden (within hours or days) decline in renal function. It is associated with increased mortality (1,2) and increased duration of hospital stay (3,4). AKI has been observed in 15–20% of hospital admissions (2,5,6). It has been estimated that the annual cost of AKI inpatient care in England is £1.02 billion, a little over 1% of the NHS budget (4).

AKI has been variably defined based on changes in serum creatinine and urine output (See Appendix 1, Table A1.1). Estimates for annual AKI hospital incidence differ based on varying AKI definitions ranging from 1,811 per million population using the RIFLE criteria (Risk, Injury, Failure, Loss and End stage) (7), to 2,400 per million population using administrative coding (8) (13), and 15,325 per million population using the AKIN (Acute Kidney Injury Network) definition (2).

Angiotensin converting enzyme inhibitors (ACEI) and angiotensin receptor blockers (ARB) are often used in the management of hypertension and cardiac failure. There is evidence that combinations of ACEI/ARBs, non-steroidal anti-inflammatory drugs, and diuretics may impair renal function (9–12). However, there is only limited evidence of an association between AKI and ACEI/ARBs alone (13,14). One study has suggested an increase in post-operative AKI in cardiovascular surgery patients taking ACEI/ARBs preoperatively (13). While other existing evidence for an association between AKI and ACEI/ARBs comes from an ecological study and is therefore limited by lack of patient level data (14). To reduce the potential adverse effects associated with ACEI/ARBs we need a better understanding of individual level risk factors for AKI associated with these drugs.

Current consensus suggests that ACEI/ARBs should be withheld during acute illness, however the evidence supporting this is limited (15,16) (11,20). This is in part because observational studies on this topic are confounded by indication. The indications for ACEI/ARB prescription are themselves associated with increased risk of AKI. Therefore an observed increased incidence of AKI may reflect increasing prevalence of comorbidities rather than a causal effect of the drugs themselves.

We aim to investigate the association between ACEI/ARBs and AKI (drug-associated AKI). We will calculate the incidence of AKI in those prescribed ACEI/ARBs and compare this to AKI incidence in a number of control groups. We have selected our control groups to avoid confounding by indication. In addition to a group of matched (age, gender and GP practice) controls with no prescriptions for medications with similar indications to those for ACEI/ARBs, we will also look at AKI incidence in those prescribed other classes of antihypertensive medications (i.e. drugs prescribed for indications similar to those for ACEI/ARBs). Our control groups will therefore be: i) those prescribed beta-blockers (BB); ii) those prescribed calcium channel blockers (CCBs); iii) those prescribed thiazide diuretics; and iv) an age, gender and GP practice matched control group not prescribed antihypertensives (ACEI/ARBs, BBs, CCBs or thiazide diuretics). We will investigate changes in AKI incidence rates over time, changes in AKI incidence when ACEI/ARBs are prescribed in drug combinations thought to be associated with impaired renal function (ACEI/ARB drugs plus other diuretics and non-steroidal anti-inflammatory drugs), and investigate the mortality, and rate of progression to end-stage renal disease (ESRD) in drug-associated AKI.

5. Study type

This will primarily be a hypothesis testing study. The null hypothesis is that ACEI/ARBs do not increase the risk of AKI compared to other antihypertensive drugs.

6. Study design

This will be a new-user cohort study with time-updating exposure status, using CPRD data and linked HES data.

7. Study population

We will use data from general practices in CPRD that have consented to Hospital Episode Statistics (HES) data linkage. The study period will cover the period for which there is HES data linkage with the CPRD database; from April 1997 to March 2012. However, if an updated version of CPRD linked HES data becomes available at an appropriate point in the project timeline (i.e. before data extraction) we will use the most recent version of the linked data available, which will result in a later end to the study period and maximise follow-up time.

We will retrieve data on all patients aged 18 or over who do not have end stage renal disease (ESRD), who have no record of a prescription for antihypertensive medication (ACEI/ARBs, beta-blockers, calcium channel blockers, or thiazide diuretics) within the 12 months prior to cohort entry, who have at least one serum creatinine result recorded at any time from 12 months prior to cohort entry onwards (in order to establish CKD status – see Section 10.4.2), and who have a new prescription for one or more of the following: i) ACEI/ARBs; ii) beta blockers (BB); iii) calcium channel blockers (CCB); or iv) thiazide diuretics, in addition to an age, sex and GP practice matched control group on none of these drugs.

7.1 Cohort entry

Cases (antihypertensive users) will enter the cohort at first prescription for an antihypertensive (new use of ACEI/ARB, BB, CCB, or thiazide diuretic). Controls will enter the cohort on the same date as their matched cases. Patients will be eligible for cohort entry from the latest of: i) one year after practice registration date; ii) date practice reached CPRD quality control standards; or iii) 18th birthday.

We have chosen to use a new-user cohort (i.e. entry to cohort on new use of the drugs of interest). If we were to include existing users of these drugs, we would introduce adherence bias, since those who have remained on the drug will be systematically different to those who stop taking a drug after the first prescription due to early adverse effects. In addition we may miss important outcomes in those who entered the cohort who were already prescribed the drugs of interest.

7.2 Cohort exit

Individuals will be eligible until the first of: i) date of death; ii) patient transferred out of practice; iii) last data collection from the practice; or iv) ESRD diagnosis.

ESRD will be defined based on hospital and primary care morbidity coding as: i) presence of an ESRD morbidity code; ii) a code for renal transplant; iii) a code for peritoneal dialysis; iv) two or more haemodialysis codes more than 90 days apart; v) stage 5 chronic kidney disease (CKD); or vi) stage 4 CKD with a fistula (this suggests rapidly worsening renal function).

8. Outcome

The primary outcome of interest will be AKI. AKI cases will be identified from three sources: i) primary care morbidity coding (Read-codes); ii) hospital admission morbidity coding (ICD-10 codes); and iii) biochemical results recorded in primary care health records.

8.1 Morbidity codes

Primary care (Read) and hospital (ICD-10) morbidity codes for AKI will be identified by a consensus exercise. A list of search terms to identify AKI will be developed using the Medline medical subject headings (MeSH) thesaurus (see Appendix 2, Table A2.1). These search terms will be applied in both Read and ICD-10 code browsers. We (Laurie Tomlinson (LT) and Kate Mansfield (KM)) will classify the two lists of codes (Read and ICD-10) returned by the search terms independently as either probably representing an episode of AKI or possibly identifying an episode of AKI. The code lists we generate will be compared and any disagreements discussed in order to generate one list of codes that definitely represent an AKI episode. A previous study has already investigated the positive predictive value of the ICD-10 code N17 in UK hospital admissions data for the KDIGO AKI definition and found it to be accurate for 95% of cases (17).

Lists of Read and ICD-10 codes for AKI are provided in Appendix 2, Tables A2.3 and A2.4, and are illustrative of the final lists that will be generated from the more rigorous code list development process described above.

8.2 Serum creatinine algorithm

We have developed an algorithm to identify community cases of AKI based on changes in serum creatinine recorded in primary care health records. Our algorithm is based on the 2013 Association for Clinical Biochemistry and Laboratory Medicine (ACB) algorithm for AKI (18). The ACB algorithm was developed to generate e-alerts using electronic hospital lab data based on the KDIGO guidelines for AKI (19). Applying this unchanged to community data would be imprudent since the frequency of renal function testing in a hospital setting is likely to be quite different to that in primary care. Compared to a hospital setting, serial serum creatinine measurements in the community are likely to be separated by longer intervals. This increases the likelihood of misclassifying a gradual decline in renal function as AKI.

To avoid this sort of misclassification, our algorithm for diagnosis of AKI in the community applies the ACB algorithm only in those recorded with morbidity codes for acute conditions likely to cause AKI (e.g. acute infections) but not sufficiently severe as to as to warrant immediate hospital admission. It is assumed that in severe acute conditions needing hospital admission (e.g. sepsis or gastrointestinal bleeding) if AKI occurs it will be recorded in hospital records and therefore picked up by ICD-10 coding in linked HES records.

To be classified as having an episode of AKI based on changes in community serum creatinine measurements a patient must have:

1. Baseline serum creatinine measurement: A minimum of one serum creatinine result prior to the recording of the index acute morbidity code in order to determine baseline serum creatinine.
2. A primary care morbidity code for an acute condition that may precipitate AKI but does not necessarily require hospital admission. Operationalised as a Read code for gastroenteritis, urinary tract infection or lower respiratory tract infection (termed the index infection).
3. A change in serum creatinine classified as AKI according to the ACB algorithm in the two weeks following the record of an acute morbidity code. The maximum interval of two weeks between acute morbidity coding and serum creatinine result has been chosen pragmatically to allow time for the practicalities of community blood testing.

4. If a hospital admission occurs within the two weeks following the index acute morbidity code then, to be defined as AKI, the recorded creatinine change must occur between the index infection morbidity code and up to and including the day of hospital admission. If a hospital admission is recorded between the index-infection morbidity code and the date of the change in creatinine, this will not be defined as AKI since we cannot assume that the index infection recorded in primary care is related to the change in creatinine.

Codes for acute infections that may precipitate AKI (urinary tract infection, gastroenteritis and respiratory tract infection) will be identified using a similar consensus approach to that presented in the previous section on identifying morbidity codes for AKI (Section 8.1).

To explore the validity of our AKI definition we will compare incidence rates calculated using our measures of AKI (using a combination of both morbidity coding and biochemical test results) with those from a study in the Canadian general population (8) and those from a recent study using secondary care biochemical data in East Kent (2).

8.3 Secondary outcomes

We will investigate mortality, hyperkalaemia and ESRD following AKI (addressing objective five: To determine rates of mortality, hyperkalaemia and dialysis following drug-associated AKI). We will investigate overall mortality and mortality at 0–3 months following the AKI episode, 4–6 months and 7–12 months. Hyperkalaemia will be established using specific Read and ICD-10 morbidity codes and potassium levels ascertained in biochemical tests. ESRD will be defined using morbidity coding as above (see Section 7.2).

9. Exposure

9.1 Exposure status

In the main analysis the exposure of interest will be ACEI/ARB use. Patients prescribed ACEI/ARB may be at greater risk of the outcome (AKI) due to prescribing indication. The main analysis will therefore investigate how AKI incidence differs in those exposed to ACEI/ARBs compared to three other classes of antihypertensive drug (BB, CCB, and thiazide diuretics), in addition to a control group exposed to no antihypertensive medication.

Prescriptions for each class of antihypertensive will be identified from coded primary care consultation data. Drugs are uniquely identified in CPRD using codes. We will identify both the generic and brand names of relevant drugs using the British National Formulary (BNF). The BNF is an authoritative guide to UK prescription drugs. We will use the drug names to create search terms to identify relevant drug codes in a data file containing codes for all the available prescription drugs. We will exclude drugs that are not taken orally from the list. A list of search terms for each class of antihypertensive is included in Appendix 2 (Table A2.2).

A prescription does not necessarily mean that a patient has taken a drug. To indicate regular use of the drug one approach would be to develop an exposure measure based on continued repeat prescription. However, our primary outcome (AKI) is an acute event that may occur as an early adverse event following initiation of therapy. Our main analysis will therefore use time-updating exposure status where individuals can move between exposure groups based on changing prescriptions.

For the main analysis, exposure status will be defined in two different ways: i) time exposed to a single class of antihypertensive drug; and ii) exposure defined by multiple binary indicator variables for each class of antihypertensive or control status. We will repeat the analysis using each exposure definition. We will also conduct a secondary analysis looking at combination antihypertensive therapy in individuals prescribed ACEIs or ARBs.

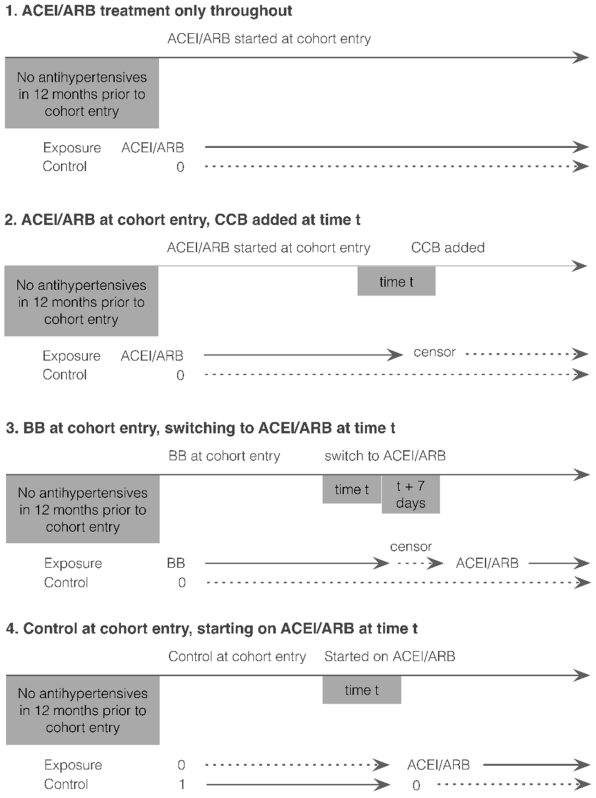
9.1.1 Exposure to a single class of antihypertensive

We will use time-updating variables to define time at risk to the different classes of drugs. Exposure will be defined in the following categories:

i. Time when prescribed an ACE/ARB (no other class of antihypertensives prescribed)
ii. Time when prescribed a BB (no other class of antihypertensives prescribed)
iii. Time when prescribed a CCB (no other class of antihypertensives prescribed)
iv. Time when prescribed a thiazide diuretic (no other class of antihypertensives prescribed)
v. Time contributed by a random sample of patients not prescribed ACEI/ARBs, BBs, CCBs or thiazide diuretics.

Figure 1 shows assignment of exposure status under four example scenarios. In scenario one the patient remains exposed to only one agent for the duration of the study. In scenario two, in the situation where a second antihypertensive is added, the patient is censored from follow-up (unless there is a further change in prescription resulting in prescription for a single agent). In scenario three, when a patient switches from one class of drug to a different class of drug, we account for a seven-day wash-out period between prescriptions, therefore start of time-at-risk to the second drug is delayed for seven days to allow for the practicalities of prescription fulfilment. In scenario four, where an individual initially selected to be part of the control group is started on an antihypertensive, the patient is censored from follow-up as a control and begins to contribute follow-up tome to ACEI/ARB exposure.

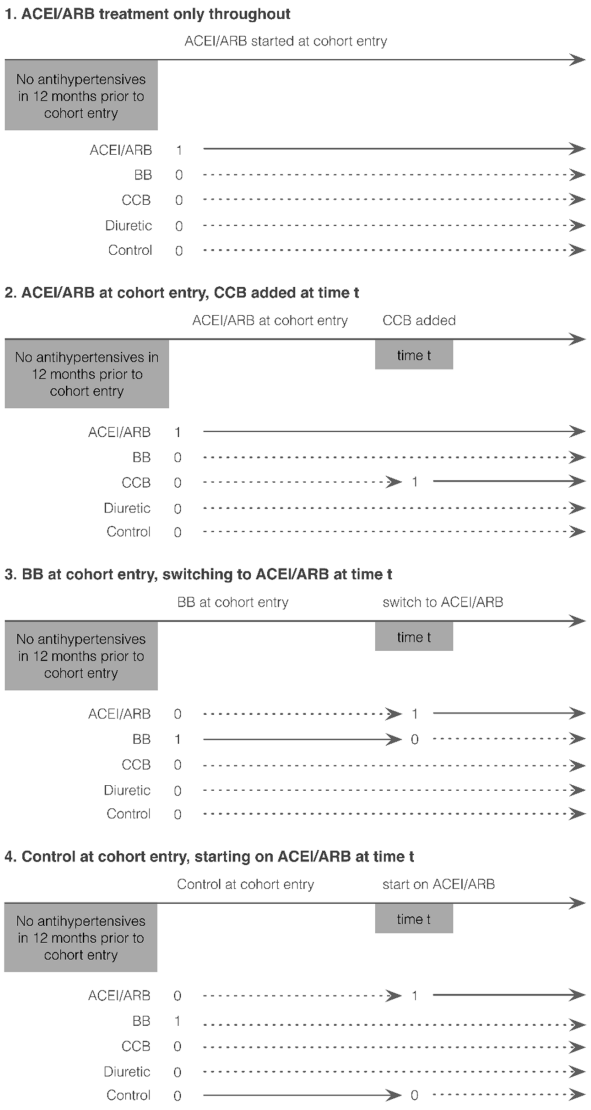
Figure 1. Assignment of time-varying exposure status during follow-up using a single exposure variable under four example scenarios.



9.1.2 Exposure defined by multiple binary indicators

As a secondary analysis we will use an alternative definition for exposure status that will allow exposure to more than one class of antihypertensive at a time. Rather than a single variable representing time exposed to a single class of antihypertensive, we will use five time-updating, binary indicator variables to indicate exposure status. Each indicator variable will identify whether the associated period of time at risk was exposed (1) or unexposed (0) to a specific class of antihypertensive. Figure 2 illustrates the same four example scenarios shown in Figure 1 using multiple binary indicator variables rather than a single exposure variable.

Figure 2. Assignment of time-varying exposure status during follow-up using multiple binary indicator exposure variables under four example scenarios.



9.2 Control group

The control group will be identified as a matched cohort. Controls will be individuals not prescribed one of the exposure drugs they will be matched on age, sex and GP practice, and will enter the cohort on the same day as a matched antihypertensive case. Controls will be matched to all cases entering the cohort who are prescribed one of the classes of antihypertensive to be investigated. We will select ten controls for each case (since we do not know whether potential controls are eligible for study inclusion, we will match a high number of controls to each case to allow for a proportion of matched controls being ineligible for inclusion). We will allow controls to be matched to more than one case (i.e. controls can be selected more than once – in order to maximise the possibility of matching cases that might occur less frequently within the dataset, for example, the very young and the very old).

During follow-up, if a control is started on one of the drugs of interest, their follow-up will be censored (with respect to the control group) from this date and they will enter the exposed cohort.

If antihypertensive users are no longer prescribed any antihypertensives, their follow-up will be censored following a seven-day wash-out period from the end of their prescription. That is, they will not move into the control group because they are likely to be systematically different to the existing control group since there will be a clear clinical reason for withdrawal of all antihypertensive medications (e.g. frailty).

10. Covariates

We aim to use a directed acyclic graph (DAG) to guide the development of covariates to be included in regression models. Examples of the covariates we will use are: ethnicity, socioeconomic status, chronic comorbidities, proteinuria, body mass index (BMI), smoking status, alcohol use, and non-steroidal anti-inflammatory drugs (NSAIDs). We will include age and sex as forced variables.

10.1 Age

Age will be categorised into the following age bands: 18–44, 45–54, 55–59, 60–64, 65–69, 70–74, 75–84, 85+. However, we will also examine the age distribution of the cohort to inform the final the age bands used in the analysis, should a high proportion (e.g. more than 40%) of the cohort fall into only one of the a priori defined age bands will we split this age band more finely (e.g. into five-year rather than 10-year intervals). For descriptive analyses age will be defined as age at cohort entry. For regression analyses age will be entered as a time-updating variable in the age bands defined above (or those informed by the age distribution of the cohort).

10.2 Socioeconomic status

Individual socioeconomic status will be measured using index of multiple deprivation (IMD). CPRD offers index of multiple deprivation as quintile, decile or twentile data for 2004, 2007 or 2010. Our study dates are from 1997 to 2012, since patients can enter and leave the study at any point during the study period, we will use the 2004 IMD data because it is as close to the midpoint of the study period as possible.

10.3 Ethnicity

Ethnicity will be classified according to both Read and ICD-10 coded data to improve data completeness (20). However, research suggests (20) that a large proportion of ethnicity data is missing. We will therefore only rely on ethnicity as a covariate in secondary analyses.

10.4 Chronic comorbidities

The following chronic comorbidities will be considered as covariates: diabetes mellitus, hypertension, ischaemic heart disease, cardiac failure, rhythm disorders, and chronic kidney disease (CKD).

10.4.1 Comorbidities recorded as present or absent

With the exception of CKD, chronic comorbidities included as potential covariates (diabetes mellitus, hypertension, ischaemic heart disease, cardiac failure, and rhythm disorders) will be recorded as present or absent based on recorded Read or ICD-10 codes. For descriptive analyses chronic comorbidities (excluding CKD) will be defined as those recorded prior to cohort entry. For regression analyses chronic comorbidities (excluding CKD) will be entered as a time-updating variables, with disease status changing with the first recorded code for each specific condition.

10.4.2 CKD

CKD stage will be established using estimated glomerular filtration rate (eGFR) calculated using serum creatinine test results. We will not use Read codes or creatinine clearance test results to identify CKD stage (in the main analysis) as adding these measures may compromise the granularity of CKD stage classification (21).

CKD will be categorised as stage two and below, and stages 3–5 based on eGFR levels (stage 2 and below: eGFR >=60; stage 3a: eGFR 45–59; stage 3b: eGFR 30–44; stage 4: eGFR 15–29; and 5: eGFR <15). Patients with no recorded serum creatinine results will be excluded from the main analysis. Estimated GFR will be calculated using the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) equation (22) with serum creatinine measures, age, gender and ethnicity. The CKD-EPI equation contains a variable for Afro-Caribbean ethnicity, however research suggests (20) that a large proportion of ethnicity data is missing. Since the proportion of people of Afro-Caribbean ethnicity in England and Wales is just over 3% (23) for the main analysis we will calculate eGFR without regard to ethnicity.

Serum creatinine measurements were not standardised until 2013, we will therefore assume that all creatinine results are not standardised and multiple results with a correction factor of 0.95 (24) before using the CKD-EPI eGFR formula (unstandardized creatinine results will give a falsely low estimate for GFR).

We will use both baseline and time-updated measures of CKD status:

a. Baseline CKD status

Baseline CKD status will be defined as: i) best of two: the highest eGFR from the most recent two serum creatinine results recorded in the 12 months prior to baseline and separated by a minimum of three months (three month timeframe chosen to correspond to the requirement for eGFR to remain at a consistent level of impairment for at least three months in order for a patient to be diagnosed at a specific CKD stage); or ii) if only one suitable creatinine result is available, the single most recent serum creatinine recorded prior to baseline (excluding patients without two serum creatinine results could systematically exclude healthier individuals, since healthy patients are less likely to serum creatinine levels monitored).

b. Time-updated CKD status

Time-updated CKD status will be defined using the 'last-carried-forward' method (25,26). Here CKD stage is defined based on the most recent creatinine result (allowing CKD stage to be updated over time). However, taking this approach may misclassify AKI episodes as worsening CKD. To avoid this, when establishing CKD status, we will ignore serum creatinine results within 28 days either side of an AKI episode identified by our AKI case definition (see Section 8.2). This approach however, will not remove the risk of misclassifying episodes AKI missed by our definition as declining CKD status. It is

hoped that increased monitoring of renal function in these cases (until serum creatinine stabilises) will minimise the amount of time patients are misclassified.

For descriptive analyses we will define CKD status using the baseline measure described above. For regression analyses we will use time-updated CKD status.

10.5 Proteinuria

Presence of proteinuria will be established using specific morbidity codes (Read and ICD-10) and biochemical test results from primary care electronic health records. Since urinary tract infection can cause a transient proteinuria, proteinuria recorded on the same day as a morbidity code for UTI will be disregarded (i.e. a record of proteinuria on the same day as UTI will not change a patient's proteinuria status to positive). The first valid record of proteinuria will change a patient's status from negative to positive for the remainder of the study.

10.6 Lifestyle

BMI, smoking status and alcohol use will be defined as those recorded prior to cohort entry. BMI will be calculated directly from weight and height records. Smoking status and alcohol use will be categorised based on primary care Read-coded electronic medical records.

10.7 Covariate morbidity code lists

Final morbidity code lists for ethnicity, smoking status, alcohol intake, proteinuria and chronic comorbidities are yet to be developed. A number of existing code lists have already been tested and used (20,21,27–30), and the CALIBER data portal (31) also offers a source of case definition algorithms and code lists. Where code lists/case definition algorithms have already been developed for existing electronic health record studies or as part of the CALIBER project we will identify relevant lists and use them to inform the development of our own code lists. Where existing code lists are unavailable, morbidity codes will be identified by a series of consensus exercises. A list of search terms to identify each relevant covariate will be developed. These search terms will then be used to search a data file containing all the available Read codes. We (LT and KM) will classify the lists of codes returned by the search terms independently as either probably or possibly representing the relevant outcome. The code lists we generate will be compared with each other and any disagreements discussed in order to generate the lists of codes to be used in the project.

We will document the decisions taken regarding code list eligibility, and where necessary we will conduct sensitivity analyses to compare codes that we feel definitely represent the relevant covariate to those that we feel only possibly represent it.

11. Sample size/power calculation

A feasibility count using CPRD (January 2014 build) (see Table 1) shows that: i) for those with a minimum of six months registration prior to their first antihypertensive prescription, there were more than 1,400,000 people with a first prescription for one or more classes of antihypertensive (ACEI/ARB, BB, CCB or diuretic) during the study period (1st April 1997 to 31st March 2012); and ii) for those with a minimum of twelve months prior registration there were more than 1,300,000 with a first prescription for an antihypertensive during the study period. Of those identified in the CPRD database with a first antihypertensive prescription during the study period, 59% were eligible for HES linkage (HES version 9) (n=839,622 patients with a minimum of six months prior registration; and n=795,464 with a minimum of 12 months prior registration).

Our study design allows a patient to move between multiple exposures if their prescription changes over time. Consequently numbers in the different exposure groups will be dynamic. Therefore, we have based our calculation of the minimum effect size that our study can detect on the most conservative sample size, i.e. the smallest group who are prescribed only one class of antihypertensive drugs during the study period, this is the group of individuals prescribed CCBs as their only hypertensive in those with a minimum of 12 months prior registration (n=64,078). If we are cautious and allow for 20% of this sample to be ineligible for inclusion in the study, we are left with a sample size of 102,524 for calculation of the minimum effect size detectable (A previous study (32) within CPRD, where follow-up started after more than 12 months of exposure to an ACEI/ARB, identified 377,649 individuals with a mean duration of follow-up of 4.6 years).

A previous study (8) has estimated the incidence of hospital admission for AKI the adult general population to be 0.7% during median follow-up of 35 months. This translates to a 1.2% probability of AKI assuming an average of 5 years follow-up. This is a conservative estimate of AKI incidence, as not all cases will lead to hospital admission. Based on a cautious estimate of a sample size (n= 102,524), we will have greater than 90% power (alpha 0.05) to detect a relative risk of 1.2 or more for incident AKI comparing each class of antihypertensive with a group of individuals not taking antihypertensives (Table 2). [Calculation done in G*Power version 3.1.9.2 (33), and cross-checked in both OpenEpi (34) using the 'Sample size and power' module for cohort studies, and Stata (35) using the stpower logrank command].

There is only a 5% drop in the number of patients identified using the more stringent requirement for 12 months antihypertensive-free interval prior to study entry compared to only 6 months. Therefore, given that there is sufficient power to detect a relative risk of 1.2 or more, we will use the more robust 12-month (prior to 'first' antihypertensive) definition for cohort entry.

Table 1. Results of feasibility count: Number of patients aged 18 years and over identified with a first prescription for a class of antihypertensive between 1st April 1997 and 31st March 2012 in CPRD (January 2014 build), and, of those identified in CPRD, the number eligible for HES linkage (HES version 9).

		ACEI/ARB	BB	CCB	Diuretic	Total
Minimum 6 months registration prior to first prescription	First prescription in study period	732,514 (52%)	576,740 (41%)	501,394 (35%)	600,759 (42%)	**1,420,953**
	First prescriptions in study period in those eligible for HES linkage	438,344 (52%)	338,518 (40%)	302,806 (36%)	355,078 (42%)	**839,622**
	First class of antiHt drug prescribed in study period	233,151 (28%)	238,785 (28%)	134,185 (16%)	233,501 (28%)	**839,622**
	Prescriptions for only one class of antiHt during the study period	129,850	147,492	66,082	109,046	**452,470**
Minimum 12 months registration prior to first prescription	First prescription in study period	700,839 (52%)	538,173 (40%)	477,549 (35%)	563,526 (42%)	**1,348,019**
	First prescriptions in study period in those eligible for HES linkage	419,343 (53%)	315,587 (40%)	288,419 (36%)	332,724 (42%)	**795,464**
	First class of antiHt drug prescribed in study period	225,555 (28%)	222,187 (28%)	129,030 (16%)	218,692 (27%)	**795,464**
	Prescriptions for only one class of antiHt during the study period	126,827	136,495	64,078	101,976	**429,376**

antiHt: antihypertensive
ACEI: Angiotensin converting enzyme inhibitor
ARB: Angiotensin receptor blocker
BB: Beta-blocker
CCB: Calcium channel blocker

Table 2. Minimum effect sizes given α=0.05, sample size=102,524, power=0.80–0.95

Power	Minimum risk ratio detectable		
	Assuming mean 5-yrs follow-up (base rate 1.2%)	Assuming mean 4-yrs follow-up (base rate 1.0%)	Assuming mean 3-yrs follow-up (base rate 0.7%)
80%	1.17	1.19	1.22
85%	1.18	1.20	1.23
90%	1.19	1.22	1.25
95%	1.22	1.24	1.28

We plan to undertake our main analysis twice by defining exposure as either exposure to a single class of antihypertensive only or using multiple binary indicators for exposure to each class of antihypertensive (see Section 9.1). Defining exposure using multiple binary indicators will allow for individuals to be exposed to more than one class of antihypertensive at a time, maximising follow-up time for each exposure. Therefore, if using the single-agent exposure definition leads to the study being underpowered (due to insufficient follow-up time) we will define exposure using the multiple binary indicator approach only.

12. Analysis

All analyses will be undertaken in Stata version 13 (35). Analysis of the data will include the following stages:

1. Basic baseline descriptive statistics for the five exposure groups:
    i. ACEI/ARB
    ii. BB
    iii. CCB
    iv. Thiazide diuretic
    v. Age, sex and GP practice matched control group
2. Main analysis: Incidence of drug associated AKI in the five exposure groups (crude and adjusted using Poisson regression).
3. Incidence of AKI when other drugs are prescribed in combination with ACEI/ARBs.
4. Comparison of AKI incidence in ACEI users versus that in ARB users.
5. Incidence of AKI when recurrent AKI is taken into account.
6. Outcomes following AKI in the five exposure groups including mortality and ESRD rates.
7. The effect of changes over time.
8. Sensitivity analyses.

12.1 Descriptive statistics

We will present basic descriptive statistics to describe the five exposure groups: i) ACEI/ARB; ii) BB; iii) CCB; iv) thiazide diuretics; and v) an age, sex and GP practice matched control group not prescribed any of the classes of exposure drug. Membership of each exposure group will be determined by any contribution of time at risk to that exposure (e.g. if a patient contributes time at risk to both the ACEI/ARB and BB exposures that patient will contribute to both the ACEI/ARB and BB exposure groups).

For each exposure group we will present the number and percentage of the group: 1) who are female; 2) in each predefined age band (18–44, 45–54, 55–59, 60–64, 65–69, 70–74, 75–84, 85+) at cohort entry; 3) have CKD stage two and below, and stages three to five at cohort entry; 4) have diabetes mellitus, cardiac failure, ischaemic heart disease, arrhythmia, hypertension or proteinuria at cohort entry; 5) in each smoking status category (non-smoker/ex-smoker, current smoker, or missing); 6) in each alcohol use category (non-problem drinker or problem drinker); 7) in each BMI category (underweight, normal, overweight/obese, or missing); 8) in each ethnic group; and 9) in each quintile of index of multiple deprivation.

We will use this information to populate the skeleton table included in Appendix 3 (Table A3.1). No statistical tests are planned at this stage of analysis we will present summary statistics alone.

12.2 Main analysis: Drug associated AKI incidence

For the main analysis, exposure status will be defined in two different ways: i) time exposed to a single class of antihypertensive drug; and ii) exposure defined by multiple binary indicator variables for each class of antihypertensive or control status.

12.2.1 Primary analysis: exposure defined as time exposed to a single class of antihypertensive

Our primary analysis will compare the incidence of AKI in the five exposure groups. Patients will contribute time only when they are prescribed the drug of interest alone (no simultaneous prescriptions for other classes of antihypertensive) and patients will swap between drug exposure groups when a prescription is changed. We will calculate crude AKI incidence rates and adjusted incidence rate ratios for each exposure group. Robust standard errors will be used to account for clustering by practice. Poisson regression will be used to calculate adjusted incidence rate ratios, initially adjusted only for time-updated age (data is split into the following age bands: 18–44, 45–54, 55–64, 65–74, 75–84, 85+) and sex, and then estimated using a fully adjusted model including covariates informed by a directed acyclic graph (DAG). Time-updated age and sex will be included as forced variables, examples of other possible covariates include: proteinuria, chronic comorbidities (CKD stage, diabetes mellitus, ischaemic heart disease, cardiac failure, hypertension and arrhythmia), and baseline smoking, alcohol intake, socioeconomic and BMI status. Rate ratios will be calculated using the control group as the reference category. Before fitting the fully adjusted model we will add each covariate independently to a Poisson regression model including the exposure, and age and sex.

The results of this analysis will be used to populate skeleton Tables A3.2 and A3.3 included in Appendix 3.

The exposure definition we have chosen for this analysis (where time at risk will be counted only when an individual is prescribed a single class of antihypertensive) allows for easier interpretation of a regression model because the coefficients returned by the model offer a direct comparison between the different types of exposure. However, it will reduce available follow-up time as we will not be able to include time when a patient is prescribed more than one class of antihypertensive agent. Therefore, we will repeat the main analysis using an alternative definition of exposure status.

12.2.2 Secondary analysis: exposure status defined by binary indicators

We will repeat the primary analysis defining exposure status using five time-updating, binary indicator variables to indicate exposure. The approach will maximize the available follow-up time for each exposure, control for exposure to other antihypertensives, allow drug combinations to be investigated through interaction terms in the model, and more closely model real life.

We would expect results from the two analyses to be broadly similar. However, rate ratios for exposure in the single-drug class exposure model will compare rates for each class of drug to incidence in the control group, while rate ratios for exposure in the multiple binary indicators model will compare rates of AKI in each anti-hypertensive class to those not exposed to that class of drug controlled for other classes of antihypertensive.

12.3 Combination prescriptions

There is evidence that combinations of ACEI/ARBs, diuretics and NSAIDs may impair renal function (9–12). We will therefore investigate combination drug therapy in ACEI/ARB users using time-updating exposure status. Time at risk will be defined in the following categories:

i. ACEI/ARB alone
ii. ACEI/ARB + thiazide diuretic
iii. ACEI/ARB + loop diuretic
iv. ACEI/ARB + loop diuretic + thiazide diuretic
v. ACEI/ARB + loop diuretic + potassium-sparing diuretic (spironolactone/eplerenone/amioloride/triamterene)
vi. ACEI/ARB + loop diuretic + potassium-sparing diuretic + thiazide diuretic
vii. ACEI/ARB + NSAID +/- any BB, CCB or diuretic.

Combination drugs (e.g. valsartan/hydrochlorothiazide, a combination of an ARB and a thiazide diuretic) will be considered as dual exposure to the classes of drug included in the preparation.

We will investigate how AKI incidence rate changes in ACEI/ARB users who are also prescribed additional medications by calculating crude and adjusted (for all covariates used in the main analysis) AKI incidence rate ratios comparing exposure to ACEI/ARBs alone (reference category) with exposure to ACEI/ARBs in addition to other drugs.

These data will be used to populate a skeleton table included in Appendix 3 (Table A3.4).

12.4 ACEI versus ARB

We will investigate AKI incidence rate in ACEI users versus that in ARB users. We will repeat the main analysis this time defining exposure as time at risk to ACEIs alone compared to ARBs alone. We will use these data to populate the skeleton table presented in Appendix 3 (Table A3.5).

12.5 Recurrent AKI

Our main analysis uses first recorded episode of AKI as the outcome measure. Patients therefore stop contributing time at risk at their first episode of AKI. To investigate the effect of multiple episodes of AKI we will repeat the main analysis including recurrent episodes of AKI. We will account for clustering in the analysis using a random effects model.

We will define recurrent AKI differently depending on whether AKI has been defined using morbidity coding or biochemistry results. Two or more successive AKI codes will be considered to represent recurrent episodes of AKI. However, if AKI is defined using biochemical data, to be classified as being a recurrent episode there would have to be a return to baseline creatinine before a second increase in serum creatinine.

12.6 Outcomes following AKI

We will investigate outcomes following AKI by investigating rates of mortality, ESRD and hyperkalaemia among those who develop AKI. We will compare rates of mortality, ESRD and hyperkalaemia in AKI patients from each of the five exposure groups in order to compare health outcomes following AKI in those on different antihypertensives compared to the control group. Exposure will be defined based on exposure status at time of an AKI event, for example, an individual will be classified as being exposed to a beta-blocker if they are prescribed a beta-blocker when they experience an episode of AKI.

We will investigate mortality following AKI episodes by calculating crude mortality rates following AKI episodes. We will also calculate crude, age and sex adjusted, and fully adjusted mortality rate ratios

for overall mortality and mortality at 0–3 months following the AKI episode, 4–6 months and 7–12 months. These data will be used to populate the skeleton table presented in Appendix 3 (Table A3.6).

12.7 Changes over time

There have been some important changes in diagnostic and administrative practices that may influence the number of cases our AKI definition identifies and the classification of CKD. These include:

i. 2004: Publication of RIFLE (36) AKI definition
ii. 2006: Introduction of standardised serum creatinine measurements.
iii. 2007: Standardised lab reporting (Pathology Messaging Implementation Project (37)) and publication of AKIN (38) AKI definition.
iv. 2012: Publication of KDIGO (19) AKI definition.

We will therefore investigate changes in AKI incidence rates over time using the following epochs: i) before 2004; ii) 2004–2005; iii) 2006–2007; iv) 2007–2011; and, v) from 2012. We will repeat the main analysis including calendar period as a covariate (after splitting the data on the calendar periods defined above).

12.8 Sensitivity analyses

We will test the validity of some of the variable definitions used in the analysis by repeating the main analysis a number of times either in select patient populations or using alternative variable definitions.

12.8.1 Person-time

The analysis will be repeated using an alternative approach to calculating person-time. In the main analysis person-time will be calculated from cohort entry to cohort exit, without taking hospital admission time into account. Our AKI outcome definition uses hospital coding, therefore, for the duration of a hospital admission, we can only define the outcome once, effectively reducing available time at risk. To check whether this influences our findings we will repeat the main analysis calculating person-time from cohort entry to cohort exit with any hospital admission time excluded from person-time.

12.8.2 Exposure status

The analysis will also be repeated defining exposure status on the basis of two or more consecutive prescriptions to investigate a subgroup with a more reliable exposure to the drugs interest.

12.8.3 Ethnicity

We will also repeat the main analysis in a group who are more likely to have complete ethnicity data. After 2006 recording of ethnicity was rewarded as part of the Quality and Outcomes Framework leading to improvements in the completeness of ethnicity recording in CPRD (20). We will therefore repeat the main analysis in new entrants to the cohort from 2006 onwards who have ethnicity data recorded in CPRD or HES. In this sensitivity analysis we will include ethnicity both as a covariate and in the equation used to calculate eGFR.

12.8.4 Renal function

Our analyses rely on serum creatinine test results both to establish biochemically defined AKI and as a measure of CKD status. We plan a number of sensitivity analyses to test the validity of both variable definitions.

a. Limited numbers of serum creatinine results

i) Limited to patients with diabetes mellitus

Patients with limited numbers of serum creatinine measures available may have their CKD status misclassified or be incorrectly identified as having biochemical AKI. Further, they are likely to be systematically different to those with multiple serum creatinine results (since regular renal function testing is more likely in those perceived to be at risk of kidney disease). Therefore, to test the validity of both our AKI and our CKD definitions we will repeat the main analysis only in those who are also recorded as having diabetes mellitus. Regular checking of renal function has been remunerated in diabetics through the Quality and Outcomes Framework. Using this group will therefore reduce information bias occurring as a result of kidney function being measured only in those perceived to be at risk of the outcome.

ii) Two serum creatinine results before index-infection code in biochemically defined AKI

We will further test the validity of our AKI definition by repeating the main analysis including only those biochemically defined AKI episodes where there is a minimum of two serum creatinine results (recorded at least three months apart) available prior to the recording of the index acute morbidity code that defines the AKI episode (UTI, gastroenteritis, URTI – see Section 8.2) in order to establish a more robust baseline serum creatinine.

b. CKD status definition

i) CKD status defined using morbidity coding and test results

Our main analysis requires a minimum of one available serum creatinine result to establish CKD status. This will limit the number of patients eligible for inclusion and is also likely to result in an unusual control group since it is unlikely that routine serum creatinine measures will be available in healthy controls. We will therefore repeat the main analysis, without the requirement for a serum creatinine result, using an alternative CKD definition. CKD will be defined as present or absent on the basis of: i) eGFR calculated using serum creatinine results (see Section 10.4.2); ii) morbidity codes for CKD; and iii) intrinsic renal disease codes (e.g. glomerulonephritis).

ii) Baseline CKD status rather than time-updated CKD status

We will also repeat the main analysis using baseline CKD status rather than the time-updated variable (see Section 9.4.2) to investigate: i) the association between CKD stage at initial prescription of antihypertensive medication and subsequent risk of AKI; and ii) to ensure that rapidly worsening CKD status has not affected the results of the analysis – AKI episodes in patients with rapidly worsening CKD status may lead to a falsely high rates of AKI at more severe stages of CKD (patients with rapidly worsening renal function will contribute less person-time at more severe levels of CKD than patients who had maintained consistently levels of CKD, therefore, episodes of AKI occurring in patients with rapidly worsening renal function would lead to a falsely high rate of AKI at higher stage of CKD).

c. Transient serum creatinine increases not representing renal disease

We may also misclassify CKD status or incorrectly identify biochemical AKI in two specific clinical scenarios that may result in a raised serum creatinine that does not represent renal disease (trimethoprim prescription or ACEI/ARB initiation). We will therefore repeat the main analysis in the following two sensitivity analyses:

i. Excluding the first recorded serum creatinine result after starting an ACEI/ARB from the algorithm used to identify biochemical AKI – Since ACEI/ARB initiation results in an acute increase in serum creatinine (39).
ii. Excluding any serum creatinine results in the two weeks following trimethoprim prescription from AKI/CKD definitions – Since trimethoprim also temporarily increases serum creatinine (40).

13. Missing data

Patients on antihypertensive drugs are likely to have other cardiovascular risk factors considered when their medications are prescribed; consequently, we anticipate the proportion of completeness to be high in this population. We will therefore undertake a complete case analysis unless missing data is greater than 30% when we will undertake further sensitivity analyses (in addition to those discussed in Section 12.8 above). For example, if necessary, we will repeat the main analysis restricting it to more recent calendar periods when BMI, alcohol and smoking data is more complete (41), in order to reduce any selection biases due to data being missing.

14. Limitations of study design, data sources and analytic methods

The main limitation in this project is one common to all studies using electronic medical records. It is the problem of accurately defining measures for outcomes, exposures and covariates. Coding may be a reflection of individual GPs' diagnostic beliefs and the patterns and context of their coding behaviour (42,43). Variation in coding practices will impact on the reliability of the definitions we use to identify outcomes, exposures and covariates in our study. However, research (44) suggests that most diagnoses within GPRD (CPRD) are recorded accurately, and, further research suggests that there have been improvements in data quality in the domains assessed by the Quality and Outcomes Framework (QOF) (45,46). We hope that any misclassification due to variability in coding practices will be mitigated by careful development of code lists and, where possible, use of previously validated code lists. In relation to our primary outcome measure, previous research has (17) has shown a high agreement between ICD-10 coding for AKI and a clinical AKI diagnosis.

The first guidelines for the diagnosis of AKI date from 2004, we might therefore expect changes in AKI recording as awareness of the diagnosis increases. Any observed increases in AKI incidence over time might therefore be attributed to changes in diagnostic awareness rather than real changes in AKI incidence. This must be acknowledged when interpreting our results. However, we will use biochemical data in addition to coding data where this temporal change should not contribute to an apparent increase in incidence.

In order to establish a more robust measure for CKD status with the ability to classify CKD into stages we have chosen to limit the study population to only those with serum creatinine measures available. Further, in those with no serum creatinine result available in the 12 months prior to cohort entry, we will only include follow-up time after the first serum creatinine result available following cohort entry. This will limit the number of patients eligible for inclusion, reduce follow-up time, and also result in select group of controls (since it is unlikely that routine serum creatinine measures will be available in healthy controls). However, it is assumed that most patients will have their renal function tested prior

to initiation of antihypertensive therapy limiting the reduction of follow-up time for cases. In addition, we plan to repeat the main analysis, without the requirement for those included to have a serum creatinine result recorded, using an alternative CKD definition that uses both biochemical test results and morbidity coding (see Section 12.8.b.i).

The ability of both our biochemical AKI definition and our CKD algorithm to reliably classify renal disease is likely to limited by the frequency of serum creatinine measures. However, we plan a number of sensitivity analyses (Section 12.8.4) to account for differences in the frequency and timing of serum creatinine measures between patients. In relation to the biochemical AKI definition specifically, since we do not have access to hospital test results, we have modified the ACB AKI e-alert algorithm (18) to account for differences in primary care biochemical data (compared to hospital lab data) and included AKI morbidity coding as part of the outcome definition.

Missing ethnicity data may limit the findings of the study. Ethnicity has been shown to be related to CKD risk (47) and should therefore should be considered as a covariate. However, research has shown that a high proportion of ethnicity data in incomplete (20), therefore rather than reducing sample size by excluding those with incomplete ethnicity data from the main analysis we will only rely on ethnicity as a covariate in a sensitivity analysis to check the validity of our findings (see Section 12.8.3). In addition, calculation of eGFR requires a variable for Afro-Caribbean ethnicity. Since census data shows the proportion of people of Afro-Caribbean ethnicity in England and Wales to be just over 3% (23) we plan to calculate eGFR without regard to ethnicity for the main analysis and check the validity of our findings in the sensitivity analysis presented in Section 12.8.3.

Undertaking the study in HES linked practices only will reduce the size of the study cohort. However, a feasibility count has shown that, while using linked practices reduces the number of patients in the cohort by 41%, there are still over 795,000 individuals with a new prescription for an antihypertensive during the study period. By using an analysis that limits follow-up time to time exposed to a single class of antihypertensive only may reduce follow-up time further and may limit the generalizability of the findings (individuals on single therapy are likely to have less morbidity than those on multiple antihypertensives). However, we hope to mitigate this by repeating the main analysis using an alternative definition for exposure that allows for multiple antihypertensive usage and maximises follow-up time (Section 12.2.2).

We aim to reduce confounding by assessing and adjusting for a covariates informed by development of a DAG. We hope to reduce confounding by indication by using prescriptions for other classes of antihypertensive agents as control groups. However, slight differences in the indications for different classes of antihypertensives may result in some degree of confounding by indication. It is hoped that this will be limited by controlling for a number of chronic comorbidities that are indications for these drugs.

There is some concern regarding the number of comparisons being made in some of our secondary analyses (particularly our analysis of combination prescriptions in ACEI/ARB users – Section 12.3). Our study will focus on the primary objective (to examine the association between ACEI/ARBs and AKI) and this will be given prominence in the write-up. We will interpret with caution the results of secondary analyses where high numbers of comparisons are made.

15. Dissemination

15.1 Patient or user group involvement

We plan to share our findings with patient/user groups via Kidney Research UK. We aim to develop some materials to communicate the balance of risks and benefits regarding the use of specific medications for patients with CKD.

15.2 Disseminating and communicating study results

The study findings will be submitted for publication in peer-reviewed scientific journals, and will be presented both at appropriate conferences and at other meetings; the latter will include scientific meetings externally, for example, The European Renal Association and European Dialysis and Transplant Association Congress, The American Society of Nephrology Kidney Week and internally within the London School of Hygiene and Tropical Medicine.

References

1. Wang HE, Muntner P, Chertow GM, Warnock DG. Acute kidney injury and mortality in hospitalized patients. Am J Nephrol [Internet]. 2012;35(4):349–55.
2. Bedford M, Stevens PE, Wheeler TWK, Farmer CKT. What is the real impact of acute kidney injury? BMC Nephrol [Internet]. 2014;15(1):95.
3. Chertow GM, Burdick E, Honour M, Bonventre J V, Bates DW. Acute kidney injury, mortality, length of stay, and costs in hospitalized patients. J Am Soc Nephrol. 2005;16(11):3365–70.
4. Kerr M, Bedford M, Matthews B, O'Donoghue D. The economic impact of acute kidney injury in England. Nephrol Dial Transplant [Internet]. 2014;29(7):1362–8.
5. Uchino S, Bellomo R, Goldsmith D, Bates S, Ronco C. An assessment of the RIFLE criteria for acute renal failure in hospitalized patients. Crit Care Med. 2006;34(7):1913–7
6. Pannu N, James M, Hemmelgarn BR, Dong J, Tonelli M, Klarenbach S. Modification of outcomes after acute kidney injury by the presence of CKD. American Journal of Kidney Diseases. Am J Kidney Dis. 2011;58(2):206–13.
7. Ali T, Khan I, Simpson W, Prescott G, Townend J, Smith W, et al. Incidence and outcomes in acute kidney injury: a comprehensive population-based study. J Am Soc Nephrol. 2007;18(4):1292–8
8. James MT, Hemmelgarn BR, Wiebe N, Pannu N, Manns BJ, Klarenbach SW, et al. Glomerular filtration rate, proteinuria, and the incidence and consequences of acute kidney injury: a cohort study. Lancet;2010;376(9758):2096–103.
9. Adhiyaman V, Asghar M, Oke a, White a D, Shah IU. Nephrotoxicity in the elderly due to co-prescription of angiotensin converting enzyme inhibitors and nonsteroidal anti-inflammatory drugs. J R Soc Med. 2001;94(10):512–4.
10. Loboz KK, Shenfield GM. Drug combinations and impaired renal function – the " triple whammy ." B J Pharmacol. 2004;59(2):239–43.
11. Fournier J-P, Sommet A, Durrieu G, Poutrain J-C, Lapeyre-Mestre M, Montastruc J-L. Drug interactions between antihypertensive drugs and non-steroidal anti-inflammatory agents: a descriptive study using the French Pharmacovigilance database. Fundam Clin Pharmacol. 2012;28(2):230–5.
12. Lapi F, Azoulay L, Yin H, Nessim SJ, Suissa S. Concurrent use of diuretics, angiotensin converting enzyme inhibitors, and angiotensin receptor blockers with non-steroidal anti-inflammatory drugs and risk of acute kidney injury: nested case-control study. Br Med J. 2013;346:8525.
13. Arora P, Rajagopalam S, Ranjan R, Kolli H, Singh M, Venuto R, et al. Preoperative use of angiotensin-converting enzyme inhibitors/angiotensin receptor blockers is associated with increased risk for acute kidney injury after cardiovascular surgery. Clin J Am Soc Nephrol. 2008;3(5):1266–73.
14. Tomlinson LA, Abel GA, Chaudhry AN, Tomson CR, Wilkinson IB, Roland MO, et al. ACE inhibitor and angiotensin receptor-II antagonist prescribing and hospital admissions with acute kidney injury: a longitudinal ecological study. PLoS One. 2013;8(11):e78465.
15. National Institute for Health and Care Excellence. Clinical guideline 169. Acute kidney injury. London: NICE;2013.
16. Feehally J, Gilmore I, Barasi S, Bosomworth M, Christie B, Davies A, et al. RCPE UK Consensus Conference Statement Management of acute kidney injury: the role of fluids, e-alerts and biomarkers. JR Coll Physicians. 2013;43:37–8.
17. Tomlinson LA, Riding AM, Payne R a, Abel G a, Tomson CR, Wilkinson IB, et al. The accuracy of diagnostic coding for acute kidney injury in England - a single centre study. BMC Nephrol;2013;14(1):58.
18. Association for Clinical Biochemistry and Laboratory Medicine. Algorithm for generating E-Alerts for Acute Kidney Injury based on serum creatinine changes with time [online]. 2013. Available from: http://www.acb.org.uk/docs/appendix-a-algorithm [Accessed 2014 Aug 21].
19. KDIGO Work Group. KDIGO Clinical Practice Guideline for Acute Kidney Injury. Kidney Int. 2012;2(1):1–138.

20. Mathur R, Bhaskaran K, Chaturvedi N, Leon D a, Vanstaa T, Grundy E, et al. Completeness and usability of ethnicity data in UK-based primary care and hospital databases. J Public Health (Oxf). 2013;1–9.

21. McDonald H. The effect of renal disease on acute infections among older people with diabetes. PhD thesis in progress. London School of Hygiene and Tropical Medicine; Forthcoming 2015.

22. Levey AS, Stevens LA, Schmid CH, Zhang Y (Lucy), Alejandro F. Castro I, Feldman HI, et al. A new equation to estimate glomerular filtration rate. Ann Intern Med. 2009;150(9):604–12.

23. Office for National Statistics. Census: Key Statistics for local authorities in England and Wales, table number KS201EW [online]. 2012. Available from: www.ons.gov.uk/Fons/Frel/Fcensus/F2011-census/Fkey-statistics-for-local-authorities-in-england-and-wales/Frft-table-ks201ew.xls [Accessed 3rd Sept 2014].

24. Fox CS, Matsushita K, Woodward M, Bilo HJ, Chalmers J, Heerspink HJL, et al. Associations of kidney disease measures with mortality and end-stage renal disease in individuals with and without diabetes: a meta-analysis. Lancet; 2012;380(9854):1662–73.

25. James M, Laupland K, Tonelli M, Manns BJ, Culleton B, Hemmelgarn BR. Risk of bloodstream infection in patients with chronic kidney disease not treated with dialysis. Arch Intern Med. 2008;168(21):2333–9.

26. James MT, Quan H, Tonelli M, Manns BJ, Faris P, Laupland KB, et al. CKD and risk of hospitalization and death with pneumonia. Am J Kidney Dis. 2009;54(1):24–32.

27. Bhaskaran K, Douglas I, Forbes H, Dos-Santos-Silva I, Leon DA, Smeeth L. Body-mass index and risk of 22 specific cancers: a population-based cohort study of 5·24 million UK adults. Lancet. 2014;6736(14):60892–8.

28. McDonald HI, Nitsch D, Millett ERC, Sinclair A, Thomas SL. New estimates of the burden of acute community-acquired infections among older people with diabetes mellitus: a retrospective cohort study using linked electronic health records. Diabet Med. 2014;31(5):606–14.

29. Quint JK, Herrett E, Bhaskaran K, Timmis a, Hemingway H, Wedzicha J a, et al. Effect of β blockers on mortality after myocardial infarction in adults with COPD: population based cohort study of UK electronic healthcare records. BMJ. 2013;347:f6650.

30. Brauer R, Smeeth L, Anaya-Izquierdo K, Timmis A, Denaxas SC, Farrington CP, et al. Antipsychotic drugs and risks of myocardial infarction: a self-controlled case series study. Eur Heart J. 2014;ehu263.

31. Denaxas SC, George J, Herrett E, Shah AD, Kalra D, Hingorani AD, et al. Data resource profile: cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER). Int J Epidemiol. 2012 Dec;41(6):1625–38.

32. Bhaskaran K, Douglas I, Evans S, van Staa T, Smeeth L. Angiotensin receptor blockers and risk of cancer: cohort study among people receiving antihypertensive drugs in UK General Practice Research Database. BMJ. 2012;344:e2697–e2697.

33. Faul F, Erdfelder E, Lang A-G, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods. 2007;39:175–91.

34. Dean AG, Sullivan KM, Soe MM. OpenEpi: Open Source Epidemiologic Statistics for Public Health, Version 3.0 [Internet]. Available from: www.OpenEpi.com

35. StataCorp. Stata Statistical Software: Release 13. College Station, TX: StataCorp LP; 2013.

36. Bellomo R, Ronco C, Kellum J a, Mehta RL, Palevsky P. Acute renal failure - definition, outcome measures, animal models, fluid therapy and information technology needs: the Second International Consensus Conference of the Acute Dialysis Quality Initiative (ADQI) Group. Crit Care. 2004;8(4):R204–12.

37. NHS Direct. Pathology Messaging Implementation Project Weekly Statistical Report [online]. 2007. Available from: http://webarchive.nationalarchives.gov.uk/20140530120528/http://www.connectingforhealth.nhs.uk/systemsandservices/pathology/edifact/pmip/reports/pmipstatus.pdf [Accessed 26th Sept 2014].

38. Mehta RL, Kellum J a, Shah S V, Molitoris B a, Ronco C, Warnock DG, et al. Acute Kidney Injury Network: report of an initiative to improve outcomes in acute kidney injury. Crit Care. 2007;11(2):R31.

39. Ahmed A. Use of angiotensin-converting enzyme inhibitors in patients with heart failure and renal insufficiency: how concerned should we be by the rise in serum creatinine? J Am Geriatr Soc. 2002;50(7):1297–300.

40. Kastrup J, Petersen P, Bartram R, Hansen JM. The effect of trimethoprim on serum creatinine. Br J Urol. 1985;57(3):265–8.

41. Vinogradova Y, Coupland C, Hippisley-Cox J. Exposure to bisphosphonates and risk of gastrointestinal cancers: series of nested case-control studies with QResearch and CPRD data. BMJ.. 2013;346:f114.

42. Peat G, Greig J, Wood L, Wilkie R, Thomas E, Croft P. Diagnostic discordance: We cannot agree when to call knee pain "osteoarthritis."Fam Pract. 2005;22:96–102.

42. Pearson N, O'Brien J, Thomas H, Ewings P, Gallier L, Bussey A. Collecting morbidity data in general practice: the Somerset morbidity project. BMJ. 1996;312:1517–20.

43. Khan NF, Harrison SE, Rose PW. Validity of diagnostic coding within the General Practice Research Database: a systematic review. Br J Gen Pract. 2010;60(572):e128–36

44. Williams PH, de Lusignan S. Does a higher "quality points" score mean better care in stroke? An audit of general practice medical records. Inform Prim Care. 2006;14:29–40.

45. Doran T, Kontopantelis E, Valderas JM, Campbell S, Roland M, Salisbury C, et al. Effect of financial incentives on incentivised and non-incentivised clinical activities: longitudinal analysis of data from the UK Quality and Outcomes Framework. BMJ. 2011;342:d3590.

46. Barbour SJ, Schachter M, Er L, Djurdjev O, Levin A. A systematic review of ethnic differences in the rate of renal progression in CKD patients. Nephrol Dial Transplant. 2010;25(8):2422–30.

Glossary of acronyms

ACEI	Angiotensin converting enzyme inhibitor
AKI	Acute kidney injury
AKIN	Acute Kidney Injury Network: the group to produce the AKIN AKI criteria (38)
ARB	Angiotensin receptor blocker
BB	Beta-blocker
BNF	British national formulary
CCB	Calcium channel blocker
CKD	Chronic kidney disease
CPRD	Clinical Practice Research Datalink
DAG	Directed acyclic graph
DM	Diabetes mellitus
eGFR	Estimated glomerular filtration rate
ESRD	End stage renal disease
HES	Hospital episode statistics
IMD	Index of multiple deprivation
KDIGO	Kidney disease improving global outcomes
NSAID	Non-steroidal anti-inflammatory drug
RIFLE	The AKI criteria produced by Acute Dialysis Quality Initiative (36): Risk, Injury, Failure, Loss and End stage renal disease.
SCr	Serum creatinine

Appendix 1 – AKI definitions

Table A1.1 The staging of acute kidney injury in adults[1] – comparing RIFLE, AKIN and KDIGO (adapted from NICE 2013 acute kidney injury clinical guidelines)(15).

Stage	RIFLE(36)[2] serum creatinine criteria	AKIN(38) serum creatinine criteria	KDIGO(19) serum creatinine criteria	Urine output
RIFLE Risk or AKIN/KDIGO 1	eGFR decrease by ≥ 25% OR 50–99% SCr rise from baseline* (1.50–1.99 x baseline)	Rise of ≥26µmol/L within 48 hours OR 50–99% SCr rise from baseline* (1.50–1.99 x baseline)	Rise of ≥26µmol/L within 48 hours OR 50–99% SCr rise from baseline* (1.50–1.99 x baseline)	< 0.5 ml/kg/h for more than 6h
RIFLE Injury or AKIN/KDIGO 2	eGFR decrease by ≥ 50% OR 100–199% SCr rise from baseline* (2.00–2.99 x baseline)	100–199% SCr rise from baseline* (2.00–2.99 x baseline)	100–199% SCr rise from baseline* (2.00–2.99 x baseline)	< 0.5 ml/kg/h for more than 12h
RIFLE Failure or AKIN/KDIGO 3	eGFR decrease by ≥ 75% OR ≥ 200% SCr rise from baseline* (≥ 3.00 x baseline) OR SCr rise to ≥354µmol/L with acute rise of 44µmol/L	≥ 200% SCr rise from baseline* (≥ 3.00 x baseline) OR SCr rise to ≥354µmol/L with acute rise of 44µmol/L any requirement for renal replacement therapy	≥ 200% SCr rise from baseline* (≥ 3.00 x baseline) OR SCr rise to ≥354µmol/L with acute rise of: ≥ 26µmol/L within 48hrs or ≥ 50% rise within 7 days any requirement for renal replacement therapy	< 0.3 ml/kg/h for 24h or anuria for 12h

1. The initial diagnosis or detection of AKI is based on a patient meeting any of the criteria for stage 1. Staging is carried out retrospectively when the episode is complete. Patients are classified according to the highest possible stage where the criterion is met, either by creatinine rise or urine output.
2. RIFLE: for simplicity the Loss and End stage categories of RIFLE are not included here (these can be regarded as clinical outcomes rather than AKI stages).

SCr - Serum creatinine.

*Increase from baseline serum creatinine is either known (based on a prior blood test) or presumed (based on the patient history) to have occurred within 7 days.

Appendix 2 – Code lists

Table A2.1 Search terms to identify morbidity codes which may represent AKI.

Condition	Symptoms	Tests	Procedures
kidney *renal* *tubular* *nep* to pick up nephritis +- nephropathy etc. *glomerulo*	*uria* to pick up oliguria, anuria, etc. *urine* to pick up reduced urine output	*creatinine* *hyperkalaemia* *electrolyte*	*dialysis* *haemofiltration*

* represents wildcard operator.

Table A2.2 Search strings to identify codes from CPRD code browsers/code data files.

Variable	Search string to identify relevant medcode/prodcode
AKI	*kidney*;*renal*;*tubular*;*nep*;*glomerulo*;*uria*;*urine*;*creatinine*;*hyperkalaemia*; *electrolye*;*dialysis*;*haemofiltration*
ACEI	*captopril*;*cilazapril*;*enalapril*;*fosinopril*;*imidapril*;*lisinopril*; *moexipril*;*perindopril*;*quinapril*;*ramipril*;*trandolapril*; *capoten*;*noyada*;cozidocapt*;*capozide*;*vascace*;*innovace*;*innozide*;*tanatril*;*zestril*;*carace*; *zestoretic*;*perdix*;*coversyl*;*accupro*;*accuretic*;*tritace*;*triapin*
ARB	*azilsartan*;*candesartan*;*eprosartan*;*irbesartan*;*losartan*;*olmesartan*;*telmisartan*;*valsartan*; *edarbi*;*amias*;teveten*;*aprovel*;*coaprovel*;*cozaar*;*olmetec*; *sevikar*;*micardis*;*diovan*;*exforge*
Beta-blockers	*propranolol*;*acebutolol*;*atenolol*;*bisoprolol*;*carvedilol*;*celiprolol*; *co-tenidone*;*esmolol*;*labetalol*;*metoprolol*;*nadolol*;*nebivolol*;*oxprenolol*; *pindolol*;*sotalol*;*timolol*; *bedranol*;*beta prograne*;*sectral*;*tenormin*;*cardicor*;*carvedilol*;*celectol*;*tenoret*; *tenoretic*;*brevibloc*;*trandate*;*betaloc*;*lopresor*;*corgard*;*nebilet*; *slow-trasicor*;*visken*;*beta-cardone*;*sotacor* with diuretic: *kalten*;*viskaldix*;*prestim* with CCB: *beta-adalat*;*tenif*
Calcium channel blockers	*amlodipine*;*diltiazem*;*felodipine*;*lacidipine*;*lercanidipine*;*nicardipine*;*nifedipine*; *nimodipine*;*verapamil*;*istin*;*exforge*;*diltiazem*;*tildiem*;*adizem*;*angitil*;*dilcardia*; *dilzem*;*slozem*;*viazem*;*zemtard*;*plendil*;*triapin*;*motens*;*zanidip*;*cardene*;*adalat*; *adipine*;*coracten*;*fortipine*;*nifedipress*;*tensipine*;*valni*;*tenif*;*nimotop*;*cordilox*; *securon*;*univer*;*verapress*;*vertab*
Thiazide diuretics	*bendroflumethiazide*;*chlortalidone*;*cyclopenthiazide*;*indapamide*;*metolazone*; *xipamide*; *aprinox*;*neo-naclex*;*hygroton*;*navidrex*;*natrilix*;*ethibide*;*tensaid*;*diurexan*
Loop diuretics	*bumetanide*;*furosemide*;*torasemide*;*frusemide*; *lasix*;*torem*
K+ sparing diuretics	*amiloride*;*triamterene*;*eplerenone*;*spironolactone*; *inspra*;*aldactone* with other diuretics: *co-amilofruse*;*frumil*;*co-amilozide*;*moduret*;*moduretic*;*frusene*
Other diuretics	*acetazolamide*;*mannitol*
NSAIDs	*aceclofenac*;*acemetacin*;*celecoxib*;*dexibuprofen*;*dexketoprofen*;*diclofenac*; *etodolac*;*etoricoxib*;*fenoprofen*;*flurbiprofen*;*ibuprofen*;*indometacin*; *ketoprofen*;*mefenamic*;*meloxicam*;*nabumetone*;*naproxen*;*piroxicam*;*sulindac*; *tenoxicam*; *tiaprofenic*;*aspirin*; *phenylbutazone*;*ketorolac*;*parecoxib*;*tolfenamic*

Table A2.3 Read codes for AKI.

Read code	Read term	Definite*
K04..00	Acute renal failure	✓
K04..11	ARF - Acute renal failure	✓
K04..12	Acute kidney injury	✓
K040.00	Acute renal tubular necrosis	✓
K040.11	ATN - Acute tubular necrosis	✓
K04y.00	Other acute renal failure	✓
K04z.00	Acute renal failure NOS	✓
K0E..00	Acute-on-chronic renal failure	✓
Kyu2000	[X]Other acute renal failure	✓
1AC1.00	Oliguria	x
8H2M.00	Admit renal medicine emergency	x
K043.00	Acute drug-induced renal failure	x
K043000	Acute renal failure due to ACE inhibitor	x
K043400	Acute renal failure induced by non-steroid anti-inflamm drug	x
K044.00	Acute renal failure due to urinary obstruction	x
K045.00	Acute renal failure due to non-traumatic rhabdomyolysis	x
K04B.00	Acute renal failure due to traumatic rhabdomyolysis	x
K0C1.00	Nephropathy induced by other drugs meds and biologl substncs	x
K0C2.00	Nephropathy induced by unspec drug medicament or biol subs	x
R085000	[D]Oliguria	x
R085z00	[D]Oliguria and anuria NOS	x
SK08.00	Acute renal failure due to rhabdomyolysis	x
SP15400	Renal failure as a complication of care	x
SP15411	Kidney failure as a complication of care	x
SP15412	Post operative renal failure	x

*Codes that possibly represent AKI (i.e. not definite AKI codes) will be used in a sensitivity analysis to test the validity of the AKI definition – the main analysis will be repeated using both possible and definition AKI codes as an outcome definition.

Table A2.4 ICD-10 codes for AKI

ICD-10 code	ICD-10 term	Definite
N17*	Acute renal failure	✓
N17.0	Acute renal failure with tubular necrosis	✓
N17.8	Other acute renal failure	✓
N17.9	Acute renal failure, unspecified	✓
N14	Drug- and heavy-metal-induced tubulo-interstitial and tubular conditions	x
N14.1	Nephropathy induced by other drugs, medicaments and biological substances	x
N14.2	Nephropathy induced by unspecified drug, medicament or biological substance	x
N17.1	Acute renal failure with acute cortical necrosis	x
N17.2	Acute renal failure with medullary necrosis	x
N19	Unspecified kidney failure	x
N99.0	Postprocedural renal failure	x
R34	Anuria and oliguria	x
R94.4	Abnormal results of kidney function studies	x

*Due to variability in coding practices between hospitals and trusts, it is difficult to place reliance on numbers after the decimal place in ICD-10 codes: suggestion is that all N17 codes are included regardless of subcategories.

Appendix 3 – Skeleton tables

Table A3.1 Characteristics of study population on ACEI/ARBs, BBs, CCBs, thiazide diuretics, and an age, sex and GP practiced matched control group of patients not prescribed any of these drugs. Data are number (%).

	ACEI/ARB (n=x)	Beta-blockers (n=u)	Calcium channel blockers (n=y)	Thiazides (n=z)	Control (n=w)
Female (%)	n (%)				
Age (at baseline)					
18–44					
45–54					
55–59					
60–64					
65–69					
70–74					
75–84					
85+					
Comorbidity (at baseline)					
CKD stage					
eGFR >=60 (stage 1/2)					
eGFR 45–59 (stage 3a)					
eGFR 30–44 (stage 3b)					
eGFR 15–29 (stage 4)					
eGFR <15 (stage 5)					
Diabetes mellitus					
Ischaemic heart disease					
Cardiac failure					
Hypertension					
Cardiac arrhythmia					
Proteinuria					
Index of multiple deprivation (quintiles)					
0–20					
21–40					
41–60					
61–80					
81–100					
Ethnicity					
White					
South Asian					
Black					
Other					
Missing					
BMI					
Underweight					
Normal					
Overweight/obese					
Missing					

Table A3.1 continued.

	ACEI/ARB (n=x)	Beta–blockers (n=u)	Calcium channel blockers (n=y)	Thiazides (n=z)	Control (n=w)
Smoking					
Non-smoker/ex-smoker					
Current smoker					
Missing					
Alcohol use					
Non-problem drinker					
Problem drinker					
Missing					

Table A3.2 AKI incidence rates and rate ratios for HES linked CPRD population with a new prescription for ACEI/ARBs, BBs, CCBs, thiazide diuretics (between April 1997 and October 2011), and a age, sex and GP practice matched control group not prescribed any of these drugs.

Exposure	Person years	AKI cases	Crude AKI incidence rate (95% CI)	Age and sex adjusted IRR (95% CI)*	Fully adjusted IRR (95% CI)**
Primary analysis – exposure to a single class of antihypertensive only					
ACEI/ARB					
BB					
CCB					
Thiazides					
Control				1	1
Sensitivity analysis – binary indicators for exposure to each class of antihypertensive					
ACEI/ARB					
BB					
CCB					
Thiazides					
Control				1	1

IRR: Incidence rate ratio.
*Adjusted for: age and sex using Poisson regression.
**Adjusted for: age, sex, and covariates informed by DAG.

Table A3.3 Poisson regression model comparing AKI incidence rate ratios (95% CIs) in each of the exposure groups prescribed ACEI/ARBs, BBS, CCBs or thiazide diuretics with the control group as the reference category – unadjusted and adjusted incidence rate ratios.

	Incidence rate ratio (95% CI)		
	Crude	Age & sex adjusted	Fully adjusted*
Exposure			
ACEI/ARB			
BB			
CCB			
Thiazides			
Control	reference	reference	reference
Sex			
Female			
Male	reference	reference	reference
Age			
18–44	reference	reference	reference
45–54			
55–59			
60–64			
65–69			
70–74			
75–84			
85+			
Comorbidity			
CKD			
eGFR >=60 (stage 1/2)	reference	reference	reference
eGFR 45–59 (stage 3a)			
eGFR 30–44 (stage 3b)			
eGFR 15–29 (stage 4)			
eGFR <15 (stage 5)			
Diabetes mellitus			
Ischaemic heart disease			
Cardiac failure			
Hypertension			
Arrhythmia			
Proteinuria			

*Adjusted for: age, sex, and covariates informed by DAG.

Table A3.4 Crude and adjusted AKI incidence rates (95% CIs) for subgroups of ACEI/ARB users taking additional medications compared to those on an ACEI/ARB alone.

Exposure	Person years	AKI cases	Crude AKI incidence rate (95% CI)	Age and sex adjusted IRR (95% CI)	Fully adjusted IRR (95% CI)*
ACEI/ARB alone				reference	reference
ACEI/ARB + thiazide diuretic					
ACEI/ARB + loop diuretic					
ACEI/ARB + loop + thiazide					
ACEI/ARB + loop + potassium-sparing diuretic					
ACEI/ARB + loop diuretic + potassium-sparing diuretic + thiazide					
ACEI/ARB + NSAID (+/- any BB, CCB or diuretic)					

IRR: Incidence rate ratio.
*Adjusted for: age, sex and covariates informed by DAG.

Table A3.5 AKI incidence rate for CPRD adult population on ACEI compared to ARBs.

Exposure	Person years	AKI cases	Crude AKI incidence rate (95% CI)	Age and sex adjusted IRR (95% CI)	Fully adjusted IRR (95% CI)*
ACEI alone				reference	reference
ARB alone					

IRR: incidence rate ratio.
*Adjusted for: age, sex and covariates informed by DAG.

Table 3.6 Mortality following AKI episodes.

	Exposure				
	AKI on ACEI/ARB	AKI on BB	AKI on CCB	AKI on thiazides	AKI control
Person years					
Deaths					
Crude mortality rate (95% CI)					
Any time following AKI					
0–3 months					
4–6 months					
7–12 months					
Mortality rate ratio (95% CI)					
Any time following AKI					
- crude					reference
- age and sex adjusted					reference
- fully adjusted*					reference
0–3 months					
- crude					reference
- age and sex adjusted					reference
- fully adjusted*					reference
4–6 months					
- crude					reference
- age and sex adjusted					reference
- fully adjusted*					reference
7–12 months					
- crude					reference
- age and sex adjusted					reference
- fully adjusted*					reference

*Adjusted for: age, sex and covariates informed by DAG.

Amendment

A. Background

A recent meta-analysis demonstrated an association between socioeconomic deprivation and risk of chronic kidney disease (CKD).[1] The review estimated that the odds of low renal function was 1.41 times greater in those with low socioeconomic status (SES) compared to high SES. CKD is also a main risk factor for acute kidney injury (AKI). Therefore, in our original protocol we requested SES as a confounding variable for our main analyses (the association of ACEI/ARB use with AKI).

As expected, a preliminary analysis for the study described in the main protocol revealed an association between SES and AKI. Risk of AKI increased with increasing level of deprivation. For example, after adjusting for age, sex, calendar period, antihypertensive drug exposure (ACEI/ARB, BB, CCB, thiazide diuretics), time-updated chronic comorbidities (DM, IHD, cardiac failure, arrhythmia, hypertension), baseline CKD stage, and lifestyle covariates (smoking, BMI, alcohol intake), those in the most deprived IMD quintile were 1.59 (95% CI 1.41, 1.66) times more likely to have AKI than those in the least deprived quintile; While those in the second quintile were 1.07 (95% CI 0.96, 1.19) times more likely to have AKI than those in the least deprived (first) quintile.

Therefore, as a secondary analysis using the cohort identified in the main protocol, we aim to investigate the association between SES and AKI in a cohort of antihypertensive users.

B. Objectives

The overall aim is to investigate the association between SES and risk of acute kidney injury (AKI) in new antihypertensive users. Specifically we aim to:

1. Describe rate of AKI by level of deprivation defined by quintiles of Index of Multiple Deprivation (IMD).
2. Assess whether there is a dose-response relationship between level of deprivation and risk of AKI.
3. Explore any variation in AKI rates in different levels of deprivation over time, and by age, sex, and geographical region (London versus the rest of England – postcode derived IMD may have a different meaning for people living in London compared to elsewhere).
4. Investigate mediators of the association between socioeconomic status AKI, for their presence and magnitude, and whether they vary with ethnicity.

C. Study type

This study will test the null hypothesis that, among patients taking antihypertensives, there is no association between SES and rate of AKI.

D. Study design and study population

This will be a population-based cohort study. We will use the same cohort as that described in the main protocol; that is, new users of antihypertensive medications aged 18 and over. However, to avoid selection bias, we will include those without serum creatinine results recorded in the 12 months prior to cohort entry. Those with baseline serum creatinine test results may represent a select group of patients (renal function is more likely to be tested in

those who are acutely unwell, or routinely monitored as part of incentivised programs; diabetics for example).

E. Sample size – power calculation

In the study documented in the main protocol, we identified a cohort of 570,445 eligible new users of antihypertensive drugs (including those without baseline serum creatinine results). During follow-up, 14,907 people developed AKI. Twenty-four percent of the cohort (n=135,536) were in the lowest quintile of Index of Multiple Deprivation (IMD). Taking those in the lowest quintile of IMD as the exposed group, we have a power of 80% to detect an effect size of 1.02 or more (Calculated using G*Power, version 3.1.9.2).

F. Selection of comparison group(s) or controls

Comparison groups will be defined within the cohort according to exposure status (see Section G below).

G. Exposures, outcomes and covariates

G1. Exposures

Our primary exposure will be socioeconomic status level defined using quintiles of IMD scores for 2004. We will use the 2004 IMD data because it is as close to the midpoint of the study period as possible.

G2. Outcomes

The outcome will be AKI as defined in the main protocol.

G3. Covariates

Based on a priori knowledge, we will consider the following pre-specified variables as potential confounders: age, sex, calendar period, region (London versus rest of England – postcode derived IMD may have a different meaning for people living in London compared to elsewhere), and ethnicity. We will include calendar period (1997–2000, 2001–2004, 2005–2008, 2009–2011, and 2012–2014) as a covariate to adjust for the many changes in clinical, diagnostic and administrative practices over the study period that may influence the measurement of baseline renal function and registration of outcomes. Research suggests[2] that a large proportion of ethnicity data is missing. We will therefore only rely on ethnicity as a covariate in secondary analyses. Ethnicity will be classified according to both Read and ICD-10 coded data to improve data completeness.[2]

We believe that the following variables are on the causal pathway between SES and AKI and we will therefore not adjust for them the main analysis: lifestyle factors (smoking, alcohol intake, and body mass index), chronic comorbidities (diabetes mellitus, hypertension, cardiac failure, ischaemic heart disease, and arrhythmia), and antihypertensive medications. However, we will consider the magnitude of their contribution in mediation analyses.

F. Analysis

We will present descriptive characteristics for individuals in the cohort by level of deprivation (quintile of IMD). We will calculate absolute rates of AKI for each level of deprivation; initially overall and then stratified by age, sex, calendar period, and region (London versus rest of England). We will calculate incidence rate ratios comparing AKI rates for each level of deprivation with the least deprived quintile using Poisson regression, adjusting for potential

confounders and using robust standard errors to account for clustering by general practice. We will initially adjust for age and sex only, and then fit an adjusted model informed by a priori knowledge including the following covariates: age, sex, calendar period, and region. Finally, we will fit a model additionally adjusting for ethnicity.

Subsequently, using a conceptual framework, we will attempt to indirectly assess the contribution of each health-related behaviour (lifestyle covariates) and time-updated comorbidity to the association between SES and AKI using a multiple regression model and investigate whether they vary by ethnicity.

As sensitivity analyses we will stratify by region (London versus rest of England – postcode derived IMD may have a different meaning for people living in London compared to elsewhere), and restrict to post 2006 data (in order to: i) limit differential misclassification of the outcome over time; ii) improve the reliability of baseline CKD stage – since 2006 GPs were reimbursed for providing a register of CKD patients); and iii) improve the reliability of ethnicity data – after 2006 recording of ethnicity was rewarded as part of the Quality and Outcomes Framework leading to improvements in the completeness of ethnicity recording in CPRD).

All data management and analyses will be performed using Stata version 14 (StataCorp, Texas).

References

1 Vart P, Gansevoort RT, Joosten MM, Bültmann U, Reijneveld SA. Socioeconomic disparities in chronic kidney disease: A systematic review and meta-analysis. *Am J Prev Med* 2015; **48**: 580–92.

2 Mathur R, Bhaskaran K, Chaturvedi N, *et al.* Completeness and usability of ethnicity data in UK-based primary care and hospital databases. *J Public Health (Oxf)* 2013; **36**: 684–92.

Amendment II

A. Background

The original ISAC protocol proposed to "undertake a complete case analysis unless missing data is greater than 30% when we will undertake further sensitivity analyses". In the data described, there are two key confounders each with missing data greater than 50%: baseline chronic kidney disease stage (a measure of kidney function) and ethnicity.

The published analysis took a complete case approach to missing data in variables other than kidney function and ethnicity (due to the large amount of missing data in these two confounders, sensitivity analyses were undertaken). The absence of a kidney function measurement was treated as a separate "unmeasured" group rather than as missing values. So with respect to this major confounder, the missing data was handled using a missing indicator approach. With respect to missing ethnicity values, a sensitivity analysis was undertaken restricting the main analysis to a subset of patients with recorded ethnicity.

We have undertaken mathematical work (not involving any data) investigating the circumstances under which the missing indicator approach and the missingness pattern approach are valid. The missing indicator approach involves creating a "missing" variable for each confounder with some missing data, indicating whether that variable was missing or not for each patient, and incorporating that indicator into the analysis model. The missingness pattern approach adds interactions between the missingness indicator variable(s) and the other confounders in the analysis model. These approaches are simple, transparent, and less computationally intensive than other popular approaches. Our theoretical work suggests that the missingness pattern approach may be a reasonable analysis option in many settings using data taken from electronic health records, and that the even simpler missingness indicator approach is likely to be approximately unbiased in many of these scenarios.

In order to demonstrate the usefulness of our proposed approach in practice, we would now like to: (i) apply these methods (missingness indicator and missingness pattern approach) to a CPRD dataset which has been used to address a substantive clinical question, (ii) compare them to other popular approaches (complete case and multiple imputation) which require very different underlying assumptions, and (iii) use our theoretical framework to explain differences between the resulting estimated treatment effects.

Our theoretical framework supports the use of the simple missingness indicator and missingness pattern approaches in a wide range of studies using data from electronic health records. By demonstrating the results of applying this theoretical framework to the clinical setting described above, we hope to illustrate the value of our methodological work for researchers using CPRD data.

B. Objectives

The overall aim is to compare the results of the original analysis and sensitivity analyses (conventional approaches to handle missing data), with the results obtained using our new approach to missing data. This will illustrate the value of our methodological work in the context of CPRD data. Specifically, we aim to determine:

1. The estimated effect of prescription of renin-angiotensin system blockers on the risk of acute kidney injury using a missing indicator approach (as published analysis).

2. The estimated effect using the missingness pattern approach (our variant of the missingness indicator approach).
3. The estimated effect obtained using multiple imputation.
4. The estimated effect obtained using full complete case analysis (i.e. additionally excluding people with unmeasured kidney function; which results in a much smaller sample than the original analysis).

C. Study type

As per original protocol.

D. Study design

As per original protocol.

E. Study population

As per original protocol.

F. Exposure, outcome and covariates

F1. Exposure

Binary baseline ACEI/ARB exposure status derived using time-updating exposure status as described in original protocol.

F2. Outcome

We will use a dichotomised version of the AKI outcome as defined in the original protocol (eg. within 5 years from cohort entry or not).

F3. Covariates

We will use covariates deemed of importance in the published analysis, as per original protocol.

G. Analysis

All analyses will be undertaken in STATA version 14 (StataCorp, Texas).

We will undertake propensity score analysis to estimate the effect of ACEI/ARB prescription on the risk of AKI in new users of antihypertensive drugs using different missing data methods:

1. Missing indicator approach
2. Missingness pattern approach
3. Multiple imputation
4. Complete case analysis

References

1 Mansfield KE, Nitsch D, Smeeth L, *et al.* Prescription of renin–angiotensin system blockers and risk of acute kidney injury: a population-based cohort study. *BMJ Open* 2016; **6**: e012690. doi: 10.1136/bmjopen-2016-012690

ISAC EVALUATION OF PROTOCOLS FOR RESEARCH INVOLVING CPRD DATA

FEEDBACK TO APPLICANTS

CONFIDENTIAL			***by e-mail***
PROTOCOL NO:	14_208A2		
PROTOCOL TITLE:	The incidence and mortality of acute kidney injury (AKI) associated with prescribing of angiotensin converting inhibitors and angiotensin-receptor blockers		
APPLICANT:	Dr Laurie Tomlinson, Lecturer, London School of Hygiene and Tropical Medicine and Honorary Consultant Nephrologist at Brighton and Sussex University Hospitals NHS Trust. laurie.tomlinson@lshtm.ac.uk		
APPROVED ☒	**APPROVED WITH COMMENTS (resubmission not required)** ☐	**REVISION/ RESUBMISSION REQUESTED** ☐	**REJECTED** ☐

INSTRUCTIONS:

Please include your response/s to the Reviewer's feedback below only if you are required to Revise/ Resubmit your protocol.

Protocols with an outcome of 'Approved' or 'Approved with comments' do not require resubmission to the ISAC.

REVIEWER COMMENTS:

Protocol amendment 14_208A2 is approved.

DATE OF ISAC FEEDBACK:	01/09/2017
DATE OF APPLICANT FEEDBACK:	

For protocols approved from 01 April 2014 onwards, applicants are required to include the ISAC protocol in their journal submission with a statement in the manuscript indicating that it had been approved by the ISAC (with the reference number) and made available to the journal reviewers. If the protocol was subject to any amendments, the last amended version should be the one submitted.

***** Please refer to the ISAC advice about protocol amendments provided below*****

During the course of some studies, it may become necessary to deviate from a protocol which has been approved by ISAC. Any deviation to an ISAC approved protocol should be clearly documented by the applicant but not all such amendments need be submitted for ISAC review and approval. The general principles to be applied in regard to the need for submission are as follows:

- Major amendments should be submitted
- Minor amendments need not be submitted (but must still be documented by the applicant and should normally be mentioned at the publication stage)

In cases of uncertainty, the applicant should contact the ISAC secretariat for advice quoting the original reference number and providing a brief explanation of the nature of the amendment(s) and underlying reason(s).

Major Amendments

We consider an amendment as major if it substantially changes the study design or analysis plan of the proposed research. An amendment should be considered major if it involves the following (although this is not necessarily an exhaustive list):

- A change to the primary hypothesis being tested in the research
- A change to the design of the study
- Additional outcomes or exposures unrelated to the main focus of the approved study*
- Non-trivial changes to the analysis strategy
- Not performing a primary outcome analysis
- Omissions from the analysis plan which may impact on important validity issues such as confounding
- Change of Chief Investigator
- Use of additional linkages to other databases
- Any new proposal involving contact with health professionals or patient or change in regard to such matters

* N.B. extensive changes in this respect will require a new protocol rather than an amendment - if in doubt please consult the Secretariat

Minor Amendments

Examples of amendments which can generally be considered minor include the following:

- Change of personnel other than the Chief Investigator (these should be notified to the Secretariat)
- A change to the definition of the study population, providing the change is mentioned and justified in the paper/output [NB previously major]
- Extension of the time period in relation to defining the study population
- Changes to the definitions of outcomes or exposures of interest, providing the change is mentioned and justified in the paper/output [NB previously major]
- Not using linked data which are part of the approved protocol, unless the linked data are considered critical in defining exposures or outcomes (in which case this would be a major amendment)

- Limited additional analysis suggested by unexpected findings, provided these are clearly presented as post-hoc
- Additional methods to further control for confounding or sensitivity analysis provided these are to be reported as secondary to the main findings
- Validation and data quality work provided additional information from GPs is not required

To submit an amendment of protocol to the ISAC, please submit the following documents to the ISAC mailbox (isac@cprd.com)
1. A covering letter providing justification for the request
2. A completed and, if necessary, updated application form with all changes highlighted; if new linkages are required the current version of the ISAC application form must be completed. Otherwise, the original form may be amended as necessary
3. **The updated protocol document containing the heading 'Amendment' at the end of it.** Please include all amendments to the protocol under this heading. No other changes should be made to the already approved document.

LSHTM Ethics Application & CARE Form

Project Information

1. Full project title

Dealing with missing confounder data in propensity score analysis

2. Is this Project in fulfillment of a degree?

(•) Yes ( ) No

2a. Degree registered for

PhD

2b. Have you completed upgrading?

(•) Yes
( ) No

2b (i). Enter date of upgrading

27/09/2016

2f(deg). Is this an original submission, or are you responding to a request for clarification from the LSHTM ethics committee?

(•) Original submission
( ) Responding to request for clarification

Student Details

3a. Student details

Title	First Name	Surname
[redacted]	[redacted]	[redacted]

Address [redacted]

City [redacted]

Postcode [redacted]

Telephone [redacted]

Email [redacted]

3c. Supervisor's name.

[redacted]

3c (i). Supervisor's email address (if more than one, please only provide the email address of your main supervisor)

Email [redacted]

Project ID: 15880

3 c(ii). Supervisor's institution

[redacted]

3c (iii). Supervisor status

Confirmed

Project Type

Note: Completing the filter will enable and disable sections of the form so you may not see all questions.

4. Does the research involve primary data collection, secondary analysis or a mix of both?

- ○ Primary
- ◉ Secondary
- ○ Mixed

4a(iii). Select type of project:

Project using data from secondary sources

Samples

6a. Does this research project involve the collection/use of human tissue samples e.g urine, stool, blood etc? (Please select yes even if the samples are not considered relevant material under the Human Tissue Act)

- ○ Yes
- ◉ No

6b. Will this project use living animals (either laboratory, livestock or wild animals) AND/OR biological material that has been obtained from animals in the experiments planned?

- ○ Yes
- ◉ No

Fast-Track

7. Does this project use anonymised and unlinkable secondary datasets only?

- ◉ Yes
- ○ No

7a. Will this project be conducted within the NHS?

- ○ Yes
- ◉ No

7b. Is this application for fast-track? Note: MSc applications are not currently available for fast-track

- ◉ Yes
- ○ No

7c. Select reason for fast-track

Using anonymised and unlinkable secondary datasets only

Vulnerable Groups

8c. Does this research project involve vulnerable groups? Vulnerable groups include: children, individuals with mental disability or learning difficulties, pregnant women, prisoners etc (see information icon for full description).

- ○ Yes
- ◉ No

8d. Does this research involve access to and/or storage of security sensitive research material? (please see information icon for what is considered security sensitive material)

- ○ Yes
- ◉ No

Geography

9. List the countries where the research project is to be conducted (For example: if you are conducting a secondary data analysis for your project and you will be based in the UK, select UK regardless of where the original data has come from):

United Kingdom

Please be aware that all primary health research conducted in the UK requires a sponsor. Please contact the RGIO at RGIO@lshtm.ac.uk for more information on sponsorship.

Outline

Note: Please do not copy and paste directly from the protocol. Applications where large portions of text have been copied and pasted directly from the protocol, and therefore do not properly answer the question, will be invalidated

10b. Give an outline of the proposed project, including background to the proposal. Include information from any systematic reviews that have been conducted. Sufficient detail must be given to allow the Committee to make an informed decision without reference to other documents.

Electronic Health Records (EHRs) are a valuable data source for investigating health related questions, and propensity score analysis has become an increasingly popular approach to address confounding bias in such investigations. However, because EHR data are typically routinely recorded as part of standard clinical care, there are often missing values, particularly for potential confounders. Although there is a wealth of knowledge dealing with missing confounder data in general, currently there is a lack of proper guidance for researchers trying to handle missing data within propensity score analyses. Methods which incorporate the pattern of missingness into the propensity score have been proposed, but have not been used much. My PhD aims to investigate these methods and provide practical guidance for researchers regarding the use of missing confounder methods incorporating missingness patterns for propensity score analyses. My PhD will mostly use mathematical theory and simulated data, but I plan to illustrate the methods using EHR data.

10b(i). Upload the study protocol, including data collection forms, questionnaires and topic guides. Please upload each document separately, ensuring that the date and version number of each document is correct.

[redacted]

11. State the intended value of the project, detailing why the topic is of interest or relevance. If this project or a similar one has been done before what is the value of repeating it? Give details of overviews and/or information on the Cochrane database.
This area is of increasing importance – please ensure you give a full response.

Missing data methods incorporating missingness patterns have been discussed in the literature, but not used much in practice. These methods are simple and may be appropriate in scenarios where other missing data methods may not be appropriate. These methods may be particularly useful in settings where data is routinely collected, for example studies using electronic health records data.

13. Overall aim of project

To provide guidance for researchers on using missingness patterns to deal with partially observed confounders, using a real clinical example for illustration.

14. Specific objectives of project

To investigate the assumptions underlying methods using missingness patterns.
To provide guidance for researchers about when it is appropriate to use missingness pattern methods.
To assess performance of these methods.

Methods

Note: Please do not copy and paste directly from the protocol. Applications where large portions of text have been copied and pasted directly from the protocol, and therefore do not properly answer the question, will be invalidated

15a. Specify the procedures/methodology to be conducted during the project. Please include outcome measures and plans for data management and analysis. For literature reviews, include details on search strategy, search terms, inclusion and exclusion criteria.

My PhD will primarily use mathematical theory and simulation studies, but will use the secondary data for illustration (details as follows).
Study design: cohort study using fully linked data from the UK Clinical Practice Research Datalink and the Hospital Episode Statistics database for new adult users of antihypertensive drugs between 1997-2014
Exposure: prescription of angiotensin-converting enzyme inhibitors and angiotensin receptor blockers
Outcome: incidence of acute kidney injury within 5 years of start of follow-up
Analysis methods: propensity score analysis, outcome regression
Missing data methods: complete case analysis, multiple imputation, missing indicator approach, missingness pattern approach, multiple imputation with missingness patterns

16. Proposed start date of the project

02/07/2018

17. Proposed end date of the project

28/09/2019

Experience

22a. State the personal experience of the applicant and of senior collaborators in the research project in the field concerned, and their contribution to this project. Indicate any previous work done related to the project topic including student and/or professional work, or publications

Helen Blake (applicant) is pursuing a PhD in Medical Statistics at LSHTM with a focus on missing data methods in propensity score analysis. Prior to starting the PhD, she completed an MSc Medical Statistics at LSHTM.

Dr Elizabeth Williamson (applicant's primary supervisor) is an Associate Professor of Medical Statistics at LSHTM. Her current research involves investigating methods for dealing with missing data within propensity score analyses, with a focus on data taken from electronic health records. Her publications include:
E. J. Williamson and A. Forbes (2014), Introduction to propensity scores. Respirology, 19: 625-635. doi:10.1111/resp.12312
E. J. Williamson, et al. (2011), Propensity scores: From naïve enthusiasm to intuitive understanding. Statistical Methods in Medical Research, 21(3):273-293. doi:10.1177/0962280210394483

Dr Clémence Leyrat (applicant's assistant supervisor) is an Assistant Professor in Statistics at LSHTM. Her current research involves handling missing data for propensity score estimation in the context of electronic health records. She is also interested in causal inference methodology. Her publications include:
C. Leyrat, et al. (2017), Propensity score analysis with partially observed covariates: How should multiple imputation be used? Statistical Methods in Medical Research. doi:10.1177/0962280217713032
C. Leyrat, et al. (2014), Propensity score methods for estimating relative risks in cluster randomized trials with low-incidence binary outcomes and selection bias, Statistics in Medicine, 33:3556–3575. doi: 10.1002/sim.6185

Professor James Carpenter (applicant's secondary supervisor) is a Professor of Medical Statistics at LSHTM and Programme Leader in Methodology at the MRC Clinical Trial Unit. His research interests include coping with missing data in complex hierarchical models, sensitivity analysis, meta-analysis and adaptive designs. His publications include:
J. R. Carpenter and M. G. Kenward (2013), Multiple imputation and its application. Wiley, Chichester.
J. A. C. Sterne, et al. (2009), Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ, 338:b2393. doi:10.1136/bmj.b2393

22b. Upload the CVs for all main investigators working on the project. For MSc students, please upload your CV only.

[redacted]

Informed Consent - Secondary Data

24. Is consent in place for secondary use of the data?

- (•) Yes
- ( ) No

24c(i). Please give details of the participant consent that was obtained when the original project(s) took place. Please upload copies of the original consent form(s). If there are no original consent forms (e.g. for audit or DHS data) please explain this.

Amendment to ISAC protocol - please see attachment to question 10b(i)

Confidentiality & Data

28. State how your data will be stored and what will be done with it at the end of the project.

Storage: On LSHTM computer system.
End of project: Deleted

Funding

33. Do you have external funding for this project?

- (•) Yes
- ( ) No

33a. If yes, please provide the name of the funder

ESRC and Bloomsbury DTC

33a(i). If yes, include details of the funding available for this project.

Bursary and stipend for my PhD

33a(ii). Are you in receipt of any funding from the United States? Or will you be collaborating with (or with individuals from) a US Institution/organisation?

- ○ Yes
- ◉ No

34. Has the project been sent out for peer/independent scientific review?

- ○ Yes
- ◉ No

34a. If no, why has the project not been sent peer/independent scientific review?

Still in progress

36. Does the Chief Investigator or any other investigator/collaborator have any direct personal involvement (e.g. financial, share holding, personal relationship etc.) in the organisations sponsoring or funding the research that may give rise to a possible conflict of interest?

- ○ Yes
- ◉ No

37. Will individual researchers receive any personal payment over and above normal salary, or any other benefits or incentives, for taking part in this research?

- ○ Yes
- ◉ No

Local Approval

49a. For projects using previously-collected human data, give details of all approvals under which the original project(s) took place. Please quote names of Ethics Committees and approval reference numbers (required even if previous approval was from LSHTM); if possible give web link to original project application. If there are no original approvals (e.g. for audit or DHS data) please explain this.

ISAC protocol number 14_208 (see attachment to question 10b(i))

49c. Will your analyses be for purposes entirely covered by the original ethics application where the data was collected, as detailed above?

- ○ Yes, this falls within the aims and scope of the original project
- ◉ No, the analyses and aims differ from the original project

49d(ii). If no, please detail how you will amend the original ethics application to include the current analysis.

I submitted an amendment to CPRD to the original protocol and this was approved (see Amendment II on page 42 of the attachment to question 10b(i))

Signature Instructions

The form should be completed and finalised prior to signing or requesting signatures. Students should ensure that the Supervisor signs prior to the Course Director/Project Module Organiser. For external supervisors, please ensure that they have registered for an account prior to requesting the signature.

Signature - Applicant

Student signature

I declare that:

- I undertake to abide by the ethical principles underlying the Declaration of Helsinki (1964, as amended) and good practice guidelines on the proper conduct of research.
- I Have read and understood, and agree to abide by the LSHTM Good Research Practice policy
- I undertake to abide by the UK Data Protection Act 1998 and any applicable local laws.
- I undertake to abide by all local rules for non-UK research.
- I agree to conduct my project on the basis set out in this form, and to consult staff (initially, my Supervisor) if making any subsequent changes – especially any that would affect the information given with respect to ethics approval.
- I undertake to adhere to all conditions set out by review bodies in giving approval and will not start the project until all required approvals are in place
- I agree to comply with the relevant safety requirements, and will submit a separate request for LSHTM travel insurance where relevant.
- I confirm that there are no conflicts of interest that preclude my participation in the project

[redacted]

Signature - Supervisor

Supervisor signature

I declare that:

- I agree that the information submitted in this application is a reasonable summary of the proposed project.
- I agree that this form correctly indicates whether or not ethics approval will be required.
- I agree that this form contains adequate information for the ethics committee to form an opinion of the proposed project.
- I agree that all required supporting documentation is attached to this application.
- (For MSc projects only) I agree that responses in the Risk Assessment section address the main risks connected with a project of this nature
- I have reviewed the risk of the project, including travel, and agree that it is an acceptable risk to the student
- I confirm that there are no conflicts of interest that preclude my role as supervisor for this project
- I Have read and understood, and agree to abide by the LSHTM Good Research Practice policy

[redacted]

Signature - Other

Note:

The form will automatically submit upon receipt of all required signatures.

After submission, you will receive a confirmation email with further details.

If you have not received a confirmation email within 5 working days please email ethics@lshtm.ac.uk (staff) or MScethics@lshtm.ac.uk (students) to check the status of your submission.

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT
United Kingdom
Switchboard: +44 (0)20 7636 8636

www.lshtm.ac.uk

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Observational / Interventions Research Ethics Committee

Miss Helen Blake
LSHTM

11 July 2018

Dear Helen,

Study Title: Dealing with missing confounder data in propensity score analysis

LSHTM Ethics Ref: 15880

Thank you for your application for the above research project which has now been considered by the Observational Committee via Chair's Action.

Confirmation of ethical opinion

On behalf of the Committee, I am pleased to confirm a favourable ethical opinion for the above research on the basis described in the application form, protocol and supporting documentation, subject to the conditions specified below.

Conditions of the favourable opinion

Approval is dependent on local ethical approval having been received, where relevant.

Approved documents

The final list of documents reviewed and approved is as follows:

Document Type	File Name	Date	Version
Protocol / Proposal	ISAC_14_208_ACE_ARB_associatedAKI_amendment_v3HB2017	31/08/2017	v3HB2017
Investigator CV	Helen Blake CV	06/09/2017	1

After ethical review

The Chief Investigator (CI) or delegate is responsible for informing the ethics committee of any subsequent changes to the application. These must be submitted to the committee for review using an Amendment form. Amendments must not be initiated before receipt of written favourable opinion from the committee.

The CI or delegate is also required to notify the ethics committee of any protocol violations and/or Suspected Unexpected Serious Adverse Reactions (SUSARs) which occur during the project by submitting a Serious Adverse Event form.

An annual report should be submitted to the committee using an Annual Report form on the anniversary of the approval of the study during the lifetime of the study.

At the end of the study, the CI or delegate must notify the committee using the End of Study form.

All aforementioned forms are available on the ethics online applications website and can only be submitted to the committee via the website at: http://leo.lshtm.ac.uk.

Further information is available at: www.lshtm.ac.uk/ethics.

Yours sincerely,

[redacted]

Professor John DH Porter
Chair

ethics@lshtm.ac.uk
http://www.lshtm.ac.uk/ethics/

Improving health worldwide

Appendix B

Resources for the planned systematic review

B.1 Protocol of the systematic review

Sources: Searches will be performed in Embase, Medline, PubMed and Scopus to select papers published between 1 January 1983 (following the publication of the first paper proposing propensity scores to account for confounding) and 13 May 2016.

Search algorithm: To restrict our review to studies using propensity scores in which missing data was a particular concern, the search strategy will be constructed for all four databases in order to search for articles referring to propensity score and missing data in either the title or the abstract. Eligible papers will have 'propensity score' as a phrase in the title, abstract or keyword fields. Variations of this phrase will be also considered in the search strategy, for example phrases with the words 'propensity' and 'match' in close proximity. In addition to mentioning propensity scores, the search strategies will require some variations of the phrase 'missing data', such as 'incomplete data' or phrases using Rubin's taxonomy of missing data. The search strategies for each database will be constructed with the aim of being as

similar as possible, with some variation resulting from differing syntax used in some databases. For example, the search strategy for Pubmed will be:

```
(("propensity-score"[MeSH Terms]) OR propensity analys*[Title/
Abstract] OR propensity match*[Title/Abstract] OR propensity adjust*
[Title/Abstract] OR propensity stratif*[Title/Abstract] OR propensity
covariate*[Title/Abstract] OR propensity weight*[Title/Abstract]) AND
(missing data[Title/Abstract] OR incomplete data[Title/Abstract] OR
missing value*[Title/Abstract] OR mcar data[Title/Abstract] OR mar
data[Title/Abstract] OR mnar data[Title/Abstract] OR missing
random[Title/Abstract])
```

Screening: After search results are obtained, references will be reviewed to identify and exclude duplicates, first using the "Find Duplicates" function in Endnote X7 to identify a preliminary list of duplicate records, then conducting a manual review to identify remaining duplicates. The criteria for deciding which record to keep and which to discard as a confounder, will be decided on the basis of the amount of information available in each record, favouring records with more information. I will review the resulting references by considering the title, abstract and keywords of the articles, retrieving the full text when further information is required.

Exclusion criteria: I will exclude articles if: they are unrelated to propensity score analysis of observational data; missing data methods focused on methods for handling missing data in the treatment or outcome, rather than missing confounder data; they are conference abstracts, commentary, editorials or letters; if they have no data example, real or simulated; or if they are published in languages other than English. Articles with a methodological focus will not included in the general review, but will be considered separately.

Data extraction: For each article included in the literature review, I will extract information on:

- propensity score method (matching, stratification, adjustment, weighting, or not reported)
- the number of confounders with missing data (number, or not reported)
- the overall proportion of missing data (percentage, or not reported)
- the missing data method(s) used (complete records analysis, multiple imputation, MPA, MIA, other, or not reported)
- whether the justification for the choice of missing data method(s) was discussed
- whether the plausibility of missingness assumptions were discussed
- whether further details were given regarding implementation of the missing data method (s)

B.1.1 Literature review: results of the screening for eligibility

From 13th May 2016: Searching the four databases listed above yielded 559 records (see Table B.1). Using Endnotes's "Find duplicates" function identified 192 duplicate records to be excluded. Manual review of the author field identified 40 more duplicate records, and manual review by title identified 4 further duplicates. After excluding these duplicate records, 323 records remained.

Table B.1: A table showing the number of records retrieved from each database searched on 13th May 2016.

Database	Number of records retreived	Duplicate records excluded	Number of unduplicated records
Embase	178	107	71
Medline	83	79	4
PubMed	47	41	6
Scopus	251	9	242
TOTAL	559	236	323

Bibliography

[1] Regulation (EU) No 1235/2010 of the European Parliament and of the Council, 2010 OJ L 348.

[2] J. M. Franklin, R. J. Glynn, D. Martin, and S. Schneeweiss, "Evaluating the use of nonrandomized real-world data analyses for regulatory decision making," *Clinical Pharmacology & Therapeutics*, vol. 105, no. 4, pp. 867–877, 2019.

[3] B. L. Strom, S. E. Kimmel, and S. Hennessy, *Textbook of pharmacoepidemiology.* Chichester, West Sussex [England]: Wiley Blackwell, second edition ed., 2013.

[4] J. Pearl, *Causality: Models, Reasoning and Inference.* New York: Cambridge University Press, 2nd ed., 2009.

[5] J. W. Jackson, I. Schmid, and E. A. Stuart, "Propensity scores in pharmacoepidemiology: Beyond the horizon," *Current epidemiology reports*, vol. 4, p. 271—280, December 2017.

[6] I. Abraham, "A definition of comparative effectiveness research [Peer commentary on "More research is needed—but what type?" by F. Godlee]," *BMJ*, vol. 341, 2010.

[7] E. Herrett, A. M. Gallagher, K. Bhaskaran, H. Forbes, R. Mathur, T. van Staa, and L. Smeeth, "Data resource profile: Clinical Practice Research Datalink (CPRD)," *Int J Epidemiol*, vol. 44, no. 3, p. 827, 2015.

[8] A. Herbert, L. Wijlaars, A. Zylbersztejn, D. Cromwell, and P. Hardelid, "Data resource profile: Hospital Episode Statistics Admitted Patient Care (HES APC)," *International Journal of Epidemiology*, vol. 46, pp. 1093–1093i, 03 2017.

[9] P. Vezyridis and S. Timmons, "Evolution of primary care databases in UK: a scientometric analysis of research output," *BMJ Open*, vol. 6, no. 10, 2016.

[10] S. Guo and M. W. Fraser, *Propensity score analysis : statistical methods and applications*. Los Angeles: Sage, 2010.

[11] K. E. Mansfield, D. Nitsch, L. Smeeth, K. Bhaskaran, and L. A. Tomlinson, "Prescription of renin–angiotensin system blockers and risk of acute kidney injury: a population-based cohort study," *BMJ Open*, vol. 6, no. 12, 2016.

[12] CPRD, "Clinical Practice Research Datalink - CPRD," Accessed 24th September 2019.

[13] R. Mathur, K. Bhaskaran, N. Chaturvedi, D. A. Leon, T. vanStaa, E. Grundy, and L. Smeeth, "Completeness and usability of ethnicity data in uk-based primary care and hospital databases," *Journal of Public Health*, vol. 36, no. 4, pp. 684–692, 2014.

[14] K. Bhaskaran, H. J. Forbes, I. Douglas, D. A. Leon, and L. Smeeth, "Representativeness and optimal use of body mass index (BMI) in the UK Clinical Practice Research Datalink (CPRD)," *BMJ Open*, vol. 3, no. 9, 2013.

[15] J. Chisholm, "The read clinical classification.," *BMJ: British Medical Journal*, vol. 300, no. 6732, p. 1092, 1990.

[16] NHS Digital, "International statistical classification of diseases and health related problems (ICD-10) 5th Edition," Accessed 24th September 2019.

[17] NHS Digital, "OPCS classification of interventions and procedures," Accessed 24th September 2019.

[18] NHS Digital, "Home - NHS Digital," Accessed 24th September 2019.

[19] S. Padmanabhan, L. Carty, E. Cameron, R. E. Ghosh, R. Williams, and H. Strongman, "Approach to record linkage of primary care data from Clinical Practice Research Datalink to other health-related patient data: overview

and implications," *European Journal of Epidemiology*, vol. 34, pp. 91–99, Jan 2019.

[20] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.

[21] Z. Luo, J. C. Gardiner, and C. J. Bradley, "Applying propensity score methods in medical research: Pitfalls and prospects," *Medical Care Research and Review*, vol. 67, no. 5, pp. 528–554, 2010. 20442340[pmid] Med Care Res Rev.

[22] P. W. Holland, "Statistics and causal inference," *J Am Stat Assoc*, vol. 81, no. 396, pp. 945–960, 1986.

[23] E. Williamson, R. Morley, A. Lucas, and J. Carpenter, "Propensity scores: From naïve enthusiasm to intuitive understanding," *Stat Methods Med Res*, vol. 21, no. 3, pp. 273–293, 2012.

[24] W. M. Holmes, *Using propensity scores in quasi-experimental designs.* Los Angeles: Sage, 2014.

[25] E. J. Williamson and A. Forbes, "Introduction to propensity scores," *Respirology*, vol. 19, no. 5, pp. 625–635, 2014.

[26] G. W. Imbens, "Nonparametric estimation of average treatment effects under exogeneity: A review," *Rev Econ Stat*, vol. 86, no. 1, pp. 4–29, 2004.

[27] S. Greenland, J. M. Robins, and J. Pearl, "Confounding and collapsibility in causal inference," *Statist. Sci.*, vol. 14, pp. 29–46, 02 1999.

[28] P. C. Austin, "An introduction to propensity score methods for reducing the effects of confounding in observational studies," *Multivariate Behav Res*, vol. 46, no. 3, pp. 399–424, 2011.

[29] D. F. McCaffrey, G. Ridgeway, and A. R. Morral, "Propensity score estimation with boosted regression for evaluating causal effects in observational studies," *Psychological Methods*, vol. 9, no. 4, pp. 403–425, 2004. Article.

[30] D. B. Rubin *et al.*, "For objective causal inference, design trumps analysis," *The Annals of Applied Statistics*, vol. 2, no. 3, pp. 808–840, 2008.

[31] M. Hernán and J. Robins, *Causal Inference*. Boca Raton: Chapman & Hall/CRC, 2020.

[32] E. Stuart, "The why, when, and how of propensity score methods for estimating causal effects," 2011.

[33] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.

[34] M. Höfler, "Causal inference based on counterfactuals," *BMC Med Res Methodol*, vol. 5, pp. 28–28, 2005.

[35] M. E. Halloran and C. J. Struchiner, "Causal inference in infectious diseases," *Epidemiology*, vol. 6, no. 2, pp. 142–151, 1995. Article.

[36] M. A. Hernán and J. M. Robins, "Estimating causal effects from epidemiological data," *J Epidemiol Community Health*, vol. 60, no. 7, pp. 578–586, 2006.

[37] R. J. Little and D. B. Rubin, "Causal effects in clinical and epidemiological studies via potential outcomes: Concepts and analytical approaches," *Annu Rev Public Health*, vol. 21, pp. 121–145, 2000.

[38] J. P. Vandenbroucke, "The history of confounding," *Sozial- und Präventivmedizin*, vol. 47, pp. 216–224, Jul 2002.

[39] T. J. VanderWeele and I. Shpitser, "On the definition of a confounder," *Ann. Statist.*, vol. 41, pp. 196–220, 02 2013.

[40] H. Cham and S. G. West, "Propensity score analysis with missing data," *Psychological Methods*, 2016. Export Date: 13 May 2016 Article in Press.

[41] J. K. Lunceford and M. Davidian, "Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study," *Statistics in Medicine*, vol. 23, no. 19, pp. 2937–2960, 2004.

[42] P. C. Austin, "Goodness-of-fit diagnostics for the propensity score model when estimating treatment effects using covariate adjustment with the propensity score," *Pharmacoepidemiology and Drug Safety*, vol. 17, no. 12, pp. 1202–1217, 2008.

[43] S. Vansteelandt and R. Daniel, "On regression adjustment for the propensity score," *Stat Med*, vol. 33, no. 23, pp. 4053–4072, 2014.

[44] J. Carpenter and M. Kenward, *Multiple Imputation and Its Application*. Statistics in Practice, Chichester: Wiley, 2013.

[45] R. Little and D. Rubin, *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics, Wiley, 2002.

[46] R. B. D'Agostino and D. B. Rubin, "Estimating and using propensity scores with partially missing data," *J Am Stat Assoc*, vol. 95, no. 451, pp. 749–759, 2000.

[47] C. A. Welch, I. Petersen, J. W. Bartlett, I. R. White, L. Marston, R. W. Morris, I. Nazareth, K. Walters, and J. Carpenter, "Evaluation of two-fold fully conditional specification multiple imputation for longitudinal electronic health record data," *Statistics in Medicine*, vol. 33, no. 21, pp. 3725–3737, 2014.

[48] S. Seaman, J. Galati, D. Jackson, and J. Carlin, "What is meant by "missing at random"?," *Statist. Sci.*, vol. 28, pp. 257–268, 05 2013.

[49] J. A. C. Sterne, I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, and J. R. Carpenter, "Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls," *BMJ*, vol. 338, 2009. doi:10.1136/bmj.b2393.

[50] J. W. Bartlett, O. Harel, and J. R. Carpenter, "Asymptotically unbiased estimation of exposure odds ratios in complete records logistic regression," *Am J Epidemiol*, vol. 182, no. 8, pp. 730–736, 2015.

[51] J. Hill, "Reducing bias in treatment effect estimation in observational studies suffering from missing data." ISERP Working Papers, 2004.

[52] M. J. Knol, K. J. Janssen, A. R. T. Donders, A. C. Egberts, E. R. Heerdink, D. E. Grobbee, K. G. Moons, and M. I. Geerlings, "Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example," *J Clin Epidemiol*, vol. 63, no. 7, pp. 728 – 736, 2010.

[53] O. Harel and X.-H. Zhou, "Multiple imputation: review of theory, implementation and software," *Stat Med*, vol. 26, no. 16, pp. 3057–3077, 2007.

[54] J. M. Robins, A. Rotnitzky, and L. P. Zhao, "Analysis of semiparametric regression models for repeated outcomes in the presence of missing data," *J Am Stat Assoc*, vol. 90, no. 429, pp. 106–121, 1995.

[55] E. A. Stuart, "Matching methods for causal inference: A review and a look forward," *Statist. Sci.*, vol. 25, pp. 1–21, 02 2010.

[56] W. Vach and M. Blettner, "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," *Am J Epidemiol*, vol. 134, pp. 895–907, 10 1991.

[57] S. Greenland and W. D. Finkle, "A critical look at methods for handling missing covariates in epidemiologic regression analyses," *American Journal of Epidemiology*, vol. 142, no. 12, pp. 1255–1264, 1995.

[58] R. H. Groenwold, I. R. White, A. R. T. Donders, J. R. Carpenter, D. G. Altman, and K. G. Moons, "Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis," *Can Med Assoc J*, vol. 184, no. 11, pp. 1265–1269, 2012.

[59] M. P. Jones, "Indicator and stratification methods for missing explanatory variables in multiple linear regression," *J Am Stat Assoc*, vol. 91, no. 433, pp. 222–230, 1996.

[60] C. Leyrat, S. R. Seaman, I. R. White, I. Douglas, L. Smeeth, J. Kim, M. Resche-Rigon, J. R. Carpenter, and E. J. Williamson, "Propensity score analysis with partially observed covariates: How should multiple imputation be used?," *Stat Methods Med Res*, 2017. doi: 10.1177/0962280217713032.

[61] E. Granger, J. C. Sergeant, and M. Lunt, "Avoiding pitfalls when combining multiple imputation and propensity scores," *Statistics in Medicine*, vol. 38, 2019. doi: 10.1002/sim.8355.

[62] P. R. Rosenbaum and D. B. Rubin, "Reducing bias in observational studies using subclassification on the propensity score," *J Am Stat Assoc*, vol. 79, no. 387, pp. 516–524, 1984.

[63] A. Mattei, "Estimating and using propensity score in presence of missing background data: An application to assess the impact of childbearing on wellbeing," *Stat Methods Appl*, vol. 18, no. 2, pp. 257–273, 2009.

[64] R. D'Agostino, W. Lang, M. Walkup, T. Morgan, and A. Karter, "Examining the impact of missing data on propensity score estimation in determining the effectiveness of self-monitoring of blood glucose (SMBG)," *Health Serv Outcomes Res Method*, vol. 2, no. 3, pp. 291–315, 2001.

[65] E. J. Williamson, Z. Aitken, J. Lawrie, S. C. Dharmage, J. A. Burgess, and A. B. Forbes, "Introduction to causal diagrams for confounder selection," *Respirology*, vol. 19, no. 3, pp. 303–311, 2014.

[66] J. Pearl, "Causal diagrams for empirical research," *Biometrika*, vol. 82, no. 4, p. 669, 1995.

[67] T. Richardson and J. Robins, "Technical report 128. Single World Intervention Graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality." http://www.csss.washington.edu/Papers/wp128.pdf, 2013.

[68] A. Balke and J. Pearl, "Probabilistic evaluation of counterfactual queries," in *Proceedings of the Twelfth AAAI National Conference on Artificial Intelligence*, pp. 230–237, 1994.

[69] I. Shpitser and J. Pearl, "What counterfactuals can be tested," in *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, pp. 352–359, 2007.

[70] Y. Qu and I. Lipkovich, "Propensity score estimation with missing values using a multiple imputation missingness pattern (MIMP) approach," *Stat Med*, vol. 28, no. 9, pp. 1402–1414, 2009.

[71] L. Malla, R. Perera-Salazar, E. McFadden, M. Ogero, K. Stepniewska, and M. English, "Handling missing data in propensity score estimation in comparative effectiveness evaluations: a systematic review," *J Comp Eff Res*, vol. 7, no. 3, pp. 271–279, 2018.

[72] R. Mitra and J. P. Reiter, "A comparison of two methods of estimating propensity scores after multiple imputation," *Statistical methods in medical research*, vol. 25, no. 1, pp. 188–204, 2016.

[73] E. von Elm, D. G. Altman, M. Egger, S. J. Pocock, P. C. Gøtzsche, J. P. Vandenbroucke, and for the STROBE Initiative, "The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies," *Annals of Internal Medicine*, vol. 147, pp. 573–577, 10 2007.

[74] M. A. Hernán, M. McAdams, N. McGrath, E. Lanoy, and D. Costagliola, "Observation plans in longitudinal studies with time-varying treatments," *Statistical Methods in Medical Research*, vol. 18, no. 1, pp. 27–52, 2009.

[75] N. Kreif, O. Sofrygin, J. Schmittdiel, A. Adams, R. Grant, Z. Zhu, M. van der Laan, and R. Neugebauer, "Evaluation of adaptive treatment strategies in an observational study where time-varying covariates are not monitored systematically," 2018. arXiv preprint arXiv:1806.11153 [stat.ME].

[76] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. Cambridge, Massachusetts: MIT press, 2009.

[77] J. Textor, J. Hardt, and S. Knüppel, "DAGitty: A graphical tool for analyzing causal diagrams," *Epidemiology*, vol. 22, p. 745, 2011.

[78] "Hypertension in adults: diagnosis and management - clinical guideline [cg127]." National Institute for Health and Care Excellence (NICE) website. https://www.nice.org.uk/guidance/cg127/, accessed 9th August 2019.

[79] H. I. McDonald, C. Shaw, S. L. Thomas, K. E. Mansfield, L. A. Tomlinson, and D. Nitsch, "Methodological challenges when carrying out research on CKD and AKI using routine electronic health records," *Kidney Int*, 2016.

[80] S. Seaman and I. White, "Inverse probability weighting with missing predictors of treatment assignment or missingness," *Commun Stat Theory Methods*, vol. 43, no. 16, pp. 3499–3515, 2014.

[81] S. Greenland, "Quantifying biases in causal models: Classical confounding vs collider-stratification bias," *Epidemiology*, vol. 14, no. 3, pp. 300–306, 2003.

[82] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[83] L. A. Stefanski and D. D. Boos, "The calculus of m-estimation," *The American Statistician*, vol. 56, no. 1, pp. 29–38, 2002.

[84] T. D. Pigott, "A review of methods for missing data," *Educ Res Eval*, vol. 7, no. 4, pp. 353–383, 2001.

[85] C. D. Nguyen, J. B. Carlin, and K. J. Lee, "Model checking in multiple imputation: an overview and case study," *Emerg Themes Epidemiol*, vol. 14, p. 8, Aug 2017.

[86] H. A. Blake, C. Leyrat, K. E. Mansfield, S. Seaman, L. A. Tomlinson, J. Carpenter, and E. J. Williamson, "Propensity scores using missingness pattern information: a practical guide," 2019. Under review in Statistics in Medicine. arXiv preprint arXiv:1901.03981 [stat.ME].

[87] A. S. Levey, L. A. Stevens, C. H. Schmid, Y. L. Zhang, A. F. Castro 3rd, H. I. Feldman, J. W. Kusek, P. Eggers, F. Van Lente, T. Greene, J. Coresh, and CKD-EPI (Chronic Kidney Disease Epidemiology Collaboration), "A new equation to estimate glomerular filtration rate," *Ann Intern Med*, vol. 150, no. 9, pp. 604–612, 2009.

[88] E. J. Williamson, A. Forbes, and I. R. White, "Variance reduction in randomised trials by inverse probability weighting using the propensity score," *Stat Med*, vol. 33, no. 5, pp. 721–737, 2014.