Implementation, evaluation and application of multiple imputation for missing data in longitudinal electronic health record research

Catherine Welch

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of the

University of London.

Department of Primary Care and Population Health

University College London

March 18, 2015

Students declaration

I, Catherine Anne Welch, confirm the work presented in this thesis is my own, except for the following collaborations.

In 2009 to 2012 I was employed by the research department of Primary care and Population Health (PCPH) as an research associate to work on the MRC funded project 'Missing data imputation in clinical databases: development of a longitudinal model for cardiovascular risk factors' led by my supervisor Dr Irene Petersen. The aim of the project was to develop, implement and evaluate multiple imputation algorithms for missing data, which take into account the dynamic and longitudinal structure of large primary care databases.

The specific objectives of the project were to:

1. Describe the patterns and structure of missing data over time in The Health Improvement Network (THIN) primary care database;

2. Develop an imputation algorithm to multiply impute the missing cardiovascular health indicators, taking account of the specific features and timing of the data recording; and

3. Evaluate the results of imputation by comparison with distributions of relevant variables seen in external research data and population surveys, and by application of imputed data in ongoing research projects.

The project was supported by a steering group including Irene Petersen, James Carpenter, Irwin Nazareth, Kate Walters, Louise Marston, Richard Morris, Jonathan Bartlett and Ian White. The steering group provided overall and individual feedback during the course of the project.

The results described in chapter 4 'Descriptive analysis of health indicator records in The Health Improvement Network' and section 6.1 'Imputing missing data in THIN', were part of the research project described above.

Nevalainen and colleagues proposed the two-fold fully conditional specification (FCS) algorithm, published in Statics in Medicine in 2009[1]. I requested the original SAS code from the authors and I wrote a Stata program to perform multiple imputation using the two-fold FCS algorithm. As described in the thesis, I further developed the two-fold FCS algorithm for imputing longitudinal electronic health records. The Stata journal published

details of the program in 2014[2] and it is public available from the Statistical Software Components archive (`http://ideas.repec.org/s/boc/bocode.html`).

Work I completed specifically for this PhD includes the simulation studies described in chapter 5 'Developing and evaluating the two-fold FCS algorithm', the first substantial study to evaluate the two-fold FCS algorithm and assess bias and precision in different settings. Some of this work was published in Statistics in Medicine in 2014[3] and, likewise, work in section 4.4 describing a new approach to deal with outliers in data with repeated measurements within patients (Published in Pharmacoepidemiology and Drug safety in 2011[4]). Finally, I applied the two-fold FCS in a substantive analysis of THIN to examine the characteristics of patients with type 2 diabetes who had a greater than average total cholesterol reduction after initiating statin treatment (section 6.2). For this study I wrote the protocol and designed this study from scratch.

Abstract

Longitudinal electronic health records are a valuable resource for research because they contain information on many patients over long follow-up periods. Missing data commonly occur in these data because it was collected for clinical and not research purposes. Analysing data with missing values can potentially bias estimates and standard errors resulting in invalid inferences.

Multiple imputation, commonly used in research to impute missing values, is increasingly regarded as the standard method for handling missing data in medical research because of its practicality and flexibility under the assumption the data is missing at random (MAR). Until now, few imputation approaches are sufficiently flexible to account for the longitudinal and dynamic structure of electronic health records. However, the two-fold fully conditional specification (FCS) algorithm was proposed to impute missing values in longitudinal data, but this methods was not currently validated in the complex setting of longitudinal electronic health records.

I propose to adapt, evaluate and implement the two-fold FCS algorithm to impute missing data from large primary care database. To achieve this, first I investigate the extent and patterns of missing data in a longitudinal clinical database for health indicators associated with cardiovascular disease risk to determine if the MAR assumption is plausible. Additionally, I develop methods to identify and remove outliers, which can potentially bias imputations, from data with repeated measurements before imputation.

Next, I adapt and develop the two-fold FCS multiple imputation algorithm to impute missing values in longitudinal clinical data for health indicators associated with cardiovascular disease risk and I validate the two-fold FCS algorithm to assess bias and precision through challenging simulation studies. I develop a new software programme which implements this adapted version of the two-fold FCS algorithm to impute missing values in longitudinal data.

Finally, I apply the two-fold FCS algorithm in THIN to (i) model cardiovascular disease risk and (ii) understand factors associated with greater total cholesterol reduction in patients with type II diabetes.

Acknowledgements

Firstly, I would like to acknowledge my supervisors Irene Petersen, James Carpenter and Jonathan Bartlett, for their support and guidance. In addition to their statistical input, I am grateful to Irene Petersen for her patience and advice to help me improve the quality of my writing. I would also like to acknowledge Jonathan Bartlett for his statistical input and advice to develop the Stata code to run the two-fold fully conditional specification algorithm.

I am also grateful to the other members of the 'Missing data imputation in clinical databases: development of a longitudinal model for cardiovascular risk factors' Steering Group; Irwin Nazareth, Kate Walters, Richard Morris, Ian White and Louise Marston, for the wealth of knowledge they brought to the project.

Next, I would like to thank the Medical Research Council for funding this project. Without their financial support this project could not have happened.

I would also like to acknowledge my colleagues at the Department of Primary Care and Population Health for their moral support throughout the whole PhD and also for kindly volunteering to proof read chapters: Linda Wijlaars, Shuk-Li Man, Ruth Blackburn, Hillary Davies, Gareth James, Nathan Davis and Rachael Hunter. Also, thank you to Brennan Kahan for his helpful advice on my introduction and discussion chapters.

Finally, I wish to thank my friends and family for their encouragement and belief.

Contents

List of Tables

List of Figures

Abbreviations

ACE	Angiotensin-converting-enzyme
ACU	Acceptable computer usage
AF	Atrial fibrillation
AHD	Additional health data
AMR	Acceptable mortality recording
ARB	Angiotensin receptor blocker
BMI	Body mass index
BNF	British National Formulary
CHD	Coronary heart disease
CI	Confidence interval
CKD	Chronic kidney disease
COPD	Chronic obstructive pulmonary disease
CPRD	Clinical Practice Research Datalink
CSD	Cegedim Strategic Data
CVD	Cardiovascular disease
DoH	Department of Health
EM	Expectation maximisation
FCS	Fully conditional specification
FIML	Full information maximum likelihood
FR	Glomerular filtration rate
GP	General practitioner
HbA_{1c}	Glycosylated haemoglobin
HDL	High-density lipoprotein
HR	Hazard ratio
HRQol	Health related quality of life
HSE	Health Survey for England
IQR	Interquartile range
IRR	Incidence risk ratio
LOCF	Last observation carried forward
LSHTM	London School of Hygiene and Tropical Medicine
MAR	Missing at random
MCAR	Missing completely at random
ML	Maximum likelihood
MI	Multiple imputation
MNAR	Missing not at random
MRC	Medical Research Council
MSE	Mean standard error
NHS	National Health Service
NICE	National Institute For Clinical Excellence
OR	Odds ratio
PCPH	Primary Care and Population Health
PVI	Postcode variable indicator
RI	Reference interval
SD	Standard deviation
SE	Standard error
SRC	Scientific review committee
THIN	The Health Improvement Network
QOF	Quality Outcome Framework
UCL	University College London
UK	United Kingdom

Chapter 1

Introduction

In healthcare research, we cannot answer important questions using either randomised controlled trials or stand-alone observational studies. For example, to monitor doctor prescribing behaviour, to study rare diseases (requires recruiting many patients) or to investigate disease risk, develop risk prediction models and investigate the long term effects of certain treatments (requires a long follow-up time). In these scenarios, electronic health records of routinely collected clinical information are a valuable resource for epidemiological investigations and health care research. Studies involving electronic health records address some of these questions, for example evaluating antidepressant prescribing in pregnant women and children[5, 6], the large patient numbers allowed research on rare conditions such as severe mental illness[7] and the long follow-up time allowed development of risk prediction models for cardiovascular disease[8, 9, 10]. Despite the potential advantages of analysing electronic health records, pressing methodological challenges remain, such as missing data. For example, studies on cardiovascular risks require regular, recorded health indicators (such as weight, height, blood pressure and smoking status) to achieve reliable inferences from statistical analyses but substantial missing data may bias the results if not handled appropriately.

An important, widely used example of electronic clinical health records is The Health Improvement Network (THIN)[11] primary care database. Primary care databases contain data on patient characteristics (such as age, sex, socioeconomic status), diagnoses, prescriptions and health indicators captured through routine consultations with a general practitioner (GP) or nurse. THIN includes electronic patient records for more than 11 million patients registered with more than 500 General Practices across the United Kingdom (UK), some collecting data since the early 1990s. GPs record each patient consultation from when patients register with the practice to the time they leave, providing a longitudinal record of health data (mean follow-up time is 6.7 years)[11]. Inevitably, information is recorded intermittently (non-monotone) in primary care databases due to irregular times between consultations and only recording information relevant to the clinical question at the consultation.

The National Health Service (NHS) introduced legislation to encourage GPs to monitor patient health indicators in primary care. New Patient Health Checks were introduced as part of the NHS Department of Health (DoH) Regulations in 1992[12], aimed to improve quality of practice by linking pay to performance. This legislation encouraged GPs to record health indicators for all newly registered patients. Despite the introduction of this

legislation, previous research showed nearly 40% of patients did not have blood pressure or weight measured in the first registration year[13]. Since 2004, patients diagnosed with specific chronic conditions or diseases are monitored at regular intervals as a part of National Health Service (NHS) care delivery plan for UK general practice (Quality Outcome Framework, QOF)[14]. Therefore, patients with previous cardiovascular events, or at risk of these, are much more likely to have several, regular measurements of health indicators recorded compared to patients without these diseases. However, limited information exists on the structure and extent of missing data in primary care records, where the procedure and motivation for recording health indicators varies over time.

Historically, 'ad–hoc' methods were commonly used for handling missing data[15]. These methods include complete records analysis, simple mean imputation methods, last observation carried forward, and creating an extra category for missing values. In a complete records analysis, only patients with complete data on all variables are included in the analysis. However, this method is only generally valid if the patients with complete data are representative of all patients, i.e. missing completely at random (MCAR). This assumption is usually implausible in primary care data because clinicians often perform more tests or record more data for high disease risk patients. Therefore, patients excluded from the analysis due to missing data may be systematically different from those included. Although easy to apply, complete records analysis and other 'ad–hoc' methods, such as last observation carried forward (the last observed response replaces the following missing response) and missing data category (creating an extra category for the missing observations) may bias estimates and standard errors. Therefore, these methods can result in misleading conclusions, because they do not appropriately consider the data structure or reasons data are missing and make implausible assumptions about the missing data[16, 17].

An alternative approach for handling missing data is multiple imputation (MI)[18], increasingly regarded as the standard method for handling missing data in medical research because of its practicality and flexibility[19]. MI creates multiple imputed datasets. In each, missing values are imputed from the conditional distribution given the observed values using an imputation model. The originally intended analysis is applied to each imputed dataset and the results combined using Rubin's rules[18]. When the imputation model is correctly specified and the missing at random (MAR) assumption is plausible, MI gives unbiased estimates and standard errors, leading to valid conclusions [20].

The original implementation of MI used a joint multivariate normal imputation model [18]. However, this approach is rarely suitable for imputing missing data in longitudinal, electronic, health records. Usually the joint multivariate normal imputation model is difficult to define or does not exist with ordered variables, and also inappropriate for binary and unordered categorical variables[21]. In this situation, fully conditional specification (FCS)[22] is a flexible alternative, which specifies a separate imputation model for each variable with missing data conditional on all other covariates. Until now, very few imputation approaches are sufficiently flexible to account for the longitudinal and dynamic structure of electronic health records. Biased estimates and standard errors can arise if the imputation model is incorrectly specified[19, 23] when imputing longitudinal patient records where patients enter and leave the database at different times. In longitudinal electronic health records where

measurements are recorded at different times, health indicator measurements may be aggregated into time blocks (e.g. one month or one year blocks) and treated as distinct variables at each time block. One approach is to use FCS to impute missing data at each time block separately conditional on other measurements recorded in the same time block. However, correlations between health indicators measurements at different time blocks will not exist in the imputed data. A second approach is to impute simultaneously all health indicators at all time blocks. This approach may converge with a few time blocks. However, in electronic health records with many health indicators and time blocks, this may potentially cause convergence problems due to co-linearity and over-fitting because of high correlations between repeated measurements of the same health indicator recorded at different time blocks. For example, co-linearity may occur if patients have the same smoking status records for many time blocks. Both the first approach, imputing each time block separately, and the second approach, including all time blocks in the imputation model, do not consider temporal ordering of the variables. The second approach does not exploit the potential simplifications achieved by considering the temporal ordering of the variables.

To address the issues described above, Nevalainen *et al.*[1] proposed a new MI approach, the two-fold FCS algorithm, to impute missing values in longitudinal data with an intermittent (non-monotone) missingness pattern. The two-fold FCS algorithm visits each time block sequentially. Within each time block, a MI procedure is followed to impute conditional on observed measurements and current imputations of missing values at the same time block and the adjacent time blocks. When all time blocks are imputed, one among-time iteration is complete. Within-time iterations and among-time iterations are repeated until estimates and standard errors become stable (i.e. converged). When completed, imputed datasets are generated for statistical analysis. Conditioning on measurements at adjacent time blocks takes into account the temporal ordering of the variables, while potentially accurately estimating correlations in the imputed data and avoiding convergence problems. Nevalainen *et al.* validated this algorithm using data simulated from a case-control study with measurements collected at three time blocks and up to 40% missing values. However, it is not known how well it will perform in the more complex setting of longitudinal, electronic health records with longer follow-up time, dynamic entry and exit from the database and more missing data.

The aims of this thesis are to (a) explore the extent of the missing health indicator data in THIN and consider removing outlier values from the data; (b) investigate potential reasons why the data are missing; (c) determine if MI is valid (i.e. the MAR assumption is plausible); (d) develop and evaluate the two-fold FCS algorithm to impute missing data in longitudinal, clinical databases, including producing a software package to apply the two-fold FCS algorithm, and (e) compare to standard methods (such as complete records analysis) for handling missing data in THIN.

Chapter 2

Aims and objectives

The overall aim of this PhD is to adapt and evaluate the two-fold fully conditional specification algorithm and implement it to impute missing data from large primary care database to account for the longitudinal and dynamic structures of these data.
The specific objectives are to:

1. investigate the extent and patterns of missing data in a longitudinal clinical database for health indicators associated with risk of cardiovascular disease;
2. develop methods to identify and remove outliers before imputation;
3. adapt and develop the two-fold FCS multiple imputation algorithm to impute missing values in longitudinal clinical data for variables associated with cardiovascular disease;
4. validate the two-fold FCS algorithm through challenging simulation studies; and
5. apply the two-fold FCS algorithm in THIN to (i) model risk for cardiovascular disease and (ii) understand factors associated with greater total cholesterol reduction in patients with type II diabetes.

2.1 Outline of the thesis

The plan for the thesis is as follows:

In the following chapter, I describe primary care databases, and specifically THIN, explain the reasons health indicators recording in primary care databases change over time and discuss the difficulties which arise when data are missing.

In section 3.2 I outline missing data theory, including the different missingness mechanisms and simple 'ad–hoc' methods routinely used for handling missing data. 'Ad–hoc' methods can lead to biased results because assumptions they require for valid inference are implausible.

An alternative method, multiple imputation (MI), is preferred to 'ad–hoc' methods because it gives unbiased results and reduces the uncertainly due to missing data under more general, plausible assumptions about the

missingness mechanism. In section 3.3, I describe MI, discuss the issues when applying MI to longitudinal clinical databases and introduce the two-fold FCS algorithm.

I describe the data included in this thesis in section 4.1. To justify using MI, I need to understand the structure and extent of missing data and, because I focus on longitudinal data, how this changes over time. In section 4.2, I investigate how recording changed by age and sex and, in section 4.3, over time. Outliers can potentially bias imputations so, before using MI, I remove outliers using a new method I developed to identify outliers in longitudinal clinical data, described in section 4.4.

Initially, I evaluate the two-fold FCS algorithm using simulated data to assess bias in different settings, described in section 5.1. In section 5.2, a second simulation study evaluates the appropriate approach using the two-fold FCS algorithm to impute missing total cholesterol measurements.

I implement the two-fold FCS algorithm, based on findings from the previous chapters, to impute missing values in THIN data and compare to other imputation methods, described in section 6.1.

Section 6.2 describes another study investigating total cholesterol reduction after initiating statin treatment for patients diagnosed with type 2 diabetes in THIN, using the two-fold FCS algorithm to impute missing values.

The discussion, chapter 7, outlines the findings in this thesis, implications of this study, the potential for applying this method to impute missing values in similar longitudinal, clinical databases and possible future work.

Chapter 3

Background

In this thesis, I investigate methods for handling missing data in longitudinal, clinical data, using a primary care database. Therefore, in this chapter, I introduce primary care databases, describe how recording of health indicators in primary care databases changes over time and explain the difficulties of analysing databases with missing data.

To understand the missing data structure, in this chapter I outline missing data theory including different missingness mechanisms. I describe simple 'ad–hoc' methods routinely used for handling missing data and their limitations and discuss handling missing data using the more suitable multiple imputation method. However, limitations can arise when applying multiple imputation to longitudinal data. I discuss these limitations and introduce the two-fold fully conditional specification algorithm, which I evaluate later in the thesis, as a possible alternative to standard multiple imputation approaches to impute missing data in longitudinal databases.

3.1 Primary care databases and missing data issues

In this section, I describe The Health Improvement Network (THIN) primary care database, explain the reasons for missing data and why health indicator recording changed over time.

3.1.1 Primary care databases

In the UK, several electronic databases exist which contain data collected from each patient consultation with a General Practitioner (GP) or other health care staff in the primary care setting. Patient information is collected at irregular times from the point of registration with the practice to the time they leave the practice, providing a longitudinal health record. The three largest databases in the United Kingdom (UK) are The Health Improvement Network (THIN)[11], the Clinical Practice Research Datalink (CPRD) (formally General Practice Research Database)[24] and QRESEARCH[25]. Available data included patient characteristics (for example, age and sex), medical records (symptoms and diagnoses), prescription information, referral to specialists, laboratory results, some lifestyle characteristics (for example, smoking status and alcohol consumption) and other health indicator measurements taken in the GPs practice (for example weight, systolic blood pressure and total cholesterol).

3.1.2 The Health Improvement Network

For this PhD project, I analysed data from The Health Improvement Network (THIN) database, a large, longitudinal, clinical primary care database widely used in epidemiological research[11]. General practices joined the THIN Quality Data recording scheme, administered by Cegedim Strategic Data (CSD), and collected data using the Vision practice management software. Medical events were coded using the hierarchical Read system of coding[26] and prescriptions were coded using multilex encrypted ID codes from the UK Prescription Pricing Authority and classified by chapter in the British National Formulary (BNF)[27]. Information on area deprivation was available as quintiles of Townsend scores[28] of deprivation, based on the patients electoral ward from the 2001 Population Census (`http://www.statistics.gov.uk/census2001/census2001.asp`).

The patient data within these practices were approximately representative of the UK population[29]. All data collected were anonymised at source before leaving the GP system and continually updated. In January 2014, THIN contained records from over 11 million patients enrolled to 562 practices, including 6.2% of the UK population. The database was created in 2003, but data for some practices dates back to the early 1990s. Recording of both consultations and prescriptions were similar to national consultation and prescription statistics[30, 31]. The data provider, CSD, obtained overall ethical approval to use THIN in scientific research from the South East Medical Research Ethics Committee (MREC/03/01/073). The 'Missing data imputation in clinical databases: development of a longitudinal model for cardiovascular risk factors' project received scientific approval from the scientific review committee (SRC)

3.1.3 Data recorded in primary care

One major limitation when analysing primary care data is the substantial missing data because, although the databases were designed for research, the data were collected for clinical management. Data collected for research purposes, such as a clinical trial or cohort study, are more likely to be recorded at regular intervals during follow-up compared to data collected for clinical management. However, in primary care, recording is intermittent (see section 3.2.2). Hence, patient information is primarily recorded as required to diagnose or monitor a condition or disease when the patient attends a GP consultation. A missing value suggests (i) the patient did not consult with a GP, (ii) the patient consulted with a GP but the value was not measured, or (iii) the patient consulted, the value was measured but not recorded. However, the data are not missing because the GP requested certain information which the patient did not provide. Therefore, the missing data may not technically be 'missing' because it was usually not intended to be recorded. For the purpose of this study, we assumed the true values would have been observed if the GP requested them for all patients so we could investigate existing methods for handling missing data.

3.1.4 Improving recording in primary care

Over the past 20 years, the National Health Service (NHS) implemented some initiatives to encourage GPs to regularly record health indicators, which included height, weight, systolic blood pressure measurements, smoking status and alcohol consumption. In 1990, the NHS Department of Health (DoH) contract between GPs and the

government was introduced[32]. The contracts aimed to improve quality of practice by linking pay to performance. In 1992, the NHS DoH contract included New Patient Health Checks[12] to encouraged GPs to record health indicators for all newly registered patients. These incentives stopped when the DoH introduced a new contract in 2004. However, many practices continued to perform new patient health checks as they often provide useful baseline information and an opportunity for GPs to identify any patient health problems. Marston *et al.*[13] investigated health indicator recording in newly registered patients from 2004 to 2006 in THIN and found, despite improved recording, nearly 40% of newly registered patients with missing health indicators records in the first registration year[13].

The revised 2004 DoH contract introduced the Quality Outcomes Framework (QOF)[14]. Under the scheme, practices received points for good quality of care, shown by recording health indicators required to monitor specific diseases. For example, GPs are required to record blood pressure measurements, total cholesterol measurements and smoking status every 15 months for patients diagnosed with coronary heart disease (CHD). It is in the practices interest to gain as many points as possible, not only for financial reward but also to achieve good patient management. Therefore, health indicators associated with risk of the diseases specified by QOF were recorded more frequently after introducing this legislation[13, 8].

3.2 Missing Data Theory

In this section, I introduce missing data mechanisms and discuss 'ad–hoc' methods, frequently used in practice for handling missing data, and their limitations.

3.2.1 Issues which arise from missing data

Missing data potentially biases statistical analyses, resulting in misleading conclusions[33]. Before considering an approach for handling missing data, first I consider different missingness mechanisms, which describe the missingness pattern in the data. Later in the thesis, I investigate which missingness mechanisms I can assume exist in the data and apply the appropriate approach for handling missing data, under this assumption, to achieving unbiased statistical analysis.

3.2.2 Missing data mechanisms

Below I describe different non-responses which exist in epidemiological and medical data and give examples from primary care data[34]:

- patient non-response - only very limited information known about a patient. For example, if a patient recently registered with a practice, we know address, date of birth and sex, but the patient did not consult so no health indicator measurements were recorded.
- item non-response - incomplete patient data. For example, the patients weight and blood pressure were recorded at regular time intervals, but not total cholesterol measurements.
- wave non-response - specific to longitudinal studies and occur when (whether by design or not) patients moved into and out of the study over time. This does not apply directly to primary care because data were not collected at fixed time points. However, usually patients did not consult for a period of time, so no information was recorded for them until they attended a consultation. Also called intermittent (non-monotone) missingness.
- attrition - again specific to longitudinal studies, refers to patients who dropped out and never returned. Attrition occurred in primary care when a patient either died or transferred out of a practice. After this time, no further information was collected. We cannot always distinguish patient attrition from wave non-respondents until death or the patient consults. Also some patients may no longer attend the practice for consultations, but are still registered. Attrition, also called monotone missing, commonly occurs in clinical trial data[15].

From the examples described for each non-response type, all of them were observed in THIN. We should also consider the process which introduced the missing data, or 'missingness mechanism'. To explore this further, consider a regression setting where Y denotes the outcome of interest, X denotes the covariate variables with missing values and $R = 1$ when X observed and 0 if missing. Below, I describe the three missingness mechanisms[17, 15, 18, 33]:

1. Missing completely at random (MCAR): the reasons for missing data do not depend on the observed or the missing values. For example, if a letter from the hospital with blood test results are lost, the chances this random event occurs is the same for all patients regardless of patient characteristics or outcome. The average outcome effect is the same among those with and without missing data.

   More formally, under MCAR, the chance $R = 1$ given $X; Y$ - using the notation $f_R[R|X; Y]$ - does not depend on X or Y, so $f_R[R|X; Y] = f_R[R]$. Therefore, the distribution of X given Y does not depend on R. Using the definition of conditional probability:

   $$\begin{aligned} f_R[X|Y; R] &= f_R[X; Y; R]/f_R[Y; R] \\ &= (f_R[R]f_R[X; Y])/(f_R[R]f_R[Y]) \text{ (because of the MCAR assumption)} \\ &= f_R[X|Y] \end{aligned} \tag{3.1}$$

   If this assumption is plausible, analysing observed data achieves unbiased but less precise estimates because patients with missing data are excluded from the analysis. The observed data may suggest a plausible MCAR mechanism, because none of the variables appear associated with data being missing, but we cannot definitely conclude a MCAR mechanism from just the observed data. However, we can prove the MCAR mechanism is implausible. For example, blood pressure values are not MCAR if recorded more often in men than women.

2. Missing at random (MAR): the reasons for missing data depend on the observed values but not the missing values. Or, we consider data MCAR conditional on the observed groups (i.e. observations within the same groups with the same probability of X missing). Within these groups, we can obtain marginal estimates from the observed data and average across the groups to achieve unbiased results.

   For example, blood pressure is missing for some patients, but recorded more frequently for female patients compared to male patients. Therefore, the probability of missing blood pressure is different for female and male patients. In this situation, blood pressure is MCAR conditional on sex. Other variables, for example age, may also determine blood pressure being recorded. Therefore, we can condition on several variables to ensure a MCAR missingness mechanism data within each strata.

   A similar argument gives the conditional probability (3.1) defined above if we assume a MAR missingness mechanism because $R = 1$ does not depend on X once we condition on Y, so $f_R[R|Y; X] = f_R[R|Y]$. Therefore, if we assume data MCAR or MAR, the distribution of variables with missing values conditional on the fully observed variables is the same for all patients.

3. Missing not at random (MNAR): The reasons for missing data depend on the underlying missing data values even after conditioning on the observed data. For example, systolic blood pressure is recorded more frequently for patients with a high systolic blood pressure.

   MNAR is the most difficult mechanism to achieve unbiased estimates and standard errors (SE), because now if X is missing, the distribution of X conditional on Y varies dependent on observing X. For valid

inferences, we require additional information or assumptions to describe the relationship between the probability of missing data and its missing value, not available from the observed data.

In other words, the conditional probability (Equation 3.1) defined above does not hold if a plausible MNAR missingness mechanism exists. $R = 1$ depends on both X and Y so the distribution of $f_R[X|Y]$ is different if X is observed or not (i.e. whether $R = 1$ or not). Therefore, handling data with a MNAR missingness mechanism is complex, as we either need to know exactly how R depends on $X; Y$ or exactly how $f_R[X|Y]$ differs according to R.

The different 'missingness mechanisms' are assumptions about the data to justify applying a specific imputation approach for handling missing data, and not a property of the data[19]. We can make plausible assumptions regarding the 'missingness mechanism' based on prior knowledge of the data, not just from the observed data, to provide additional evidence for a specific mechanism[34].

In real life, 'missingness mechanisms' are a continuum between MAR and MNAR. The three categories of missingness are not mutually exclusive, all can occur in datasets like THIN. MCAR, pure MAR, and pure MNAR do not exist because the pure form requires almost universally untenable assumptions. All missingness is MNAR to an extent (i.e., not purely MAR). We need to understand the reasons why data are missing and which assumptions are plausible before performing analyses on datasets with missing data. However, rather than focusing on whether the assumptions are violated, instead we should consider if the violation is big enough to matter to a practical extent[35].

Of the three missingness mechanisms, MAR provides a reasonable primary working assumption for an analysis of data with missing values. It assumes the distribution of partially observed variables given the fully observed variables is the same for all individuals. Given a sufficiently large dataset with many covariates like THIN, this seems a plausible starting point to impute missing values in THIN.

3.2.3 'Ad–hoc' methods for handling missing data

Various, commonly used 'ad–hoc' methods exist for handling missing data, discussed below[15, 16, 17]:

1. Complete records analysis: we only include patients with complete data for all variables in the analysis. A popular method for handling missing values because standard statistical software automatically excludes missing data and performs complete records analysis.

   Under a MCAR missingness mechanism assumption, this method achieves unbiased estimates, but we always reduce precision because we exclude records. However, the MCAR assumption is rarely plausible, but sometimes justified in its simplicity if it introduces minimal bias and precision reduction. Although, if missing data are MNAR, even a small proportion of missing data may bias estimates and SE.

   We can achieve unbiased complete records analysis if the explanatory variables missing values are not missing conditional on the outcome given the covariates[33, 36, 37], discussed in more detail in section 3.3.7.

2. Last observation carried forward (LOCF): the last observed response replaces the following missing response. This method handles loss to follow-up. LOCF is often applied to longitudinal clinical trial data, but also used in cohort studies.

   A single value replaces each missing response for either outcome or exposure variables, used as an estimate of a distribution. The subsequent analysis gives imputed responses the same status as actual observed responses which underestimates the SE. A LOCF analysis is equivalent to analysing the last measured responses. One study suggests regression coefficients are not just biased under LOCF but test of treatment effect may suffer from greatly inflated Type I error rates (null hypothesis is accepted when it is false) and confidence intervals may have a far from the nominal coverage probability[38]. LOCF is based on strong and often implausible assumptions about the data, such as the status of patients is unchanged. However, patients who leave the study are possibly different from those that stay. Also, if loss to follow-up is informative, LOCF is biased[39].

3. Missing category method: creating an extra category for the missing observations. Dissimilar classes can be grouped into one category which is not meaningful and the results are difficult to interpret and possibly biased[40]. In multivariate analysis, it is difficult to explain the relationship between a missing indicator category and other variables. To illustrate the drawbacks of the missing category method, we generated different systolic blood pressure distributions for patients with each smoking status category (Figure 3.1 - left). We randomly selected 30% and changed the smoking status to missing. If we group together all patients missing smoking status into a 'missing' category, the systolic blood pressure distribution in this category is not similar to any one of the smoking status categories because it includes patients from all three categories (Figure 3.1 - right). This 'missing' category is difficult to interpret because the measurements don't belong together.

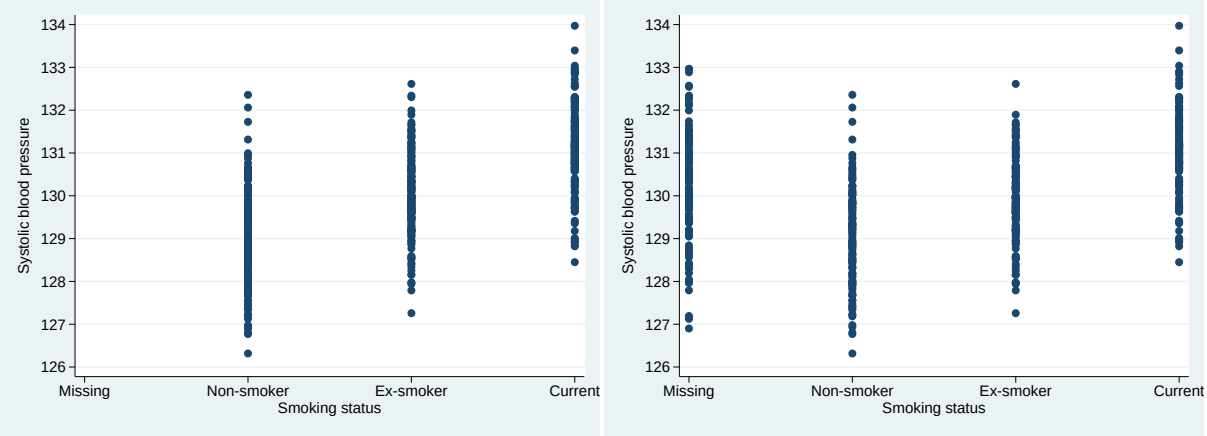

Figure 3.1: Distribution of systolic blood pressure measurements for each smoking status with missing data category

4. Marginal mean imputation: missing values imputed with the average of the observed values for that variable. This method is difficult to apply to categorical variable where the 'average category' has no meaning, like sex. This method ignores all the other variables in the dataset, reducing the variation in the data. Also,

imputing missing data to the same value will underestimate the marginal SE and reduce the association between variables. To illustrate the drawbacks of the marginal mean imputation method, we generated systolic blood pressure distribution for patients (Figure 3.2 - left) and changed 30% of systolic blood pressure values to missing. Next, we replaced the missing values with the mean of the original systolic blood pressure distribution. Now this distribution spikes at the mean (Figure 3.2 - right), unlike the original distribution. Even though the mean systolic blood pressure stays the same for both distributions, but the SE reduces from 0.06 to 0.05 because of this spike at the mean.

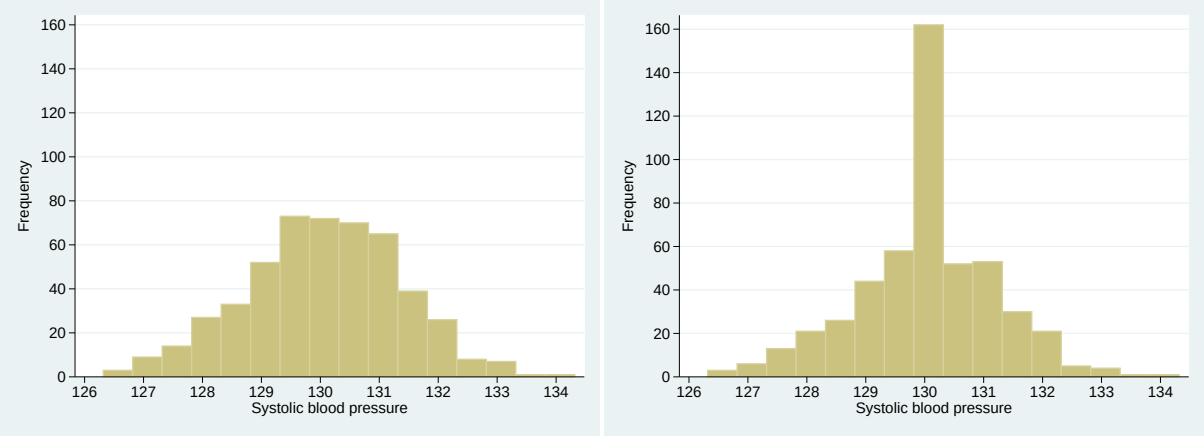

Figure 3.2: Distribution of systolic blood pressure measurements before and after marginal mean imputation

5. Conditional mean imputation: we fit a regression model to the complete data to obtain unbiased mean estimators under the MAR assumption. Even though this method is an improvement over marginal mean imputation, it still underestimates SE and gives the impression of too precise estimates because it treats the imputed values as 'true' values.

6. Hot deck imputation: the data is divided into subgroups and missing values are selected with replacement from the observed measurements (donor values) in the same subgroup. The hot deck literally refers to the deck of matching computer cards for the donors available for a missing value (i.e. this is a non-parametric method)[15]. Potential difficulties arise when deciding how to form appropriate donor groups. Even if we select valid donor 'decks' and achieve unbiased estimates, we will underestimate the SE because we treat the donor values as 'true' values[20].

Some of the 'ad–hoc' methods achieve unbiased estimates if the appropriate assumptions are plausible but they do not correctly account for the uncertainty due to the missing data and will underestimate the SE, potentially leading to a Type I error[41]. Except for complete records analysis which overestimates the SE. We can adjust the SE, but the conditions for applying these adjustments are restrictive[15]. Frequently, 'ad–hoc' methods are applied and data is analysed before considering if the underlying assumptions are plausible, but we should first make assumptions about the data before performing the appropriate analysis under those assumptions.

3.2.3.1 Maximum likelihood

Alternatively, we can consider the non-'ad–hoc' maximum likelihood (ML) approach for handling missing data. With or without missing data, we first construct the likelihood function. To calculate the ML estimates, we find the parameter values which make the likelihood function as large as possible and integrate the likelihood function over the variables with missing data, obtaining the marginal probability of observing the observed variables[42]. Estimates obtained through maximum likelihood are asymptotically unbiased if missing data are MAR and the model is specified correctly[43]. We can obtain SEs directly because a by-product of the maximisation is a numerical approximation to the matrix of second derivatives at the maximum, i.e the SEs. ML can obtain estimates and SEs for a linear regression of Y on binary X when X has missing values.

Sometimes we cannot calculate the first or second derivatives of the log likelihood, for example when the analysis includes many covariates or non-monotone missing data (and no direct way to incorporate auxiliary variables). In this situation, a useful alternative is the expectation maximisation (EM) algorithm, an iterative procedure which uses ML to obtain a mean vector and covariance matrix, so does not require the first or second derivatives of the log likelihood. The EM algorithm relates the ML estimation of a parameter θ from the log-likelihood of θ given the observed values to the ML estimation based on the complete-data log-likelihood[15].

Expectation: conditional expectation (i.e. mean) of the missing data given the observed and current estimated parameters are found and missing values are replaced with the expectation.

Maximisation: perform ML estimation of θ as if there were no missing data.

The disadvantages of the EM algorithm are:

1. difficult to produce general code;
2. it can be slow to converge, especially with many missing values, many covariates or non-monotone missing data;
3. difficult to obtain SE directly; and
4. the maximisation step is complex (i.e. has no closed form).

Alternatively, instead of the EM algorithm, we can use full information maximum likelihood (FIML), a parameter estimation method which appropriately incorporates missing data. It directly maximizes the likelihood for the specified analysis model rather than doing the two steps like the EM algorithm, giving estimates and SEs[42]. Applications which use FIML, such as structural equation modelling, allow models to include auxiliary or latent variables[44], which can improve accuracy and precision of estimates. However, FIML may not be computationally feasible when analysing databases like THIN because, with many observations and time points, these complex models may not converge.

3.3 Multiple Imputation

The previous section described missing data theory and some frequently used 'ad–hoc' methods for handling missing data. In this section, I describe multiple imputation (MI) and explain why this approach is more suitable than 'ad–hoc' methods. I emphasise the limitations of applying MI to longitudinal clinical databases. Finally, I describe a new approach, two-fold fully conditional specification MI algorithm, and explain why this method may be more suitable for handling missing data in longitudinal clinical databases compared to other imputation approaches.

3.3.1 Introduction to multiple imputation

The key step in MI is to fit our model of interest to the data to obtain unbiased, precise estimates. To achieve this, we must correctly estimate the distribution of the variables with missing values conditional on the observed data and account for the uncertainty due to the missing data (rather than estimating the exact values of the missing data, which we can never know).

We can use MI to create multiple imputed datasets: say we have a model of interest and a dataset we wish to fit this model to, but some values of explanatory variables in the model are missing. First, we specify the imputation model (different from the model of interest). MI imputes the data by selecting draws from the predictive distribution of the missing data given the observed values. Each imputed dataset is 'complete'. We obtain unbiased estimates and SEs when we fit the model of interest to each imputed dataset and correctly combine the results, provided: (i) the MAR assumption is plausible and (ii) the model of interest and in the imputation model are congenial (i.e. the imputation models conditional distribution is the same as the model of interests predictive distribution)[20].

To implement MI, we need a model for the partially observed variables given the fully observed (i.e. the imputation model). We can either achieve this directly, using a joint model, or approximate this process indirectly using a sequence of conditional models, called fully conditional specification, which I describe in more detail below[15].

The direct approach proceeds as follows; fit a joint, multivariate normal distribution to the data conditional on observed variables in the imputation model, draw random values from the joint conditional distribution and replace the missing values with these draws (imputed) values.

We adapted the following technical justification from Rubin's Multiple Imputation for Non-response in Surveys, 1987[18] and applied the theory to a clinical data setting.

Y refers to outcome variables $j = 1, \ldots, p$. These variables are fully observed for all N patients in the data:

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix}$$

where Y_i is a row vector for patient i with p fully observed covariates. X refers to $k = 1, \ldots, q$ covariates, or predictors of the outcome (e.g. age, weight, blood pressure, etc.), with missing values recorded for all $i = 1, \ldots, N$ patients in the database

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_N \end{pmatrix}$$

where X_i is a row vector for the i^{th} patient and q covariates with missing values.

$$R = \begin{pmatrix} R_1 \\ \vdots \\ R_N \end{pmatrix}$$

is the response indicator for observed or missing values where R_i is a row vector for patient i and q covariates with missing values. R_{ij} is binary and $R_{ij} = 1$ indicates an observed value in X for variable j recorded for patient i and $R_{ij} = 0$ indicates a missing value for X.

3.3.2 Bayesian inference

Q=Q(Y,X) is an estimation of a function of Y and X. With scalar X_i, a common quantity of interest is the population mean $\bar{X} = \sum_1^N X_i/N$.

Let $obs = [(i,j)|R_{i,j} = 1]$, where $i = 1, \ldots, N$ refers to patient i in the database and $j = 1, \ldots, p$ refers to variable j, such that X_{obs} indicates the observed measurements of X in the database. Therefore, $mis = [(i,j)|R_{i,j} = 0]$ such that X_{mis} indicates the missing measurements of X in the database.

A Bayesian inference for $Q(Y,X)$ follows from its posterior distribution. Its conditional predictive distribution given the observed values (Y, X_{obs}) is calculated under specified models. The posterior distribution is $Pr(Q|Y, X_{obs})$. As $Q(Y,X)$ is a function of the observed values (Y, X_{obs}) and missing values in X_{mis}, the posterior distribution of Q is calculated from the observed values and the posterior distribution of X_{mis}:

$$Pr(Q|Y, X_{obs}) = \int Pr(Q, X_{mis}|Y, X_{obs}) dX_{mis}. \tag{3.2}$$

General expressions for the posterior distribution of X_{mis} are less complex than those for Q because they do not involve the integration in equation (3.2). Therefore, we consider the posterior distribution of X_{mis} instead of Q. The posterior distribution of Q is obtained from (3.2), often evaluated directly. If we can assume the missing data are MAR, we do not condition on R when drawing X_{mis}.

3.3.3 Maximum likelihood v.s Bayesian

The frequentist theory of maximum likelihood (ML) assumes large samples. If this assumption is implausible, we can use Bayesian methods and obtain inferences using the exact posterior distribution for a particular choice

of prior. However, inferences with small samples are more sensitive to the choice of prior distribution than inferences with large samples. When prior information is available, better inferences can result from formally incorporating this information. When prior knowledge is limited the Bayesian approach with disperse priors often gives better frequentist inferences than large-sample approximations. For complex problems involving missing data, the Bayesian approach is attractive as it provides answers in situations where no exact frequentist solutions are available. The Bayesian answer, once derived, can be evaluated from a frequentist perspective. MI is based on Bayesian principles and in general MI under a realistic model has excellent frequentist properties[15].

3.3.4 Fully conditional specification (FCS)

To perform imputation we need a joint multivariate normal model for the variables with missing values conditional on the variables with no missing values. This model has parameter η say. We can estimate η, under the MAR assumption, and impute the missing values, taking full account of the uncertainty due to the missing data. In many cases, full joint multivariate normal model is difficult to define or is not multivariate normally distributed, i.e. not congenial with the model of interest. These problems arise for the following reasons[21]:

- More than one variable is imputed;
- The predictor has missing values;
- Co-linearity occurs, especially with many variables and few observations;
- Rows or columns are ordered, e.g. longitudinal data;
- Different variables types (e.g. binary, unordered categorical, ordered categorical, continuous), so the application of theoretically convenient models, such as the multivariate normal, are theoretically inappropriate; or
- The relationship between the predictors is complex, e.g. non-linear, or subject to censoring.

Fully conditional specification (FCS), or chained equations, is a computationally convenient, indirect approach which approximates the joint multivariate normal model. This method implicitly imputes missing values using a series of conditional regression models[21], under a plausible MAR assumption and assuming the joint multivariate normal distribution exists, known as compatibility of the conditionals[20].

If the joint distribution assumed by FCS exists, this process is a Gibbs' sampler[45], an example of a 'Markov Chain Monte Carlo' (MCMC) algorithm[15]. The Gibbs' sampler generates a draw from the distribution $P(x_1, \ldots, x_p)$ of a set of p random variables $x_1, \ldots, x_p$ when draws from the joint distribution are more difficult to compute compared to draws from conditional distributions $p(x_j|x_1, \ldots, x_{j-i}, x_{j+1}, , x_p), j = 1, \ldots, p$. Initial values are chosen and new values sequentially selected from the conditional distribution. Under quite general conditions the sequence of iterates converges to a draw from the joint distribution of $X_1, \ldots, X_p$.

FCS proceeds as follows:

1. For each variable in the imputation model in turn, fills in missing values with randomly chosen observed values.

2. Discards the 'filled-in' values in the first variable and specifies a regression model for this variable conditional on other variables in the imputation model and replaces the missing values with random draws from the conditional distribution.

3. step 2 is repeated for each variable in turn and completes one 'cycle' once each variable is imputed.

4. step 3 is repeated cycles until estimates and standard errors become stable (i.e.convergence).

FCS is a flexible alternative MI approach when we cannot specify a convenient and realistic joint distribution[45] and requires less iterations compared to the joint model approach[21]. Despite a lack of theoretical justification, simulation studies showed this method works acceptably well[46]. Investigations of direct vs. indirect imputation found we can expect similar results from both[47]. However, recent work by Liu *et al.* defined sufficient conditions (including compatibility of the conditionals) under which, as the sample size tends to infinity, the stationary distribution of the Markov chain generated by the chained equations algorithm (assuming that this stationary distribution exists and the chain converges to it) converges to the posterior predictive distribution of the missing data implied by a joint Bayesian model[48]. Hughes *et al.* and Liu *et al.* both proved a 'non-informative margins' condition, which guarantees the imputed values obtained using chained equations (at convergence) are drawn from the posterior predictive distribution of the missing data implied by a Bayesian joint model if, together with the compatibility of conditionals (and assuming that the Markov chain generated by the chained equations converges to a stationary distribution), the joint prior distribution factorises into independent priors[48, 49].

3.3.5 Summary of Repeated Imputations

Analysing a single imputed dataset treats the imputed values as 'true' values, and without adjustments it cannot reflect variability under the imputation model. However, when the imputation model repeatedly draws imputations, we can combine the multiple complete data inferences to form one inference and increase the efficiency of the estimate over that obtained from a single dataset and properly reflect uncertainty due to missing data[15, 50]. This method to correctly combine multiple inferences, called Rubin's Rules[18], proceeds as follows:

If Q is the quantity of interest in the study, generally a k-dimensional row vector, X the database mean and $\bar{X}$ the population mean. If all data are observed, inferences for Q follow the distribution:

$$(Q - \hat{Q}) \sim N(0, U) \tag{3.3}$$

where $\hat{Q}$ is a statistic estimating Q, U is a statistic estimating the variance ($k \times k$ covariance matrix) of $(Q - \hat{Q})$, and $N(0, U)$ is the k-variate normal distribution with mean 0 and variance U. Q is a 1-1 function of the quantity of interest making the normal approximation reasonable.

We draw M sets of repeated imputations and construct m complete data sets, where $\hat{Q}_{\star 1}, \ldots, \hat{Q}_{\star M}$ and $\hat{U}_{\star 1}, \ldots, \hat{U}_{\star M}$ are the values of the statistics, $\hat{Q}$ and U for each imputed datasets.

The $\hat{Q}$ and U statistics provide the mean and variance of Q given (Y, X_{obs}, R). We combine the M repeated estimates and associated variances for Q as follows:

$$\bar{Q}_M = \sum_{m=1}^{M} \frac{\hat{Q}_{\star m}}{M} \tag{3.4}$$

be the average of the M estimates,

$$\bar{U}_M = \sum_{m=1}^{M} \frac{\hat{U}_{\star m}}{M} \tag{3.5}$$

be the average of the M variances, and

$$B_M = \sum_{m=1}^{M} \frac{(\hat{Q}_{\star m} - \bar{Q}_M)}{M-1} \tag{3.6}$$

be the variance between (among) the M estimates, where the superscript t indicates transpose when Q is a vector. The quantity

$$T_M = \bar{U}_M + (1 + M^{-1})B_M \tag{3.7}$$

is the total variance of $(Q - \bar{Q}_M)$.

Rubin[18] defined an imputation procedure with $m = \infty$ as 'proper' (properly reflecting uncertainty due to the missing data) for the set of complete data statistics $(\hat{Q}, U)$ when we satisfy the following conditions:

1. $\bar{Q}_\infty \sim N(\hat{Q}, U)$
2. $B_\infty \approx B$ where B is the variance of $\bar{Q}_\infty$
3. $\bar{U}_\infty \approx U$
4. B is stable under repeated sampling

Interval estimates and significance levels for scalar Q are formed using a t reference distribution with

$$v = (M-1)(1 + r_M^{-1})^2 \tag{3.8}$$

degrees of freedom, where r_M is the relative increase in variance due to missing data:

$$r_M = (1 + M^{-1})\frac{B_M}{\bar{U}_M}. \tag{3.9}$$

A $100(1-\alpha)\%$ interval estimate of Q is

$$\bar{Q}_M \pm t_v(\alpha/2)T_M^{1/2}, \tag{3.10}$$

where $t_v(\alpha/2)$ is the upper $100\alpha/2$ percentage point of the student t distribution on v degrees of freedom (e.g. if $v = \infty$ and $1 - \alpha = 0.95$, $t_v(\alpha/2) = 1.96$). With infinite imputations $m = \infty$, the total variance reduces to the

sum of the two variance components and the confidence interval is based on a normal distribution $v = \infty$[50]. Also, the significance level associated with the null values Q_0 is

$$Prob\{F_{1,v} > (Q_0 - \bar{Q}_M)^2/T_M\} \tag{3.11}$$

where $F_{1,v}$ is an F random variable on one and v degrees of freedom. The fraction of missing information about Q is:

$$\gamma_M = \frac{r_M + 2/(v+3)}{r_M + 1} \approx \frac{B_m}{\bar{U}_m + B_m}. \tag{3.12}$$

3.3.6 Specifying the imputation model

The MI procedure includes certain variables in the imputation model to fully account for uncertainty in predicting the missing values and create appropriate variability in the multiple imputed data[18]. Most importantly, all variables in the model of interest, including the outcome variable[51], must be included in the imputation model. It may appear circular to condition imputations on the outcome when the final objective is to regress the outcome on the full set of covariates: however, imputing draws from the conditional distribution of the missing covariates given the observed covariates and the outcome yields consistent estimates of the regression coefficients[15]. If the outcome variable is excluded from the imputation model, relationships with this variable may not exist in the imputed data because the imputations are generated assuming the variables are independent of the outcome[35].

In addition to the variables in the model of interest, we can obtain more precise estimates by including additional variables in the imputation model, known as auxiliary variables[17]. If we only include variables from the model of interest in the imputation model, the standard errors from analysing the imputed data are consistent with the same analysis using the complete records. MI cannot add any additional information compared to the observed records.

We can include auxiliary variables either predictive of the missing values and/or the probability of values being missing in the imputation model[20]. When MI fits a regression model to the variable with missing values conditional on other variables in the imputation model, only auxiliary variables predictive of the missing values will affect the draws from the conditional predictive distribution, not auxiliary variables only predictive of the probability of the values being missing[20]. Therefore, as far as computationally feasible, we should include as many auxiliary variables predictive of the missing values as possible[36]. However, including auxiliary variables predictive of the missing values but not the probability of those values being missing will improve efficiency, but not address bias[52]. We also want to include auxiliary variables predictive of the probability of values being missing because this will increase the plausibility of the MAR assumption and reduce bias[52]. Therefore, we should also include auxiliary variables predictive of both the missing values and the probability of values being missing to obtain unbiased estimates and SEs from statistical analysis of the imputed data[52]. Over-fitting (including redundant predictors) may reduce the precision of the final estimates to some extent but should not increase bias because imputations are Bayesian, so it may be preferable to over- rather than under-fit in the

imputation model[46], but a lack of correlation between the variables can bias the results towards the null[53]. However, MI tends to be forgiving when the model of interest and the imputation model are not congenial[50].

Schafer *et a.* recommended the imputation model should be rich enough to preserve the associations and relationships among variables in the model of interest and ensure compatibility[50]. van Buuren *et al.*[21] early work investigated FCS using a simulation study and found FCS produced essentially unbiased estimates with appropriate coverage when using incompatible models. However, more recent work by Seaman *et al.*[54] found FCS imputation model needs to be compatible, if the model of interest had non-linear or interaction relationships, to achieve unbiased estimates and SEs. To ensure compatibility, we can impute passively: impute the main effect and passively calculate the non-linear or interaction term. This method ignores the non-linear relationship or interaction when imputing other variables with missing data. A second method is to impute the non-linear or interaction term as 'just another variable' in the imputation model, and is the recommended method for linear regression with quadratic or interaction effect[54]. Another method, substantive model compatible FCS, uses rejection sampling to explicitly specify the imputation model for each variable with missing data[55]. When data were MAR, substantive model compatibility FCS gave less biased estimates compared to including the term as 'just another variable'[55].

Another consideration is when ratio variables are included in the model of interest. A recent study recommended entering the log–transformed ratio terms separately into the imputation model[56]. Entering them without the log-transformation can bias the results in some situations by imputing some very large or very small values outside the range of acceptable values, including negative values. Using a log transformation ensured imputed values were positive[56].

If using a model of interest, such as a generalized linear model, the outcome variable should be consistent in the imputation model[46]. If the model of interest is a time-to-event model, we include the length of time each patient is at risk and an event indicator in the imputation model. For a Cox model, White *et al.*[57] investigated different functions of the time-to-event variable and demonstrated that the Nelson-Aalen estimator for the cumulative hazard function and the event indicator included together in the imputation model provided the least biased results.

3.3.7 MI vs complete records analysis

A complete records analysis relies on unrealistic missing data assumptions (MCAR) and can biased estimates and reduce power, as explained in section 3.2.3. In most situations, the MAR assumption required for MI is a better approximation to reality than either the MCAR assumption[15] or assuming missingness is independent of the response given the covariates, as explained in the following paragraph. MI improves efficiency as well as reducing bias compared with complete records analysis[58]. However, due to the added complexity, MI analysis is more susceptible to human error (i.e. specifying an incompatible imputation model) than complete records analysis[19].

One study investigated the bias incurred for complete records and MI for different missingness mechanisms[37].

MI appeared superior to complete records across a wider range of settings[37]. MI was more robust to different missing data mechanisms compared to complete records in the bivariate case, though not in the univariate case. When the data was MCAR, MI was more efficient than complete records for estimating the coefficient of an incomplete variable X if X was strongly associated with outcome[37]. The complete records analysis was unbiased provided that missingness is unrelated to the outcome variable, given the covariates[37]. Therefore, we introduced bias in the complete records analysis if a covariate associated with missingness was excluded from the model of interest[37].

It is informative to perform both MI and complete records analysis. It bridges the gap between current and future practice[59] and reassuring if the results are similar. Although, where complete records and MI analyses give different results, it is important to try and understand why[19]. We should also consider subject-matter context to assess carefully whether MCAR, MAR, and/or other missing data mechanisms (in particular missing conditional on X) are plausible, and therefore whether complete records analysis or MI are valid. If it is plausible that both are valid, then MI is preferred because of its greater efficiency. If the missing data mechanism seems to fit neither approach perfectly, further investigations are required to understand the results from two approaches. While MI is more robust to departures from assumptions, it is not guaranteed and reporting findings obtained from data with substantial missingness in covariates should be appropriately cautious[37].

When data are missing for reasons beyond the control of investigators, the MAR assumption may not hold. Sometimes it is not possible to relax the MAR assumption in any meaningful way without replacing it with other equally untestable assumptions. However, in this situation, Schafer *et al.* recommends using principled methods like MI that assume MAR anyway because they will tend to perform better than 'ad–hoc' procedures[50].

3.3.8 How studies using primary care data handle missing data

In this section, I evaluate how recently published studies using primary care data report and handle missing data. Other papers[19, 60, 61, 62] report findings from systematic reviews exploring how studies handle missing data, but not in a primary care data context.

To keep this review contemporary I selected all papers from the last two years (from the period January 2011 to June 2013) using primary care data and reported how they handled missing data. However, this is a descriptive review rather than systematic or comprehensive review. A limitation of this is that I may not include all relevant studies. However, the studies I selected are typical of those using primary care data, which gives an idea of current practice.

I found 10 recently published papers which use primary care data I describe how these papers handle missing data issues (Table 3.1) to show the problems with commonly used methods.

Table 3.1: Summary of how 10 papers published from January 2011 to June 2013 using primary care data handle missing data

Paper	Description	Missing data	How missing data were handled
Currie *et al.* 2013[63]	Patients with type I diabetes in THIN	Not reported	Complete records
Ose *et al.* 2012[64]	Primary care data for patients with coronary heart disease from 8 European countries	849 patients out of 3,505 with missing data	Complete records
Hirst *et al.* 2012[65]	Patients with prostate cancer in GPRD	Not reported	Complete records
Currie *et al.* 2012[66]	Patients with type II diabetes in THIN	Not reported	Complete records
Collins *et al.* 2012[67]	Patients with moderate to severe kidney disease end-stage Kidney disease in THIN	Reported missing data for key variables, approximately 85% of patients missing for serum creatinine	Imputed missing smoking status, amount of cigarettes smoked, systolic blood pressure, BMI and serum creatinine using multiple imputation
Khan *et al.* 2011[68]	Data on patients with breast, colorectal and prostate cancer	Not explicitly reported but can calculated them from given information	Repeated analysis using complete records, multiple imputation and missing indicator category. All gave similar results so only report complete records
Vinogradova *et al.* 2011[69]	Patients with pneumonia prescribed statins in QRESEARCH	Missing Townsend deprivation score, BMI and smoking status	Repeated analysis using complete records and missing data category. Both gave similar results so reported only missing data category
Miller *et al.* 2011[70]	Patients with chronic obstructive pulmonary disease in GPRD	Missing BMI and/or smoking status	Not explicitly explained, but study appears to use missing data category
Vinogradova *et al.* 2011[71]	Patients with cancer prescribed statins in QRESEARCH	Missing Townsend deprivation score, BMI and smoking status	Used multiple imputation, complete records analysis and missing data category. All gave similar results so reported results from multiple imputation
Suissa *et al.* 2011[72]	Women with type II diabetes treated with insulin in GPRD	Not reported	Used complete records at baseline, missing data category during follow up

None of these studies explained if the necessary assumptions for each method for handling missing data were plausible, or how using these methods can potentially bias the results. Next, I discuss the studies that considered missing data in more detail.

Some papers attempted to justify the chosen method for handling missing data. For example, the aim of Ose *et al.*[64] study was to identify factors associated with health-related quality of life (HRQol, the discrepancy between actual and desired functional capacity) in primary care at practice and patient-level in patients with coronary heart disease. They used a multilevel analysis to identify associations with 14 potential explanatory variables and HRQof at patient and practice level. The authors excluded patients missing any explanatory variables. The authors did not report how many values were missing for each variable, but they did report that the HRQol was similar between those patients included and excluded from the analysis, which suggests the patients included in the complete records analysis were representative of the population.

Hirst *et al.* reduced the missing data by selecting a time period with more complete recording[65]. The aim of the study was to develop a method to more accurately identify castration resistant prostate cancer patients using primary care data. They investigated the incidence and prevalence of selected comorbidities before and after castration resistant prostate cancer. First, they performed a complete records analysis, next they stratified the analysis to before and after the introduction of QOF as a sensitivity analysis (see section 3.1.4). The authors reported the results were more 'reliable' (precise) when analysing data recorded after the introduction of QOF due to more complete recording.

Three studies compared different methods for handling missing data. Collins *et al.*[67] validated a risk prediction model, the QKidney score, for patients in a primary care database and evaluated its discrimination and calibration. Initially, they performed multiple imputation. One variable, serum creatinine is important for predicting kidney disease but, as this variable had substantial missing data, the authors created two scores, with and without serum creatinine. They explained multiple imputation and the imputation model used, but they only use variables from the model of interest in the imputation model. As discussed earlier, if no additional auxiliary variables are included in the imputation model the standard errors are consistent with a complete records analysis. Also, data were extracted for each patient over many years and it is not clear if they imputed missing values over time or just at one time point. Collins *et al.* compared the MI results to a complete records analysis and they found the same results for the model without serum creatinine but different results for the model with serum creatinine. They suggested it was due to a possible MNAR mechanism but did not explain why.

Khan *et al.*[68] investigated the burden of health outcomes associated with treatment of British long-term cancer survivors by finding the incidence of treatment associated health outcomes using a Cox proportional hazards model. The authors found similar results for complete records, multiple imputation and missing indicator category, so chose to report the complete records analysis. As explained earlier, it is preferable to report the results from MI because it is more efficient compared to a complete records analysis and the assumptions for MI are

more plausible compared to a complete records analysis. However, the authors did not explain MI, whether the MAR assumption is plausible or the imputation model used. If, like Collins *et al.*, they excluded auxiliary variables from the imputation model, this could explain why the results following MI are similar to a complete records analysis.

Vinogradova *et al.*[69] investigated pneumoinia risk in patients prescribed statins using logistic regression analysis, handling missing data using missing data category. The authors justify this method because the results using a missing data category were the same as a complete records analysis. Both these methods require strong, untestable assumptions, therefore, both methods showing the same results does not necessarily prove either method is correct. Both methods may be incorrect and could yield the similar (incorrect) results.

In a different study, Vinogradova *et al.*[71] also investigated cancer risk in patients prescribed statins and performed a logistic regression analysis, imputing missing data using MI. However, the authors did not explain MI, if the MAR assumption is plausible or describe the imputation model used so it is not possible to determine if MI was implemented correctly. The authors did compare the results to complete records analysis and found similar results, but did not report the complete records analysis results. The authors used missing data category as a sensitivity analysis, which also gave similar results. Again, MI and complete records analysis possibly gave similar results because the authors possibly did not include auxiliary variables in the imputation model.

This small review of studies using primary care data with missing values suggest that most of the studies do not fully consider the implications of missing data or how the methods used could affect the results. These studies also did not consider if the method for handling missing data is appropriate for longitudinal clinical data. The MI methods I investigate in this thesis could be used as an appropriate method for handling missing data in longitudinal clinical data: the studies reviewed in this section could use these methods to impute missing values to obtain unbiased results under a plausible MAR assumption.

3.3.9 Problems with imputation in longitudinal clinical database

For the reasons discussed in the earlier section 3.3.4 (i.e. longitudinal data), we chose not to use a joint multivariate normal distribution to impute missing values in THIN. Therefore, we investigated using FCS.

Say the longitudinal database consisted of $i = 1, \ldots, n$ independent subjects with $v = 1, \ldots, p$ variables and data collected at $j = 1, \ldots, q$ equally spaced time points. Let $X = (X_1, ..., X_q)$ denote the vector-valued random variable consisting of the q repeated measurements of the p explanatory variables $X_j = (X_{j1}, ..., X_{jp})$ and X^{obs} and X^{mis} denote the observed and the missing elements in X, respectively. Lastly, let Y denote the vector of outcome variables. To apply MI, we used a suitable imputation model:

$$f(X^{mis}|X^{obs}, Y, \theta) \tag{3.13}$$

where θ was a parameter describing the association between the Xs. We assumed a missing at random (MAR) missingness mechanism so the model (3.13) does not condition on the missingness pattern.

We considered correctly applying MI, and specifically FCS, to data with repeated measurements to account for the longitudinal and dynamic structure of the data. One approach was to impute missing values by fitting MI models separately to each time point. Using FCS, a series of $(v = 1, ..., p)$ conditional density models were defined at time j.

$$f(X_{jv}^{mis}|X_{j1}^{*}, ..., X_{j(v-1)}^{*}, X_{j(v+1)}^{*}, ..., X_{jp}^{*}, Y, \theta_{jv}) f(X_{jv}^{mis}|X_{j,-v}^{*}, Y) \tag{3.14}$$

where X^* refers to the observed and current imputed values of X.

If we imputed missing values from a large, dynamic longitudinal database with intermittent missingness[46] and a long mean follow-up time like THIN, the correlations between measurements recorded at different times may be distorted in the imputed data. Therefore, we wanted to impute each time point conditional on information at other time points. Another option was to impute conditional on past and future observations at $j = 1, ..., q$ time points.

If we included measurements of variables recorded at each time point as separate variables into the imputation model, imputations became computationally intensive and increased the possibility of co-linearity due to overfitting because, in large, dynamic database with many variables, each imputation model had $(p-1)+(q-1)\times p$ potential covariates. Also, this imputation model is potentially difficult to specify correctly because of the large number of variables and time points:

$$f(X_{jv}^{mis}|X_{1}^{*}, ..., X_{j-1}^{*}, X_{j,-v}^{*}, X_{j+1}^{*}, ..., X_{q}^{*}, Y) \tag{3.15}$$

3.3.10 Description of two-fold fully conditional specification (FCS) algorithm

Nevalainen *et al.*[1] proposed an alternative method as a compromise between 3.14 and 3.15. The two-fold FCS algorithm addresses the computational issue by restricting the imputation to a narrow time window. This method

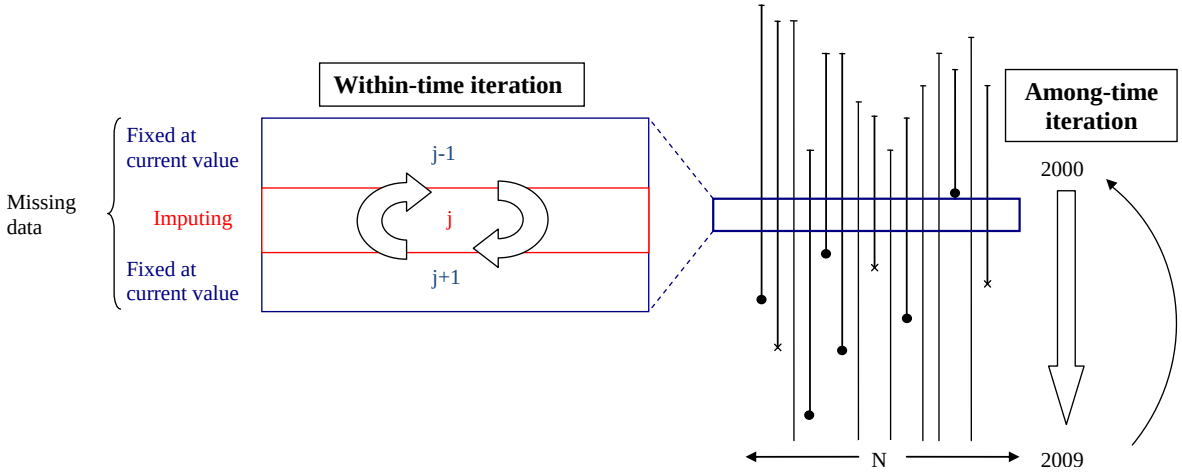

Figure 3.3: Illustration of the two-fold fully conditional specification (FCS) algorithm imputing simulated data. The vertical lines represent the time-lines of N patients in THIN. Patients enter (horizontal bar at the top of the line) and leave THIN at different times. A circle at the end of the time-line indicates the patient was transferred out of the practice, a cross indicates the patient died.

is an approximation to the 'full' imputation model given by equation 3.15. The two-fold FCS algorithm proceeds in two stages:

1. Within-time iterations: at time j. X_{jv}^{mis} is imputed conditional on
   $X_{j-1}^{*}, X_{j,-v}^{*}, X_{j+1}^{*}$
   and the outcome Y. Observed and previously imputed values adjacent to time j are predictors in the imputation model and not imputed at this stage. We specify b_w within-time iterations and one within-time iteration applies the FCS algorithm to all p variables at time point j (Figure 3.3).

2. Among-time iterations: The within-time iteration is repeated sequentially for all time points. When values were imputed at all times, an among-time iteration is complete. We specify b_a among-time iterations. Among-time iterations are required to ensure inter-correlation among repeatedly measured variables are maintained in the imputed data (Figure 3.3).

Restricting the imputation models to a small time window reduces computational problems due to over fitting, but may still consider the longitudinal and dynamic structure because the imputation model conditions on measurements at adjacent time points. However, the two-fold FCS algorithm assumes conditional independence, i.e. measurements outside of the time window are independent given that we are conditioning on measurements closer in time. The two-fold FCS algorithm also takes advantage of the intermittent missingness pattern because the imputation models condition on not only observed measurements before the time point of interest but also after the time point of interest.

The previous implementation of the two-fold FCS algorithm imputed time-dependent variables only and all subjects entered and exited the study at the same time points with only 3 time points[1]. For this PhD study we implemented a more flexible two-fold FCS algorithm which imputes time-independent variables, increase the

time window width and imputes missing data for subjects with different entry and exit time points. We assumed compatibility of conditionals (i.e. we can assume an underlying joint multivariate model exists) when using the two-fold FCS algorithm to impute missing values in longitudinal clinical data, and we attempted to correctly specify imputation models, to ensure congeniality. Under these assumptions, the analysis of data imputed using the two-fold FCS algorithm obtains unbiased estimates and unbiased SEs calculated using Rubin's rules[73].

3.3.11 Summary

In this chapter, I described the theory required to understand the results of the analysis in the following thesis chapters. I discussed the THIN data used in this thesis, particularly the issues raised by missing data. I discussed the concepts of missing data theory, including the limitations of frequently used 'ad–hoc' methods for imputation and why MI is a preferred imputation method. I described MI and I discussed issues that arise when imputing longitudinal data using existing approaches and I introduced the two-fold FCS algorithm and explained how this approach may overcome the limitations of the other approaches. Next, chapter 4 reports a descriptive analysis of health indicator recording in THIN so I can understand when and why values are missing and if the MAR assumption is plausible, which justifies using MI to impute missing values.

Chapter 4

Descriptive analysis of health indicator recording in The Health Improvement Network

Up to this point, this thesis focused on analysing data from primary care, such as The Health Improvement Network (THIN), and in particular the issues raised by missing data. I discussed how Rubin's missingness mechanism taxonomy applies in this context and described commonly used 'ad–hoc' methods, which are only valid when implausible assumptions about the missing data are true. I described multiple imputation (MI), which is valid under more plausible assumptions about the missing data compared to the 'ad–hoc' methods. I explained why certain MI approaches are not suitable to impute missing data in primary care databases and how the two-fold fully conditional specification (FCS) algorithm has potential to impute missing data in longitudinal, clinical database but not yet validated in this setting. Before I investigate using the two-fold FCS algorithm to impute missing data in THIN, I explore the extent and patterns of missing data in THIN for health indicators (such as weight, systolic blood pressure and total cholesterol) associated with risk of cardiovascular disease (CVD) to understand if the missing at random (MAR) assumption is plausible, and if it holds at all times of health indicator recording. In this chapter, I first describe the data extracted from THIN, and a few other data sources from the Health Survey for England, used in this thesis. Secondly, I describe how health indicator recording for patients varies by age and sex. Next, I describe how health indicator recording changed over time and in particular how National Health Service (NHS) legislation influenced changes in health indicator recording. These investigations allow me to evaluate the plausibility of the MAR assumption for health indicators recorded in THIN and how this changes over time. Incorrectly recorded values (i.e. outliers), which can adversely influence imputations, are a common occurrence in primary care data. Therefore, in my final investigation before using MI, I introduce a new method to identify and exclude outliers in longitudinal data.

4.1 THIN data included in this study

Many general practices became computerised in the 1990s to manage patient information more effectively, starting with Vamp software before gradually changing to Vision software after 1995 (Figure 4.1). However, practice mortality was inaccurate during the time immediately after changing to new software, which raised a few data quality issues because practice mortality rate is derived from mortality data. The inaccurate mortality rates

occurred because practices excluded patients from the transfer to Vision if they died before practices changed to Vision. Therefore, the mortality rate was artificially reduced when practices changed to Vision software, and increased over time as patients who transferred to Vision died. One study examined the mortality recording in these practices and derived a marker, the Acceptable Mortality Recording (AMR) date, to indicate when practices recorded a mortality rate similar to national statistics[74]. Analysing data recorded before the AMR date biases mortality estimates because the patients mortality rate did not represent the general UK population.

Another issue which arose from practices becoming computerised was inaccurate data recording because some practices still used paper records at the same time as electronic records. Horsfall *et al.* examined this issue and derived a marker, the Acceptable Computer Usage (ACU) date, to indicate the earliest date when the average annual recording rate included at least one medical record, one additional health data (AHD) record and two prescription records per patient[75]. Horsfall *et al.* showed that the ACU date in combination with the AMR date produced incidence of disease and prescribing trends consistent with external data sources[75]. Also, using only the ACU date did not capture sudden peaks in mortality recording due to data updates or conversions, resulting in periods of low mortality included in the analysis. Therefore, I included patient records in this study after the latest of the AMR date and ACU date.

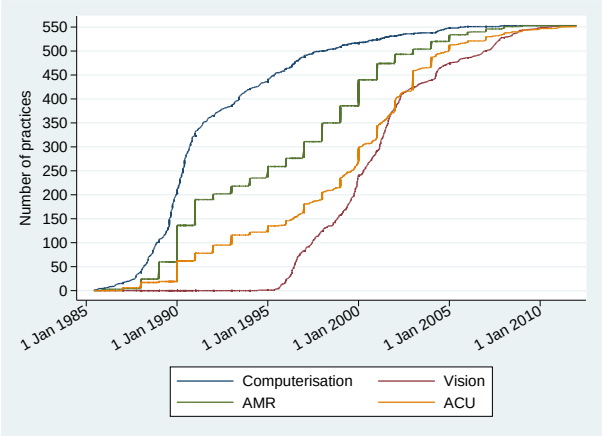

Figure 4.1: Cumulative distribution of dates when practices became computerised (computerisation), changed to Vision software, records showed acceptable mortality rate (AMR) and acceptable computer usage (ACU) of 553 practices included in THIN version 1201

Practices record postcodes in the patient records and the data provider Cegedim Strategic Data (CSD) link correctly entered postcodes to area level information, known as postcode variable indicator (PVI) data, which includes Townsend deprivation score quintile[28]. The Townsend deprivation score quintiles were derived using 2001 census data. In total, 16 practices stopped contributing data to THIN before the PVI data were generated in 2006. As PVI data were not recorded for patients in these practices, I excluded them from this study. I also excluded an additional 4% of practices because less than 80% of patients had a Townsend deprivation score recorded, which occurred if the patients postcode was not a valid postcode in 2001.

In summary, to form the cohort for analysis in this thesis, I selected data for each patient from the latest of i) AMR date, (ii) ACU date, (iii) registration date (the date the patient registered to the practice) or (iv) 1 January 1995. I followed patients up and censored them at the earliest of (i) death, (ii) the patient transfers out of the practice or (iii) the practice is no longer contributing data to THIN (i.e. they may have chosen to use a different data provider).

I included patients permanently registered to a practice (in contrast to 'temporary' registration status, the 'permanent' registration status implies the patient does not intend to leave the practice and the patients registration status is expected to continue indefinitely) any time from 1 January 1995 to 31 December 2012, with all relevant dates complete and in a logical order and not missing date of birth, sex or Townsend deprivation score quintile. I included patient information recorded when patients were aged between 18 and 100 years.

THIN is updated once a year, so I included the latest data available at the time I performed each analysis, which explains the different patient numbers and date ranges included in the different analyses.

I developed extensive Stata code to extract and manage the data before analysis, which is available on request.

4.1.1 Other data sources

For comparative purposes I used the Health Survey for England (HSE). The HSE is a series of annual surveys designed to measure health and health-related behaviours in adults and children in England and is considered nationally representative. Approximately 15,000 adults were included in each survey. I extracted data from three surveys in different calendar years, 1998[76], 2006[77] and 2008[78].

4.2 Recording of health indicators in THIN by age and sex

In this section, to understand the reasons GPs record health indicators, find differences in health indicator recording and evaluate the extent of the missing data, I describe health indicator recording by age and sex.

4.2.1 Methods

I included THIN data described in section 4.1 in this analysis from 1 January 1999 to 31st December 2011. I censored patients at 5 years after their last consultation if they did not consult for at least 5 years.

I extracted the annual recording of routine health indicators (height, weight, systolic blood pressure and total cholesterol) if patients had at least one measurement recorded during each calendar year. I calculated the annual recording per 100 person years and per 100 consultations. A consultation occurred if the patient visited the practice and consulted with a doctor or nurse. Health indicators were recorded when patients attend a consultation, so fewer health indicators recorded per 100 person years might be explained by patients attending fewer consultations, which I can determine by also reporting health indicator recording per 100 consultations. However,

many consultations do not record these health indicators.

4.2.2 Results

In total, 5,374,834 patients to 504 practices were included in the analysis. The height measurement annual recording per 100 person years peaked at age 20-25 years and again at age 75-80 years with a rapid decrease after this peak (Figure 4.2). The peak at age 20-25 years was approximately double for female patients compared to male patients, but was the same for both at age 75-80 years (Figure 4.2). The height measurement annual recording per 100 consultations decreased as age increased after the peak at age 20-25 years (Figure 4.2). The peak at age 20-25 years was apparent for height measurements per 100 person years and per 100 consultations, but the peak at age 75-80 year for height measurements per 100 person years was no longer apparent for height measurements per 100 consultations (Figure 4.2). Male patients had slightly more height measurement annual recording per 100 consultations compared to female patients in the younger ages groups, but this difference reduced as age increased (Figure 4.2).

The weight measurement annual recording per 100 person years were recorded more frequently by age and sex compared to height measurements (Figure 4.2). The weight measurement annual recording per 100 consultations also peaked for female patients at age 20-25 years, same as height measurements, but constantly increased for male patients up to age 75 years (Figure 4.2). After age 75 years, the weight measurement annual recording per 100 consultations decreased as age increased for both male and female patients (Figure 4.2).

The systolic blood pressure measurement annual recording was greater than for height and weight (Figure 4.2). The highest peak in systolic blood pressure measurement annual recording per 100 person years for female patients occurred at age 20-25 years, but no peak at this age for male patients (Figure 4.2). This peak for female patients, but not male patients, aged 20-25 years was also apparent for the systolic blood pressure measurement annual recording per 100 consultations (Figure 4.2).

The total cholesterol measurement annual recording per 100 person years was lowest for patients aged 18 years and increased to a peak for patients aged approximately 70 years (Figure 4.2). Total cholesterol measurement annual recording per 100 person years was very similar for male and female patients, but the total cholesterol measurement annual recording per 100 consultations was higher for male patients compared to female patients (Figure 4.2).

I report my interpretation of these results in the following discussion.

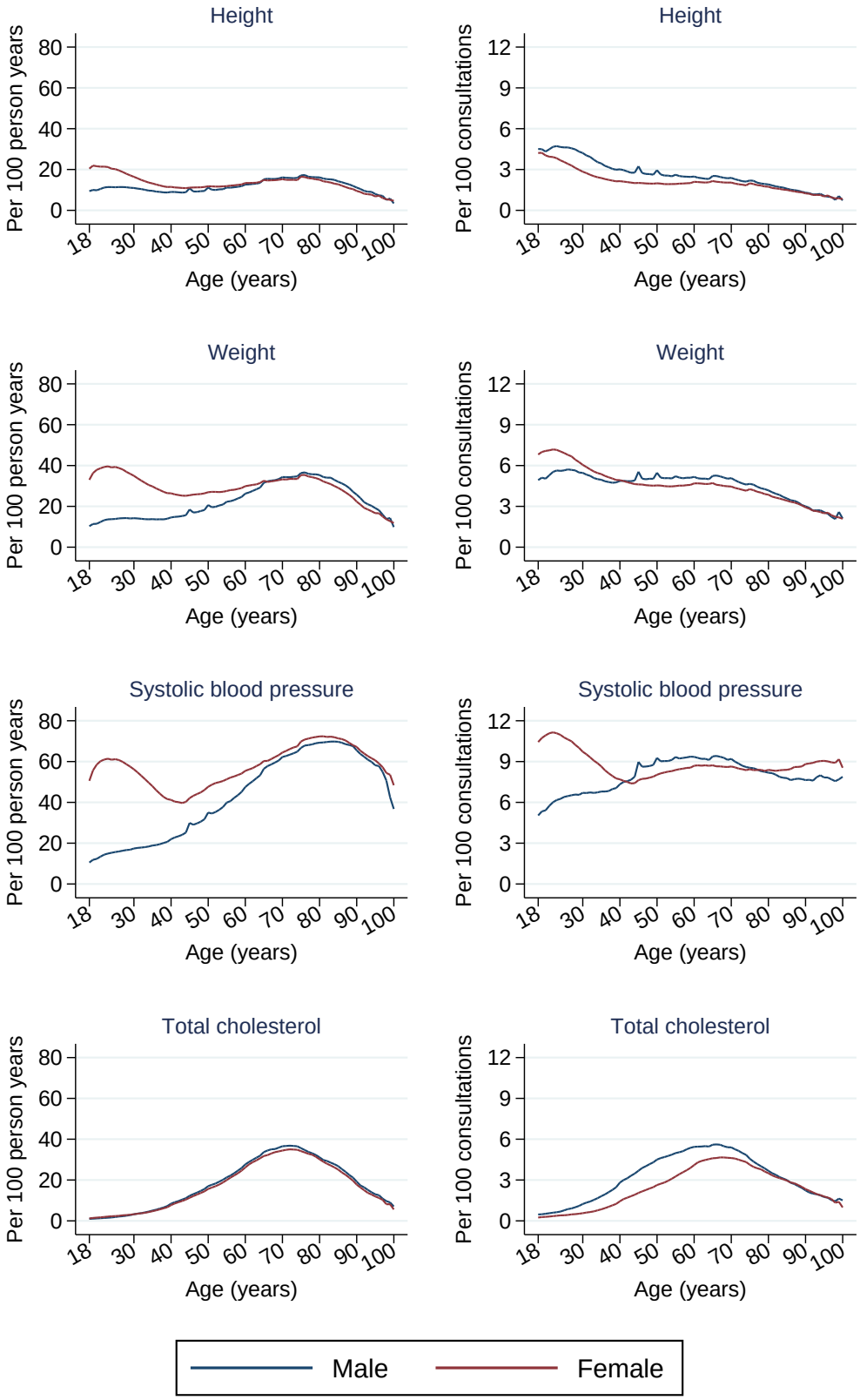

Figure 4.2: Height, weight, systolic blood pressure and total cholesterol measurement annual recording for patients in THIN from 1999 to 2011 by age and sex, per 100 person years (left) and per 100 consultations (right)

4.2.3 Discussion

In summary, the above findings showed that height, weight and systolic blood pressure measurement annual recording per 100 person years was higher for younger female patients (aged less than 50 years) compared to younger male patients. After age 75 years, annual recording per 100 person years was similar for male and female patients, and decreased after this age. The total cholesterol measurement annual recording per 100 person years was similar for male and female patients. However, male patients had a higher recording of height and total cholesterol measurements per 100 consultations compared to female patients. The annual recording of weight and systolic blood pressure measurements per 100 consultations was similar for male and female patients, except recording was higher for female patients compared to male patients less than age 40 years. Next, I consider possible reasons for these findings.

Patients may be less likely to have multiple height measurements recorded because adult patient's heights are unlikely to change over time. The decrease in height measurement recording as age increased may occur because a height measurement was recorded for most patients at younger ages and this measurement was kept in the patient record and referred to when the patient consulted with the GP at older ages. However, weight and systolic blood pressure measurements can vary over time so, to monitor patients, GPs may record them more frequently than height measurements. Bhaskaran *et al.* found a similar distribution of weight measurement recording, by age and sex, in CPRD [79], with weight measurements recorded more frequently for female patients compared to male patients and increased weight measurement recording up to age 75 years, which decreased after age 75 years.

Young female patients may of had more health indicator measurements recorded compared to young male patients because they consult more (Figure 4.3)[80]. Many female patients aged 20-25 years possibly consulted for pregnancy or contraception related reasons, when height, weight and systolic blood pressure measurements were measured.

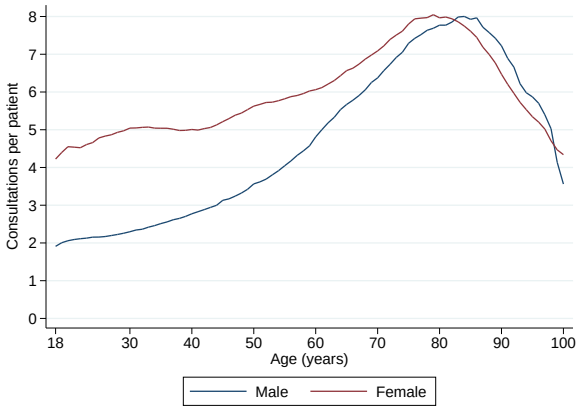

Figure 4.3: Consultations per patient by age and sex

The increased consultation rates for all patients potentially explains the peak at age 75-80 years in height and

weight measurement recording. As patients become older, disease risk increases and patients attend more consultations that require the GPs to record these measurements. Also, retired patients, particularly retired men, may be more likely to consult during retirement because they have more free time compared to before retirement. However, the height and weight measurements annual incidence per 100 consultations still decreased slightly as patients got older suggesting fewer height and weight measurements recorded as patients age increased after accounting for the increased consultations. Adjusting for consultation did not explain the high systolic blood pressure measurement recording at older ages (80+ years) because consultations per patient also decrease (Figure 4.3). Systolic blood pressure is an important health indicator for many diseases and easily measured in the GP practice, so when patients were older with an increased disease risk, blood pressure may be measured more frequently to monitor these patients.

Before age 70 years, total cholesterol measurement recording per 100 consultations was more frequent for male patients compared to female patients which may be because male patients have a higher CVD risk, which total cholesterol is an important health indicator for[10, 84]. Total cholesterol measurement recording per 100 consultations decreased steeply after the peak at age 70 years for male and female patients, which was similar to the findings for height and weight and the overall decrease in consultations per patients (Figure 4.3). This decline was possibly a survival effect. Health indicator recording initially increases because the patients were older and became sicker and required monitoring. The sickest patients that required the most monitoring died before the healthier patients, who stayed alive and required fewer health indicators recorded. Another explanation is the elderly may find it more difficult to visit the GP practice for a consultation, and some may be housebound. Alternatively, elderly patients may move to a nursing home and no longer visit their local GP or require a home visit, which would exclude them from this analysis. Some elderly patients may transfer to the practice that provides medical care for that nursing home, but that practice may not contribute data to THIN.

4.2.4 Summary

In this section, I described health indicator recording by age and sex for health indicators associated with CVD. These results suggest missing health indicator values are not MCAR because health indicator recording varies by age and sex.

In the next section, I describe health indicator recording over calender time to assess the plausibility of the MAR assumption over the years of follow-up included in the data analysis.

4.3 Longitudinal recording of health indicators in THIN

In the previous section, I described the differences between health indicator recording by age and sex for health indicators associated with CVD. In this section, I describe health indicator recording over time in THIN. As described in section 3.1.4, the Department of Health (DoH) introduced 'new patient health checks' in 1992, to measure some health indicators for all patients when they register with a practice, and the Quality Outcomes Framework (QOF) in 2004, to encourage GPs to monitor patients with QOF specified diseases by measuring health indicators. I investigate how this legislation affected health indicator recording in THIN, assess the MAR assumption plausibility and evaluate if it holds at all times of health indicator recording to justify using MI to impute missing data.

4.3.1 Methods

For the purpose of this descriptive analysis, I identified four patient cohorts who's registration year with a general practice in THIN was 1995, 2000, 2005 or 2010. These patients were included in the analysis from the year of registration until they transferred out of the practice, died or censored at 31 December 2011. For each patient, I identified the date measurements were recorded for each of the following health indicators: height; weight; smoking status (current/ex-smoker/non-smoker); systolic blood pressure; and total cholesterol. I performed the following analysis for each health indicator and patient cohort:

- The proportion of male and female patients with at least one of each health indicator measurement recorded each calendar year after registration.
- A Kaplan-Meier failure curve to investigate the time from registration to the first of each health indicator measurement for male and female patients.
- The proportion of male and female patients with at least one of each health indicator measurement recorded each calendar year after registration by disease status.

The QOF guidance outlined the specific health indicators to monitor certain diseases, outlined in the following section 4.3.2. I analysed each health indicator by the diseases the health indicators were monitored for. I used QOF specified Read codes to identify a disease diagnosis (coronary heart disease (CHD, Appendix A); stroke or transient ischaemic attack (stroke, Appendix B); diabetes (Appendix C); chronic obstruction pulmonary disease (COPD, Appendix D) in patient records.

4.3.2 QOF guidance

QOF points were awarded to general practices for evidence of ongoing management of patients with specific diseases.

First introduced in 2004, QOF (`http://www.nhsemployers.org/PayAndContracts/GeneralMedicalServicesContract/QOF/Pages/Rewarding-QOF.aspx`) awarded points for 'the percentage of patients with a diabetes diagnosis whose notes recorded body mass index (BMI) in the previous

15 months'. GPs monitored BMI for patients with a diabetes diagnosis, aiming to reduce weight in overweight patients with a diabetes diagnosis and improve glycaemic control. Both height and weight measurements were required to calculate BMI.

Also in the 2004 guidance, QOF awarded points for 'the percentage of patients with CHD, stroke or diabetes whose notes have a record of blood pressure in the previous 15 months'. GPs monitored systolic blood pressure for patients with a coronary heart disease (CHD), stroke or diabetes diagnosis as it is important for these patients to maintain a normal systolic blood pressure.

In 2006, a new QOF guidance awarded points for 'the percentage of patients with CHD, stroke or diabetes who have a record of total cholesterol in the previous 15 months'. QOF derived this guidance from evidence that lower total cholesterol reduced vascular risk for patients diagnosed with CHD, stroke or diabetes (see website).

The 2004 QOF guidance awarded points for recording smoking status every 15 months for patients diagnosed with at least one of the following diseases: CHD, stroke, diabetes or COPD. This guideline excluded patients who never smoked. In 2008 (`http://www.nhsemployers.org/PayAndContracts/GeneralMedicalServicesContract/QOF/Pages/ChangesToQOF200809.aspx`), QOF modified this guidance to include a smoking status record for all patients diagnosed with these diseases, regardless of previous smoking status.

4.3.3 Results

In total, 33,454 patients registered in 1995 and this increased to 188,847 in 2010 (Table 4.1). Patients were younger in cohorts registered earlier. More female patients than male patients registered, and this percentage was similar for each cohort (Table 4.1).

Table 4.1: Number, median and mean age at registration and sex distribution of newly registered patients in 1995, 2000, 2005 or 2010.

Cohort	Number of	Age (years)	Sex, n (%)	
	patients	median (IQR)	Male	Female
1995	33,454	31.4 (21.4, 45.9)	15,813 (47.3)	17,641 (52.7)
2000	113,047	31.5 (23.7, 44.3)	52,659 (46.6)	60,388 (53.4)
2005	193,963	31.9 (24.2, 44.1)	89,885 (46.3)	104,078 (53.7)
2010	188,847	32.3 (24.8, 45.4)	86,719 (45.9)	102,128 (54.1)

IQR: interquartile range

4.3.3.1 Percentage of patients with at least one measurement recorded

For height, weight, systolic blood pressure and smoking status, the percentage of patients in each cohort with at least one measurement recorded was highest in the registration year (Figure 4.4), from 50% to 80%. This percentage decreased to the lowest point in the years immediately after registration, and increased again to its highest point in the year 2011; 10–15% for height measurements, 30–35% for weight measurements, 50–55% for systolic blood pressure measurements and 45–55% for smoking status (Figure 4.4).

The percentage of female patients with at least one height, weight, systolic blood pressure and smoking status measurement recorded was higher than male patients (Figure 4.4).

4.3.3.2 Time from registration to first measurement

Most patients had at least one height and weight measurement recorded during follow-up; at least one height measurement recorded for 80% of male patients and 85% of female patients and at least one weight measurement recorded for 85% of male patients and 95% of female patients (Figure 4.5). During follow-up, at least one systolic blood pressure measurement was recorded for almost all female patients and approximately 90% of male patients (Figure 4.5).

For height, weight and smoking status recording, female patients in cohorts registered later had a first record slightly sooner after registration compared to female patients in cohorts registered earlier (Figure 4.5). This difference between the cohorts was greatest for smoking status recording (Figure 4.5). Approximately 90% of male patients and almost all female patients had at least one smoking status recorded during follow-up (Figure 4.5).

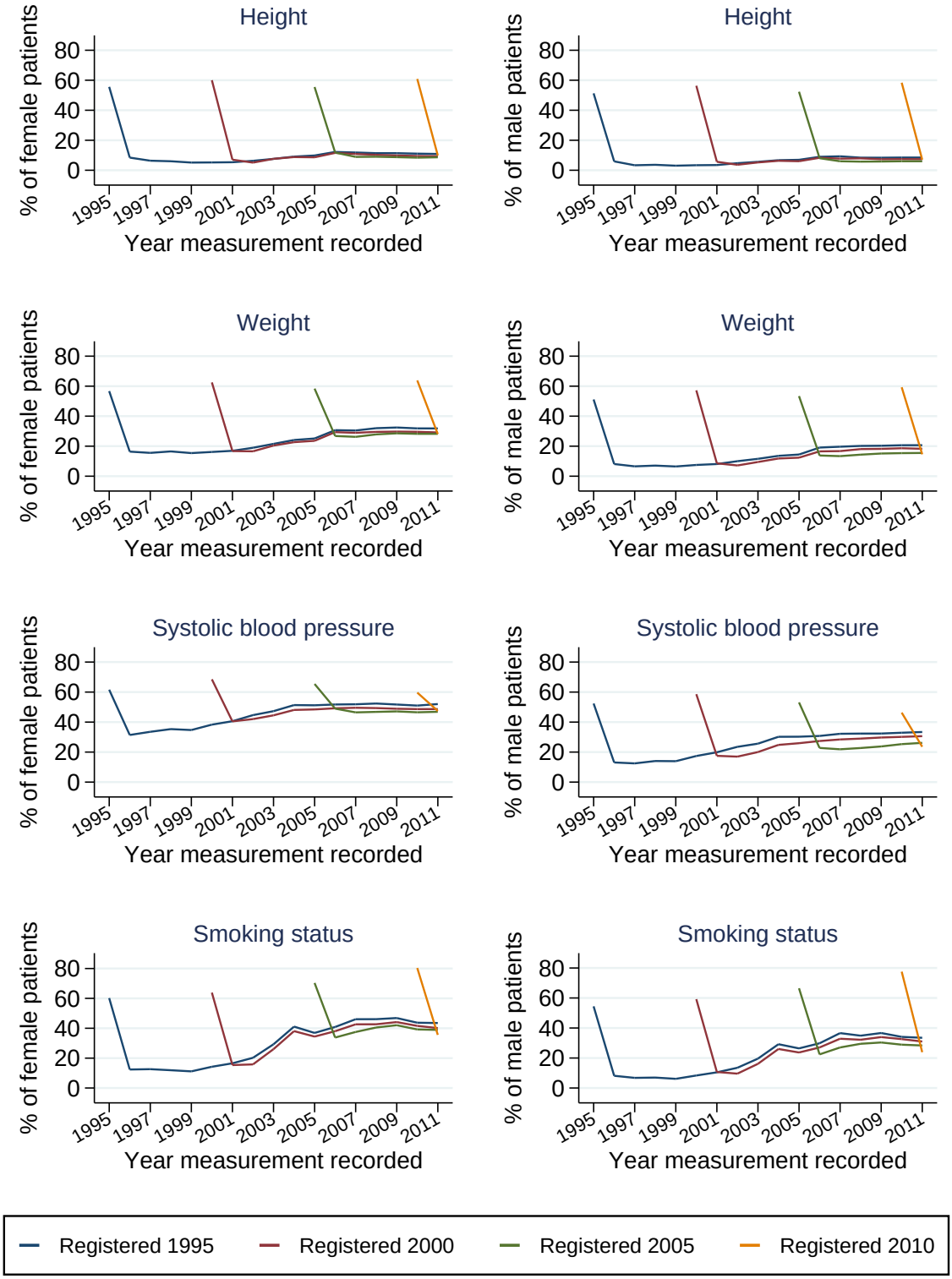

Figure 4.4: Percentage of female patients (left) and male patients (right) in each cohort with at least one height, weight, systolic blood pressure or smoking status measurement recorded each year during follow-up.

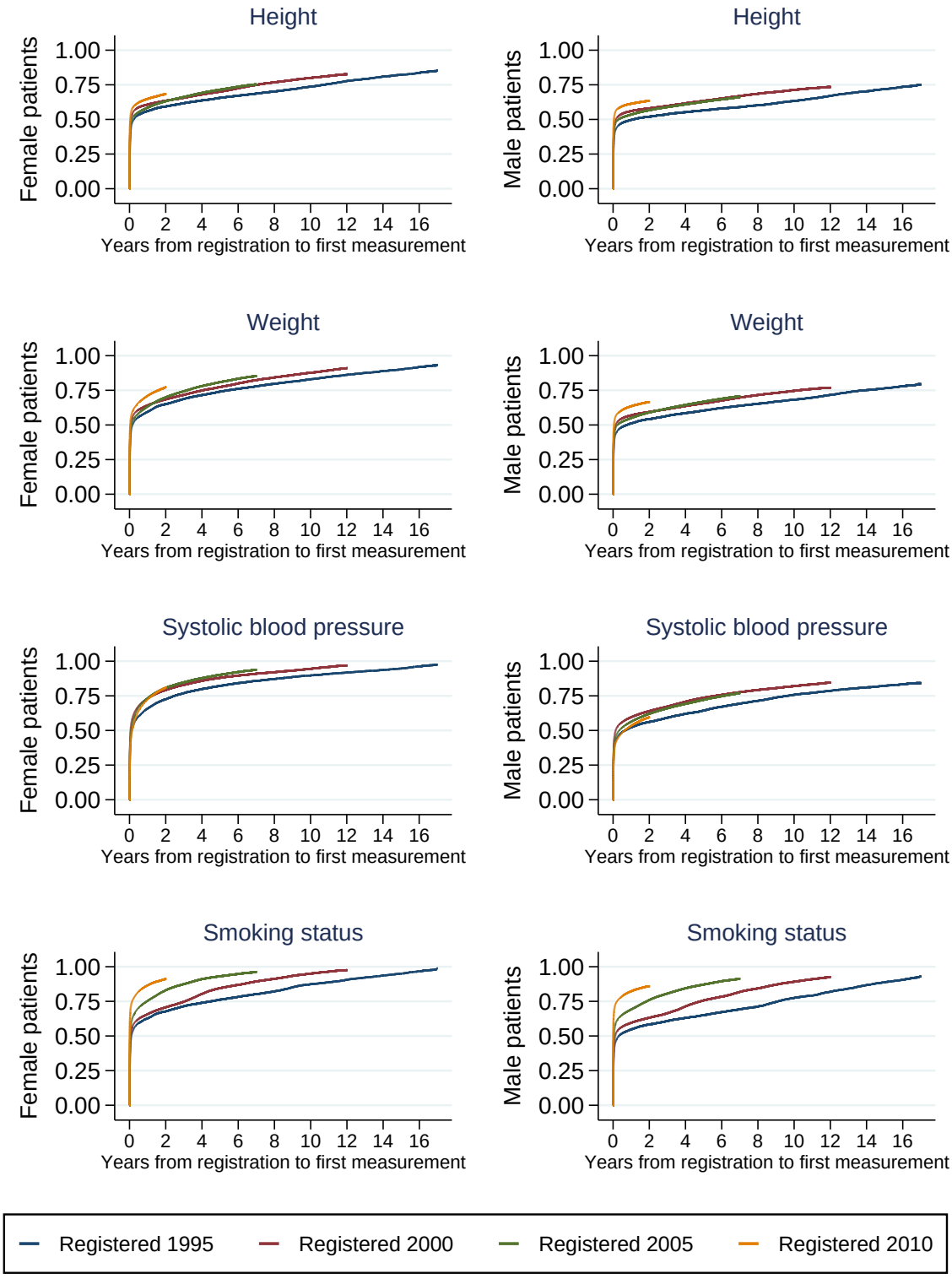

Figure 4.5: Time from registration to first height, weight, systolic blood pressure or smoking status measurement for female patients (left) and male patients (right) in each cohort.

4.3.3.3 Percentage of patients with and without diseases identified by QOF that have at least one measurement of the required health indicators

The percentage of patients with at least one height, weight, systolic blood pressure or smoking status measurement recorded was higher for patients with diabetes compared to patients without diabetes (Figure 4.6). This difference was greater in the later calendar years. I found similar results for at least one systolic blood pressure measurement recorded for patients with CHD or stroke diagnosis (Appendix E). After the QOF guidelines were introduced in 2004, approximately twice as many patients with diabetes had a smoking status recorded compared to patients without diabetes (Figure 4.6). I found similar results for patients with at least one smoking status measurement recorded and CHD, stroke and COPD diagnosis (Appendix E).

Generally, more female patients had at least one systolic blood pressure measurement or smoking status recorded compared to male patients, but recording was very similar for male and female patients diagnosed with diabetes (Figure 4.6), CHD, stroke or COPD (Appendix E).

4.3.3.4 Total cholesterol

Total cholesterol measurement recording was very different to the other health indicators. The percentage of patients with at least one total cholesterol measurements was lowest in the registration year (Figure 4.7). The percentage of patients with at least one total cholesterol measurement recorded increased during follow-up, and was greater for patients in cohorts registered earlier (Figure 4.7).

In each cohort, I observed a longer time between registration and recording the first cholesterol measurement compared to other health indicators. For example, in the first year of registration approximately 75% of patients had a first height, weight, systolic blood pressure or smoking status measurement record but approximately 15% of patients had a total cholesterol measurement recorded (Figure 4.7).

The percentage of patients with diabetes and at least one total cholesterol measurement was substantially greater than patients without diabetes (Figure 4.7). Total cholesterol recording increased steadily over time to approximately 80% of patients with diabetes in 2011 compared to approximately 20% of patients without diabetes in 2011. I found similar results for patients with other diseases (CHD and stroke) (Appendix E).

Overall, total cholesterol recording was very similar for male and female patients.

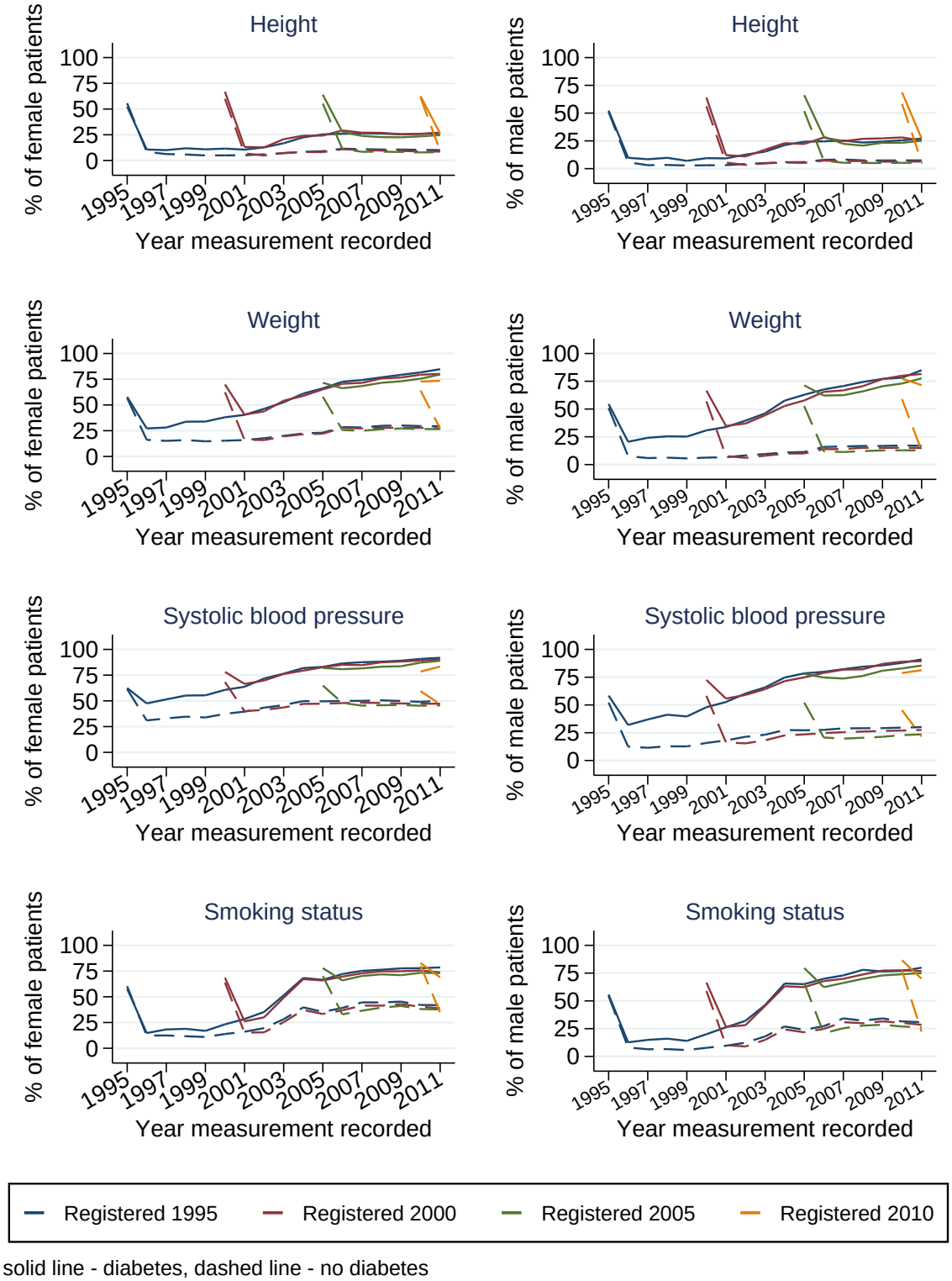

Figure 4.6: Percentage of female patients (left) and male patients (right) with and without diabetes in each cohort with at least one height, weight, systolic blood pressure or smoking status measurement recorded each year during follow-up.

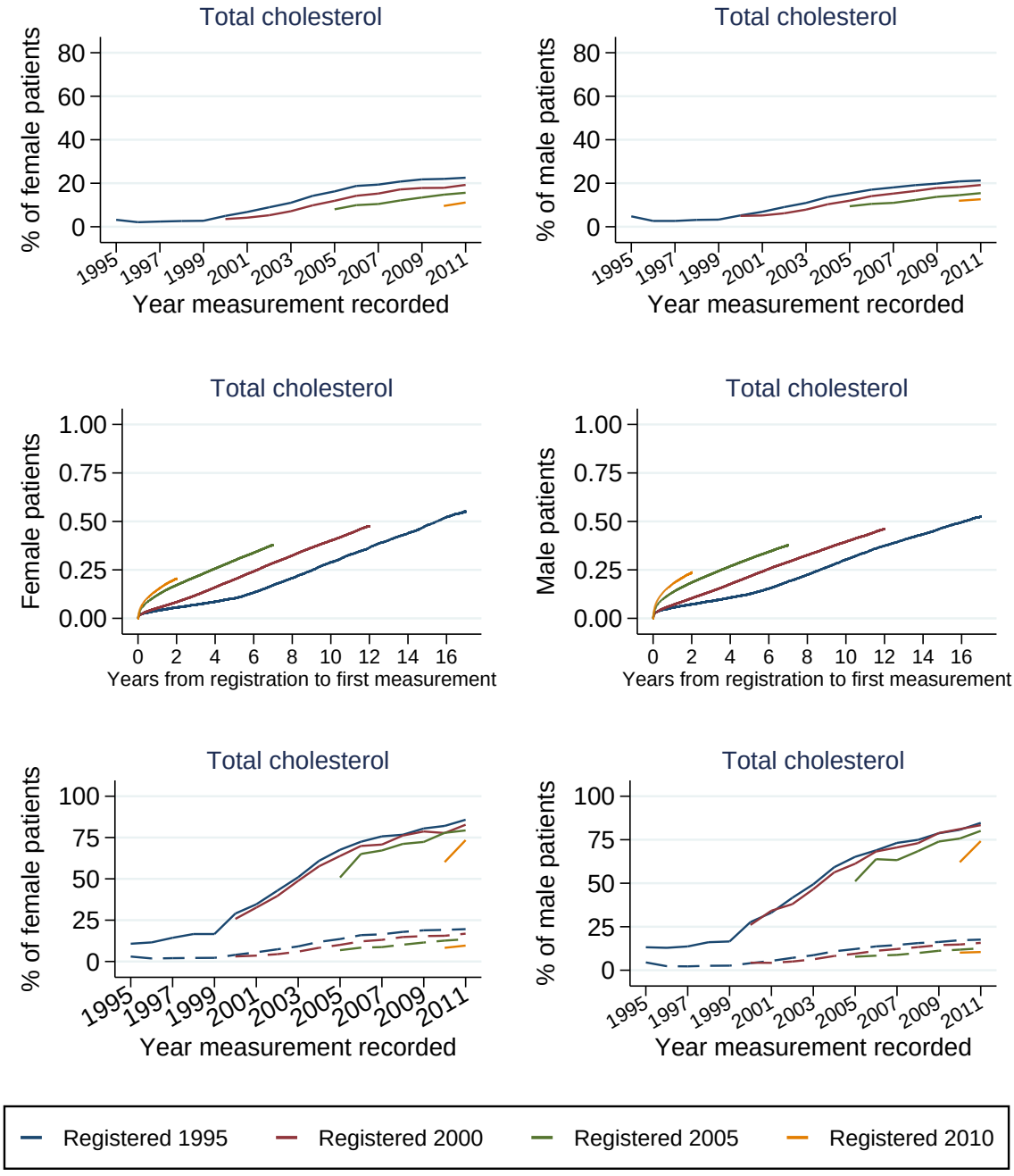

Figure 4.7: Percentage of female patients (left) and male patients (right) in each cohort with at least one total cholesterol measurement recorded each year during follow-up (top), time from registration to first total cholesterol measurement for female patients in each cohort (middle) and percentage diabetic and non-diabetic patients in each cohort with at least one total cholesterol measurement recorded each year during follow-up (bottom; solid line - diabetes, dashed line - no diabetes).

4.3.4 Discussion

In summary, the above findings showed height, weight, systolic blood pressure and smoking status recording was highest in the registration year, due to DoH introducing new patient health checks (discussed earlier in section 3.1.4). Height measurements mostly stay the same for patients after age 18 years, which explains why height was less frequently measured, but weight, systolic blood pressure and smoking status can vary over time so GPs repeatedly measured these health indicators to monitor patient health.

Overall, health indicator recording increased after DoH introduced the QOF guidance in 2004. Recording particularly increased for patients with QOF specified diseases. Therefore, many disease free patients had few health indicator measurements recorded after the registration year. Bhaskaran *et al.* also observed similar weight measurement recording in CPRD, with increased recording in more recent years and for patients with type II diabetes compared to patients without type II diabetes[79].

Female patients consulted more than male patients[81, 82]. One study that investigated reasons why men and women consult found psychiatric chronic illness and psychological distress were more strongly associated with consultation in women than in men. Current somatic symptoms and cognitive factors (illness attitudes related to illness behaviour and health anxiety) were more strongly associated with consultation in men than women, despite women reporting more physical symptoms and negative illness attitudes overall[83]. Systolic blood pressure was recorded more frequently for female patients compared to male patients, possibly because female patients visited the GP for the contraceptive pill or pregnancy related consultations. During these consultations, GPs require a blood pressure measurement.

Fewer patients had a total cholesterol measurement recorded compared to other health indicators. One explanation is because a blood sample was required to measure total cholesterol, usually only taken if the GP suspects the patient has a disease, indicated by a high BMI or high blood pressure, and requires total cholesterol measurement to confirm diagnosis. However, total cholesterol recording increased after the introduction of QOF in 2004 and also after 1999 when statins (a lipid-modifying drug) became more commonly prescribed. Lipid-modifying drugs lower total cholesterol so GPs measure and record total cholesterol for patients prescribed lipid-modifying drugs to monitor total cholesterol reduction.

I found no differences in recording between sex for patients diagnosed with any one of the QOF specified conditions discussed in this section.

From the analysis, I observed only a few patients had health indicators recorded in a given year and missing data was dependent on age, sex and disease status. Therefore, the results from a complete records analysis in a given year would probably be biased because the MCAR assumption is violated.

The MAR assumption is plausible in the registration year. Many patients had data recorded in this year, because

of new patient health checks, which depended less on unobserved data, i.e. the MNAR assumption is unlikely. However, MCAR is also unlikely because observing health indicators is conditional on registration status in each year and, as I have shown in this section, newly registered patients were more likely to have health indicators recorded. I can include registration status in the imputation model to increase the plausibility of the MAR assumption.

The MAR assumption was less plausible in the years after registration because data was mainly recorded for patients with a disease diagnosis. The MAR assumption can be made more plausible by including disease status variables in the imputation model. Diseases status is probably predictive of both missingness and the underlying missing values so, as explained in section 3.3.6, it is important to include these variables in the imputation model to achieve unbiased estimates and SEs and to increase the MAR assumption plausibility. Therefore, when I impute missing health indicator values in THIN in chapter 6, I will condition on disease status.

The MAR assumption was least plausible for total cholesterol measurements compared to other health indicators because recorded measurements were probably abnormal so the MNAR assumption was more plausible (total cholesterol recording more likely for patients with higher total cholesterol) than the MAR assumption. However, as well as conditioning on disease status in the imputation model to make the MAR assumption more plausible, I can also condition on lipid-modifying drugs which is also predictive of missing total cholesterol and the underlying total cholesterol values.

Many patients had at least one recording of the health indicators described in this section during follow-up, suggesting the best imputation approach exploits the longitudinal data, such as the two-fold FCS algorithm. These measurements could be predictive of both missingness at future times (for example, patients with high weight measurement recorded may be more likely to have future weight measurements recorded) and the underlying missing values themselves (patients with a high weight measurement recorded may be more likely to have a high weight measurement at other times) and, therefore, important to include these variables in the imputation model to both improve the plausibility of the MAR assumption and increase the accuracy of the imputed values.

4.3.5 Summary

From the analysis in this section, I found health indicator recording changed over time and by disease status. Therefore, I can increase the MAR plausibility of missing health indicators over time by including disease status and lipid-modifying drugs for total cholesterol, as well as age, gender, in the imputation model. Also, the two-fold FCS algorithm is potentially a suitable method because it can incorporate measurements at other times and many patients had at least one health indicator measurement, except total cholesterol, recorded during follow-up. Before I apply the two-fold FCS algorithm, I first consider identifying and excluding outliers in THIN. The next section described a new approach to identify outliers in data with repeated measurements, like THIN.

4.4 Two stage method to remove outliers from longitudinal data

Any incorrectly recorded values may influence the imputation of the missing values and potentially bias the associations between the variables and the imputed values drawn from the conditional distribution based on the data. Therefore, I investigate removing incorrectly recorded values before imputation and retain extreme but genuine values in the dataset to represent true variation in the population.

In this section, I compare a new method to identify outliers for health indicators repeatedly recorded in THIN to existing methods. I illustrate the methods using height measurements. The results for other health indicators (weight, total cholesterol, high-density lipoprotein cholesterol, systolic blood pressure and diastolic blood pressure) are reported in Appendix F. I submitted these findings for publication and Pharmacoepidemiology and drug safety published the manuscript describing this method in 2011[4] (Appendix G).

4.4.1 Introduction

THIN consist of repeated measurements of health indicators such as height, weight, blood pressure and cholesterol for several million patients. Inevitably, some values for these health indicators were recorded incorrectly. However, some extreme values were genuine because the database covers a large, diverse population. The challenge for researchers is to distinguish between incorrect records and those which represent true, but extreme values.

Existing methods to identify outliers include removing values more extreme than the mean $\pm$ a multiple of the standard deviation[85]. Also, the Dixon Test[86] compares the most extreme value with the second most extreme to identify outliers and is the recommended method to exclude outliers before using the data to define a Reference Interval (RI) for laboratory tests using a sample of healthy individuals to calculate a suitable range. Outliers potentially affect the calculation of this range. The Dixon Test was previously validated for use with cross-sectional data, but not longitudinal data.

In longitudinal records, repeated measurements were recorded for individual patients so outliers occurred at two levels: population level and individual level. Population level outliers were values substantially outside the range of the distribution of measurements from the general population e.g. height measurements above 2.4 metres. Individual level outliers were highly implausible values in the context of measurements collected over time for each patient. E.g. an patient with height repeatedly measured as 1.8 metres was unlikely to suddenly shrink to 1.5 metres. However, 1.5 metres was a plausible measure for another patient. Therefore, existing methods did not capture this incorrectly recorded value because they only detected population level outliers. Ignoring individual level outliers potentially biases further analyses, such as calculation of an patient's body mass index (BMI). For example, if a patient weighed 60 kg and had a height of 1.8 metres, this patient's body mass index (BMI) is 18.5 kgm^{-2}; within the normal range. However, if height was incorrectly recorded as 1.5m, the BMI was 26.7 kgm^{-2}; the patient is overweight.

4.4.2 Methods

For comparative purposes, I chose to select a 10% random sample of practices from THIN. As THIN consisted of many patients, the analysis of a subset of patients was more manageable. I included patients in the analysis if they met the inclusion/exclusion criteria described in section 4.1. Patients aged 16 years and over were included and data from the time period 1995 to 2009.

4.4.2.1 Population level outliers defined using Health Survey for England

To identify potential population level outliers, I interrogated the Health Survey for England (HSE). I extracted data from two surveys in different calendar years, 1998[76] and 2008[78], to capture changes in the distribution of the health indicators in the general population over the study period.

I identified and removed potential population level outliers using three different approaches.

1. Application of acceptable ranges from external sources

   I identified age and sex specific ranges from the HSE data. For each health indicator, I added/subtracted 10% to the most extreme values in HSE. Data in THIN were identified as outliers and changed to missing if they were outside these age and sex specific ranges. For the purposes of this study, I assumed the distribution of heights within sex and age range is approximately Gaussian. The 10% range was considered conservative, to ensure I retained extreme but true measurements.

2. Standard deviation from mean

   For each practice, I found the height measurements distribution and standard deviation. Any measurements more extreme than the mean $\pm$ 3 times the standard deviation were identified as outliers and changed to missing.

3. The Dixon Test

   I applied The Dixon Test[86] to each practice included in the study. If the difference between the most extreme (large or small) value and the next most extreme value divided by the range of values exceeds $\frac{1}{3}$, the extreme value in question was identified as an outlier[86].

4.4.2.2 Individual level outliers

After excluding population level outliers identified using HSE data, a series of multilevel linear regression models with a patient specific random intercept and slope were fitted to the remaining data to identify individual level outliers for patients with at least two measurements recorded on different dates. I assumed patients with single measurements were correct unless it was outside the ranges identified at population level. The following model was fitted to the height measurements adjusted for age and sex.

i = patient

j = date measurement was recorded

y_{ij} = height measurement recorded on j for patient i

x_{ij} = age of patient i when height measurement was recorded on j

z_i = sex of patient i

$y_{ij} = \alpha + \beta_x x_{ij} + \beta_z z_i + \mu_i + \upsilon_{ij} + \epsilon_{ij}$

where $\mu_i \sim N(0, \sigma_\mu^2)$ was the random intercept for patient i

and $\upsilon_{ij} \sim N(0, \sigma_\upsilon^2)$ was the random slope for patient i and height measurement recorded on j

and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ was the random error for patient i and height measurement recorded on j.

This model allowed height to vary linearly with age, consistent with research identifying a small decrease in height with age over the typical period of observation[87]. Allowing for this potential linear trend had little computational or inferential cost. This model was desirable to identify outliers for measurements more likely to vary over time, such as weight.

Random effects models take account of the variability within patients and the standardised residuals of the random effects (observed value minus predicted value divided by the square root of the residual mean square) gauge the distance the measurement was from the fitted value predicted by the model. If the measurement was very different from the mean of all the observations for that patient, they had a larger positive/negative standardised level 1 residual. Measurements were identified as outliers and excluded if the standardised residuals were more extreme than ± 10. The model was refitted to the remaining data and outliers were identified again using the same 'cut-off'. This process was repeated until no more outliers were identified.

Adjusting for age and sex took account of the different mean and variability between men and women. The purpose of this approach was to identify and exclude measurements that fell outside the main bulk of the sex and age specific distribution. To do this I identified outliers on the basis of the standardised level 1 residual, not the level 2 (within patient) residual. I also investigated alternate standardised residuals ± 5 and ± 15 because the ± 10 'cut-off' was selected arbitrarily.

Due to the size of THIN, I analysed data from each practice separately and the results were combined. On average, there were approximately 9,000 patients registered to each practice at any one time.

If a patient had two very different measurements, it was more difficult to ascertain which point was erroneous. The method described was unlikely to identify either measurement as an outlier because the line of best fit was in between these two points. Since I fitted a multilevel model, the line of best fit was not exactly in the middle of both points, due to shrinkage, and the two points had different size standardised residuals. Therefore, I proceeded as follows: if patients had two measurements and at least one standardised residual of these measurements was more extreme than 8, the measurement with the most extreme standardised residual was identified as an outlier.

4.4.2.3 Validation of standardised residual 'cut-off' points

To validate the 'cut-off of ± 10, two volunteers (a general practitioner and a statistician) were raters, blinded to potential outlier status and asked independently to use professional judgement to determine if height measurements plotted against age for 100 randomly selected patients were an extreme outlier, a possible outlier, a low possibility outlier or not an outlier. I compared the raters' judgement to the calculated standardised residuals, categoriesed as follows prior to validation:

- No outlier: $|$ standardised residual $| < 2$;
- Low possibility outlier: $2 \leq |$ standardised residual $| < 5$;
- Possible outlier: $5 \leq |$ standardised residual $| < 10$; and
- Extreme outlier: $10 \leq |$ standardised residual $|$.

4.4.3 Results

In total, 43 practices were included in the analysis and 548,450 patients contributed data after they were 16 years old. Of these patients, 315,944 had one or more height measurement available and 513,367 height measurements.

4.4.3.1 Population level outliers

- Application of acceptable ranges from external sources

  Altogether, 1,550/513,367 (0.3%) measurements were outside the population level age and sex specific ranges derived from HSE (Table 4.2). This included a number of data entry errors e.g. the record 176 meters was probably incorrectly entered in centimetres. After excluding population level outliers, the age and sex specific minimum, maximum and mean measurements were similar to the HSE data.

- Standard deviation from mean

  Excluding height measurements more extreme than 3 standard deviations from the mean identified 751/513,367 (0.1%) outliers. The range of retained measurements still included implausible values (0.01m to 10.02m) after applying this method.

- The Dixon Test

  The Dixon Test did not identify any outliers. In this scenario, The Dixon Test was insensitive.

4.4.3.2 Individual level outliers

After excluding population level outliers identified using boundaries derived from HSE data, I fitted the random effect model once and 75 measurements had corresponding standardised residuals more extreme than $\pm$ 10. Thirteen measurements were the most extreme of two measurements when at least one measurement was more extreme than $\pm$ 8. After fitting the model twice more, I identified 93 measurements as individual level outliers, 17 were the most extreme of two measurements.

Table 4.2: Maximum + 10% and minimum -10% of height measurements found in Health Survey for England (HSE) data from 1998 and 2008 by age and sex.

Age range	Male		Female	
	Min-10%	Max+10%	Min-10%	Max+10%
$16 \leq$ and < 25	1.26	2.22	1.26	2.03
$25 \leq$ and < 35	1.37	2.23	1.27	2.06
$35 \leq$ and < 45	1.36	2.21	1.24	2.06
$45 \leq$ and < 55	1.35	2.18	1.22	2.02
$55 \leq$ and < 65	1.27	2.16	1.25	1.99
$65 \leq$ and < 75	1.33	2.10	1.26	1.96
75 +	1.32	2.08	1.11	1.97

Using a $\pm$ 5 'cut-off', I identified 825 individual level outliers when I fitted the random effects model once to the data. However, this did not effect the range of values retained. A $\pm$ 15 'cut-off' was less sensitive and I only identified 9 individual level outliers after fitting the random effect models 3 times.

In total, I identified 1,643 (0.3%) height measurements as outliers after application of both methods to identify population level outliers and individual level outliers using a 'cut off' of $\pm$ 10 standardised residuals.

4.4.3.3 Validation

Using a sample of 100 patients with 40 extreme individual level outliers, one rater identified 30 (75.0%) of these as extreme outliers and the other identified 32 (80.0%). The raters and the random effects model were consistent (Table 4.3). Out of 310 measurements which were not extreme outliers, rater 1 identified 18 (5.8%) as extreme outliers and rater 2 identified 27 (8.7%) as extreme outliers.

The measurements the raters classified as no outlier, low possibility outlier, possible outlier and extreme outlier were in concordance with the standardised residuals calculated by the random effects model (Figure 4.8). However, they did not identify a number of extreme outliers.

4.4.4 Discussion

In this section, I examined methods to remove outliers from electronic health care records. Removing population outliers based on sensible boundaries defined by representative survey data was more efficient compared to other methods such as the Dixon Test and three times the standard deviation.

Historically, it was common practice to select an acceptable range of values based on a multiple of the standard deviation. The earliest formulation of an outlier test criterion a rule of thumb, and involved 'reject(ing) any observation whose residual exceeds in magnitude five times the probable error' (i.e. 3.37 times the standard

Table 4.3: Height measurements identified as outliers by the random effects model adjusted for age and sex and the two raters.

Outlier		Random effects model			
		No outlier	Low possibility	Possible	Extreme
Total		150	95	65	40
Rater 1,	No outlier	150 (100.0)	94 (99.0)	30 (48.4)	4 (10.0)
n (%)	Low possibility	0 (0.0)	1 (1.0)	7 (11.3)	0 (0.0)
	Possible	0 (0.0)	0 (0.0)	7 (11.3)	6 (15.0)
	Extreme	0 (0.0)	0 (0.0)	18 (29.0)	30 (75.0)
Rater 2,	No outlier	126 (84.0)	74 (77.9)	26 (41.9)	4 (10.0)
n (%)	Low possibility	11 (7.3)	16 (16.8)	3 (4.8)	0 (0.0)
	Possible	13 (8.7)	5 (5.3)	6 (9.7)	4 (10.0)
	Extreme	0 (0.0)	0 (0.0)	27 (43.6)	32 (80.0)

deviation) because if the Gaussian law of error was truly satisfied, only about one observation in a thousand was rejected, 'and therefore little damage will be done in any case'[85]. For cross-sectional data, using fixed thresholds calculated as a multiple of the sample standard deviation was inadequate for identifying outliers. They can inflate the estimated standard deviation so their presence was not detected[88]. Recent work has also showed that outlier detection using fixed multiples was inefficient[89]. In addition, the proportion of measurements identified as outliers was influenced by the size of the sample. This method had shortcomings with longitudinal data because it did not identify implausible height measurements as outliers. A possible explanation for this because THIN is clinical data, rather than research data, with more data entry errors.

Another method investigated was The Dixon Test. This method did not identify any outliers in THIN, maybe because masking affected The Dixon Test, which occurs when the less extreme ones mask the aberrance of the most extreme.

In addition to removing population outliers, I repeatedly fitted a random effects model to longitudinal clinical data to further identify a small number of individual level outliers. This two stage method may reassure longitudinal data users that records were consistent within patients by identifying implausible changes between successive observations.

Initially, the random effect model was fitted to the data without removing population level outliers, but the method had difficulties converging in some situations. However, after removal of population level outliers, the method converged successfully as the most extreme values were no longer present to influence the random effects models.

If a patient had an extreme measurement recorded, in most cases this resulted in the GP taking further measure-

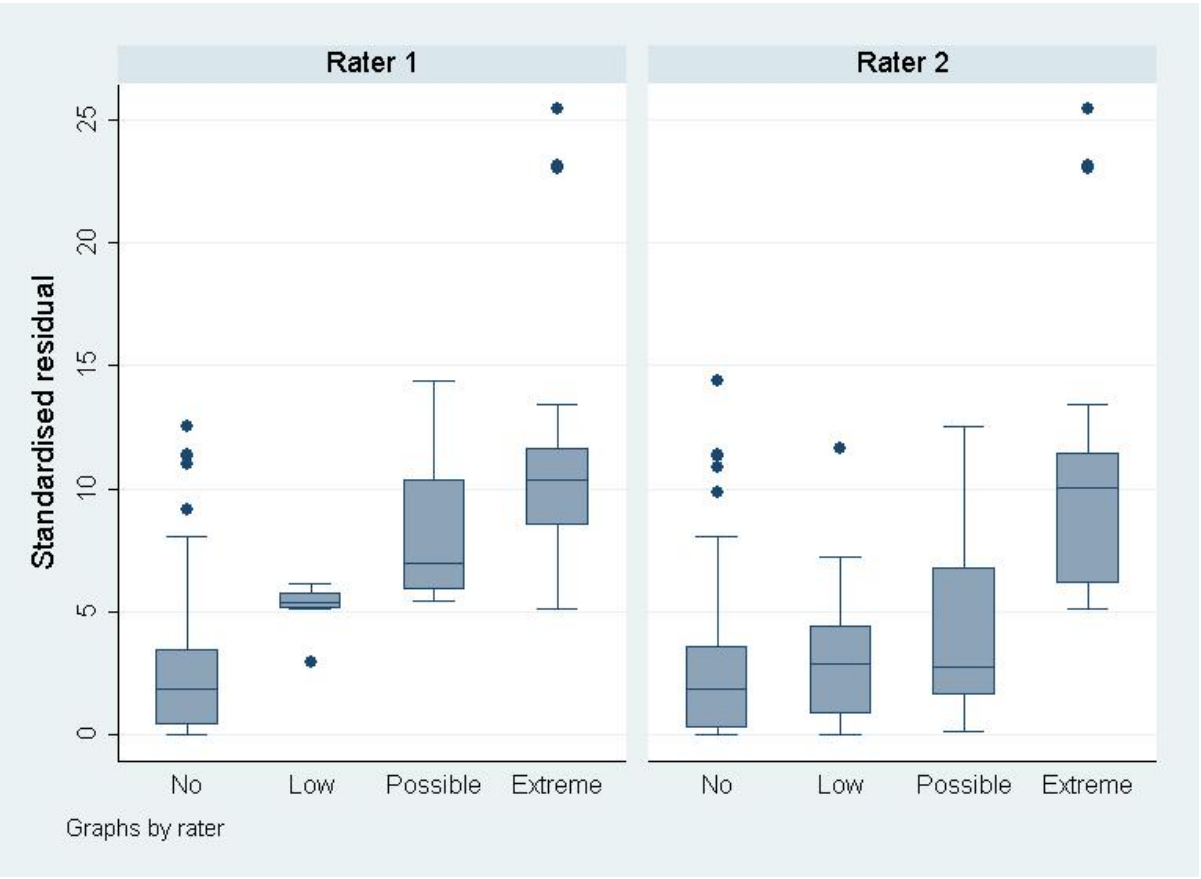

Figure 4.8: Distribution of standardised residuals for height measurements classified as no outlier, low possibility outlier, possible outlier or extreme outlier by the two raters (negative standardised residuals were recoded to positive)

ments, either to ascertain it was a true measurement or to monitor the patient. Extreme measurements were less likely to be identified as outliers when repeatedly recorded. Therefore, this method accurately identifies an outlier in the situation where the measurement was very different to all the other measurements recorded for a patient. With three or more measurements, assuming the majority were reasonably accurate, individual level outliers were identified with minimum reference to the population distribution. I used population level information for only two measurements to identify the most likely outlier.

As with all these methods to identify outliers, the 'cut-off' was an arbitrary choice and sometimes difficult to justify. The aim of this study was to attempt to identify the most likely outliers and retain the true but extreme values. Therefore, even though the boundaries described in Table 4.2 were wide, the random effects models were a 'double check' for the remaining values to identify individual level outliers.

A 'cut-off' point of standardised residuals more extreme than $\pm$ 5 identified more outliers each time the random effects model was re-fitted to the data compared to using $\pm$ 10 as a 'cut-off'. However, using a 'cut-off' of $\pm$ 5 could artificially reduce the variation too much. A 'cut-off' of $\pm$ 15 was too large as it did not identify some

impossible values as outliers. A small validation study was conducted to justify using $\pm$ 10 standardised residuals as a suitable 'cut-off' point for height. Extreme outliers from the random effects model did not always agree with the points selected by the raters. This result was not surprising as the rater's judgement was compared against a defined set of rules. However, there was a good agreement shown as 75% and 80% of extreme individual level outliers were identified by the raters.

The appropriate 'cut-off' choice might vary for other variables. A limitation of using height as an example was limited variation within individual measurements. However, I used the two-step method to identify outliers for other health indicators in THIN and I found excluding weight measurements with standardised residuals more extreme than $\pm$ 10 excluded approximately the same number of outliers as for height.

In a similar process to this study, other studies used a combination of a defined range ('cut-off' points) and identify any remaining outliers using an iterative re-weighting process as a robust estimator[90, 91]. These studies didn't use repeated measurement data.

The two stage outlier removal method was efficient, low cost and computationally relatively easy to apply to longitudinal data. Removing outliers increases users' confidence in the data. If very large samples of data were used, a few outliers may have little impact on the analysis. However, many studies used a subset of primary care databases and other electronic health records, e.g. comparing a group of patients with disease X against a group of patients with disease Y. In these samples, outliers due to measurement error may substantially bias the results. Regression dilution will occur if outliers were present in the data as they increase the variance of the covariate[92].

If no external source of data exists to compare against, use common sense as an alternative way to exclude the most extreme, impossible population level measurements, e.g. exclude height measurements greater than 3m or less than 1m. I applied this method to identify outliers in longitudinal data from THIN primary care database, but it is easily adapted to identify outliers in other data sources with repeated measurements on individuals.

4.4.5 Summary

In this section, I described a method to identify outliers in data with repeated measurements. I identified and removed these outliers before using multiple imputation so they do not incorrectly influence the fit of the imputation model to the data.

In the following chapter, I will evaluate how well the two-fold FCS algorithm performs using a simulation study to impute missing weight, systolic blood pressure, smoking status and total cholesterol measurements, using methods informed by the findings in this chapter.

Chapter 5

Developing and evaluating the two-fold fully conditional specification algorithm

In the previous chapter, I described health indicator recording in The Health Improvement Network (THIN) by age, sex and over time and explained how the missing at random (MAR) assumption was plausible if I condition on certain variables when imputing, which justifies using multiple imputation (MI). As explained in section 3.3, the two-fold fully conditional specification (FCS) algorithm may be a suitable method to impute missing data in longitudinal, clinical databases. However, before applying the two-fold FCS algorithm in THIN, in this chapter I assess bias and precision using simulated data.

In the next section, I simulate health indicator variables to mimic the essential features of their distributions in THIN. Then I change weight, systolic blood pressure and smoking status values to missing using a MCAR missingness mechanism and compare the bias and precision of estimates from fitting the model of interest to the original simulated data, complete records analysis and data imputed using a 'baseline MI' (i.e. impute missing values at one time point) and using the two-fold FCS algorithm. Simulation studies are advantageous because they allow me to compare the imputed data to the original data to evaluate the two-fold FCS algorithm because the original coefficients are known from the data generation mechanism. Also, I can re-run the analysis with different among-time iterations and window width to investigate how varying these parameters affects convergence.

In the last section of this chapter, I repeated the simulation process but include total cholesterol and lipid-modifying drugs variables in the simulated data. As described in chapter 4, total cholesterol is recorded less frequently compared to weight, systolic blood pressure and smoking status so I change more total cholesterol values to missing compared to other health indicators. I find the bias and precision of total cholesterol estimates from fitting the model of interest to the original simulated data and compare to different approaches to impute missing total cholesterol values using the two-fold FCS algorithm.

5.1 Simulation study to evaluate the two-fold fully conditional specification algorithm

In this section, I evaluate the two-fold FCS algorithm, by comparing bias and precision, using a simulation study to assess its potential for routine application to large primary care databases. I explore how biased and precision varies with two key parameters: the among-time iterations and the time window width. I submitted part of this work for publication as part of the project funded by the MRC and Statistics in Medicine published this article[3] (Appendix H).

5.1.1 Methods

Later in this thesis, to impute THIN data, I stratify patients by practice and sex to allow relationships among health indicators to vary by practice and sex and impute each strata separately to reduce the data included in each imputation. But, before imputing THIN data, I evaluate the two-fold FCS algorithm using simulated data which accurately reflected the planned THIN imputation in section 6.1. Specifically, I simulated data to mimic an extracted THIN cohort to investigate the association between health indicators recorded at a baseline time block and a future coronary heart disease (CHD) event. I generated samples of $N = 5,000$ male patients in each simulated dataset. This sample size should have enough power to obtain meaningful results. Also, I chose to impute male patients only because, in chapter 4, I showed the risk of CHD and health indicator recording varied by sex. Also, as the statistical properties of the method should be similar across different linear relationships between the variables, the results found from a simulation study based on males should apply to females. I used the following steps to perform the simulation study:

1. simulated full data including complete observations for patients based on a cohort of patients extracted from THIN;
2. fitted the model of interest to the full data, recording parameter estimates and standard errors;
3. made observations missing completely at random (MCAR);
4. imputed missing data to create multiple imputed datasets, and
5. fitted the model of interest to each imputed dataset, and combined the results using Rubin's rules[18].
6. Recorded the imputation-based parameter estimates and standard errors.

I chose to use a MCAR mechanism instead of MAR mechanism, even though earlier in this thesis I showed the MCAR assumption was not plausible, because the statistical properties (estimates and SE) will be similar. Also, a MCAR mechanism allowed me to detect bias introduced by the imputation method and (if there is no bias) use comparisons of SEs from full data, complete record and MI analysis to directly compare efficiency.

I repeated the process 1,000 times so I could estimate bias in the parameter estimates, compare the empirical standard errors of the parameter estimates with the estimated standard errors and estimate the coverage of nominal

95% confidence intervals.

5.1.1.1 Data generation mechanisms

I built up simulated data for patients using a sequence of conditional steps, starting with the simulation of the patients and their registration times. The probabilities for the algorithm were derived from a cohort of male patients extracted from THIN, as described in section 5.1.1.3. For the simulation study, I created 1,000 simulated datasets, each consisting of $N = 5,000$ male patients and 10 years of follow-up time from 2000–2009. For each patient i, where $i = 1, ..., N$, and for each year (or time block) t from 2000 to 2009, indicated by $t = 0, ..., 9$, I simulated the following categorical and continuous health indicator and patient characteristic variables associated with CHD risk: time of registration (1999 or before); smoking status; age at baseline; Townsend deprivation score quintile (measurement of social deprivation); systolic blood pressure; weight and anti-hypertensive drug treatment. I investigated two different data generation mechanisms:

Data generation mechanism I
For each patient $i = 1, \ldots, N$:

1. I marked patient i as registered with the general practice in 1999 with probability p_1, and created binary variable where $reg_i = 1$ denoted patients registered in 1999 and $reg_i = 0$ for patients registered before 1999;

2. I generated baseline age (i.e. age in 2000), denoted age_i with values $a = 1, \ldots, 10$, corresponding to 5-year age categories from 40 to 89, using probabilities $q_{0,a}$ for patients registered before 1999 ($reg_i = 0$) and probabilities $q_{1,a}$ otherwise;

3. I generated smoking status (time-independent), denoted $smoke_i$, from a multinomial logistic model conditional on age category in 2000:

$$\text{logit}\{\Pr(smoke_i = b)\} \quad = \quad \beta_{0,b}^{smoke} + \sum_{a=2}^{10} \beta_{a,b}^{smoke} \, [age_i = a]$$

   where $b = 1$ (non-smoker - reference category), 2 (ex-smoker), 3 (current smoker);

4. I generated social deprivation (Townsend deprivation score quintile, from 1 (least deprived) to 5 (most deprived), denoted $townsend_i$, from an ordinal logistic regression model conditional on registration (1999 or other), age category in 2000 and smoking status:

$$\begin{aligned}\text{logit}\{\Pr(townsend_i = c)\} &= \beta_{0,c}^{town} + \beta_1^{town}\,[reg_i = 1] + \sum_{a=2}^{10} \beta_{2,a}^{town}\,[age_i = a] \\ &\quad + \sum_{b=2}^{3} \beta_{3,b}^{town}\,[smoke_i = b]\end{aligned}$$

where $c = 1, ..., 5$, with reference category $c = 1$;

5. I generated continuous time-dependent systolic blood pressure and weight measurements, denoted $systolic_i$ and $weight_i$ respectively, for calendar years 2000 to 2009 (denoted $t = 1, \ldots 10$).

   For time block t:

$$\begin{aligned}systolic_{i,t} &= \beta_{0,t}^{sys} + \beta_{1,t}^{sys}[reg_i = 1] + \sum_{a=2}^{10} \beta_{2,t,a}^{sys}[age_i = a] + \sum_{b=2}^{3} \beta_{3,t,b}^{sys}[smoke_i = b] \\ &\quad + \sum_{c=2}^{5} \beta_{4,t,c}^{sys}[townsend_i = c] + \sum_{l=0}^{t-1} \beta_{5,t,l}^{sys} systolic_{i,l} + \sum_{l=0}^{t-1} \beta_{6,t,l}^{sys} weight_{i,l} + \epsilon_{1,i},\end{aligned}$$

   and

$$\begin{aligned}weight_{i,t} &= \beta_{0,t}^{weight} + \beta_{1,t}^{weight}[reg_i = 1] + \sum_{a=1}^{10} \beta_{2,t,a}^{weight}[age_i = a] \\ &\quad + \sum_{b=1}^{3} \beta_{3,t,b}^{weight}[smoke_i = b] + \sum_{c=1}^{5} \beta_{4,t,c}^{weight}[townsend_i = c] \\ &\quad + \sum_{l=0}^{t} \beta_{5,t,l}^{weight} systolic_{i,l} + \sum_{l=0}^{t-1} \beta_{6,t,l}^{weight} weight_{i,l} + \epsilon_{2,i},\end{aligned}$$

   where $\epsilon_{1,i} \overset{iid}{\sim} N(0, \sigma^2_{\epsilon_1})$ and $\epsilon_{2,i} \overset{iid}{\sim} N(0, \sigma^2_{\epsilon_2})$.

6. I generated binary time-dependent anti-hypertensive drug treatment variables, denoted $drug_{i,t}$ for time block t, from logistic regression models:

$$\begin{aligned}\text{logit}\{\Pr(drug_{i,t} = 1)\} &= \beta_{0,t}^{drug} + \beta_{1,t}^{drug}[reg_i = 1] + \sum_{a=2}^{10} \beta_{2,t,a}^{drug}[age_i = a] \\ &\quad + \sum_{b=2}^{3} \beta_{3,t,b}^{drug}[smoke_i = b] \\ &\quad + \sum_{c=1}^{5} \beta_{4,t,c}^{drug}[townsend_i = c] \\ &\quad + \beta_{5,t}^{drug} systolic_{i,t} + \beta_{6,t}^{drug} weight_{i,t}.\end{aligned}$$

7. I generated a time to CHD event outcome T_i from an exponential distribution (with constant proportional hazard) with log hazard equal to:

$$\begin{aligned}&\theta_0 + \sum_{a=2}^{10} \theta_{1,a}[age_i = a] + \sum_{b=2}^{3} \theta_{2,b}[smoke_i = b] \sum_{c=1}^{5} \theta_{3,c}[townsend_i = c] \\ &\qquad + \theta_{5,1} systolic_{i,1} + \theta_{6,1} weight_{i,1} + \theta_{7,1}[drug_{i,1} = 1]\end{aligned}$$

5.1.1.2 Imputing smoking status

Apart from smoking status, all the health indicators investigated in chapter 4 were continuous variables, assumed approximately normally distributed, even though many measurements are rounded to the nearest integer (Figures 5.1 and 5.2). However, smoking status is a categorical variable (with 3 categories; non-smoker, ex-smoker and current smoker) so a normal distribution is inappropriate for imputing missing smoking status values. Additionally, when choosing an appropriate approach to impute smoking status, I considered how smoking status changes over time because observed values depend on previously recorded values. (For example, if a patient is recorded as a current smoker, I did not want to impute the patient as a non-smoker at a future time point.)

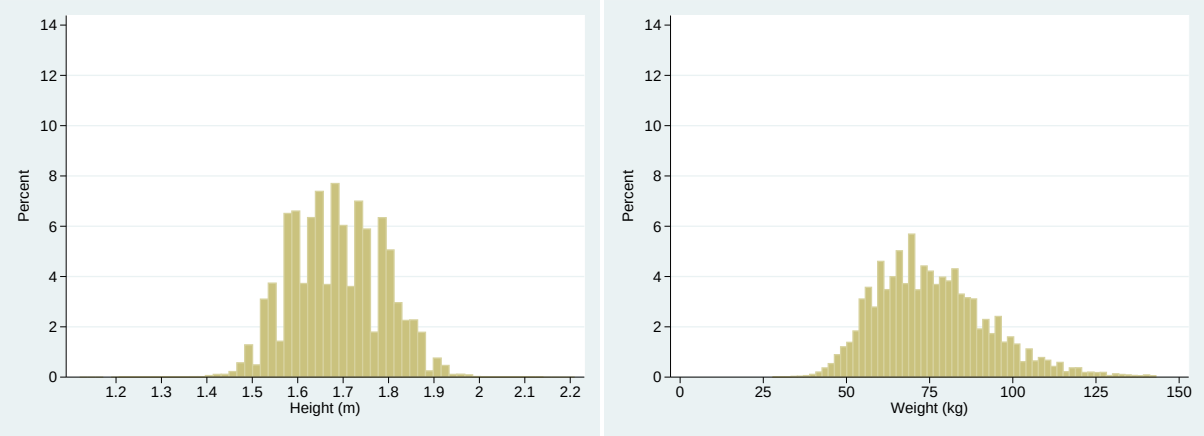

Figure 5.1: Distribution of height and weight measurements recorded at registration

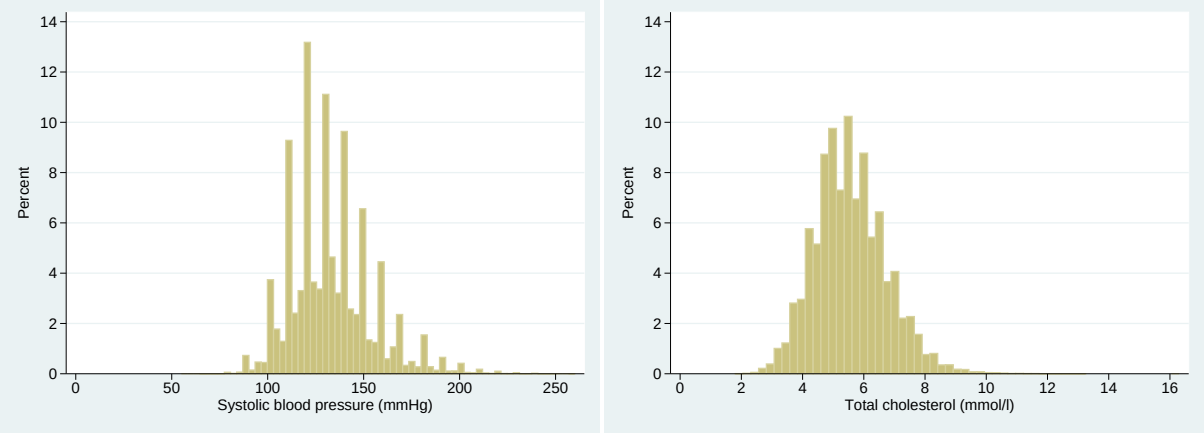

Figure 5.2: Distribution of systolic blood pressure and total cholesterol measurements recorded at registration

A few studies used MI to impute missing smoking status in primary care databases, but the authors did not compare the results after imputation with population data so they may not be representative[9, 8, 93]. Some studies dichotomised smoking status into 'never' and 'ever' smokers because of difficulties defining ex-smokers, even though this approach is subject to misclassification and loses the richness of the variable[9, 8]. Marston *et al.*[13] imputed missing health indicator values for newly registered patients in THIN and compared the results to the data recorded in the Health Survey for England (HSE). This study found that the distribution of each continuous

health indicators in THIN was similar to the same variables in the HSE, but not smoking status (with 3 categories) and recommended further investigation to determine the best approach to impute missing smoking status[13].

For this thesis, I investigated smoking status recording in detail to determine a suitable imputation approach. Smoking status recording at registration improved since the introduction of new patient health checks[94] and now regularly recorded for patients diagnosed with a QOF specified disease since the introduction of QOF[95] in 2004, but not regularly recorded for those without a QOF specified disease (discussed in section 4.3). However, the QOF guidelines regarding smoking status recording changed since their introduction in 2004. Originally, GPs were not required to record smoking status if the patients previous smoking status was non-smoker. This was amended in 2008 so GPs were required to record smoking status for all patients diagnosed with CHD, stroke, hypertension, diabetes, COPD, asthma, CKD or psychosis, regardless of previous smoking status. Therefore, the reason and incentives for recording smoking status changed over time.

From discussing smoking habits with GPs, I discovered GPs recorded only a few smoking status records for non-smokers during follow-up because they considered it unlikely that adult non-smokers would start smoking so thought it unnecessary to update smoking status regularly for known non-smokers. However, ever smokers may change smoking status so GPs update smoking status regularly for these patients; ex-smokers would start smoking again and current smokers would stop smoking.

I investigated if smoking status recording in THIN fitted these patterns. I extracted the data described in section 4.1 for this analysis, from 1 January 1995 to 31 December 2009 for patients aged between 18 and 100 years with at least two years follow-up. I investigated smoking status recording each year and compared these percentages to the HSE 2006 survey[77]. First, I calculated the percentage of patients in THIN with each smoking status per year. Next, I changed the records such that patients who only ever had a non-smoker record were non-smokers at all time points and recalculated the percentage with each smoking status (Figure 5.3). The recalculation increased the percentage of non-smokers to 45-50% each year, slightly less than the percentage of non-smokers in HSE in 2006 (52%)[77], which provides additional support for the GPs' hypothesis. Therefore, from these findings, before imputing smoking status, assuming patients who only ever had a non-smoker record were non-smokers at all times points captures most of the non-smoker records and greatly reduces the missing data from 60%-90% to 40%-50%.

The percentage of ex-smokers and current smokers in HSE was respectively 26% and 22%. Since these percentages were much higher than those observed in THIN (Figure 5.3), I imputed the remaining missing smoking status values to either ex-smoker or current smoker.

Data generation mechanism II

I used a second data generation mechanism, identical to the first except for the way smoking status was generated and conditioned on, to investigate if estimates were less biased if I imputed based on the findings in section 5.1.1.2. Specifically, I used step 3 from Data generation mechanism I, as described previously, to generate smoking status

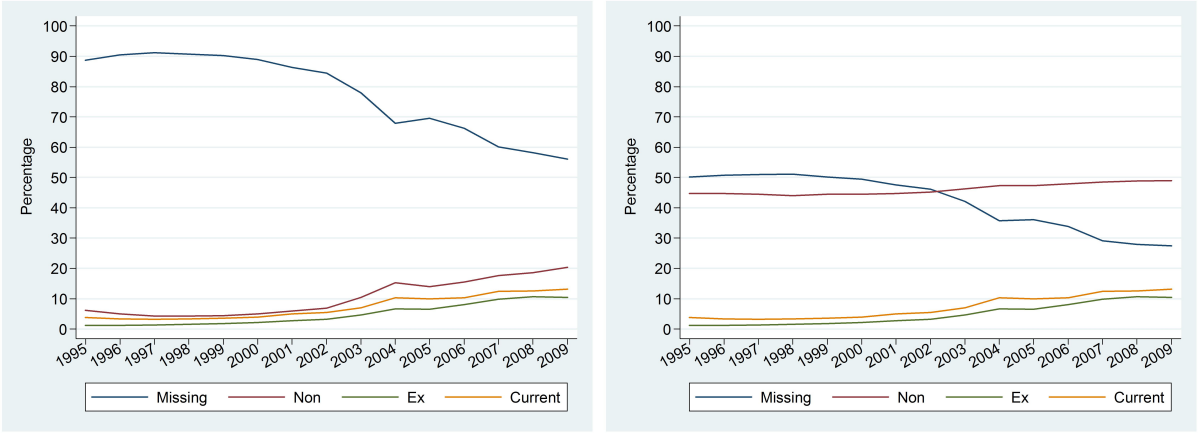

Figure 5.3: Percentage of patients aged 40 years and over with each smoking status per year. The figure on the left shows the distribution of smoking status records and patients missing smoking status in THIN from 1995 to 2009. The figure on the right shows the same distributions when patients who only ever had non-smoking records are assumed to be non-smokers in all years.

in the first time block (year 2000), $smoke_{i,0},$. Patients who were non-smokers in that time block were marked as non-smokers for the next time blocks of follow-up. For the remaining patients (smokers or ex-smokers) for $t = 1, \ldots, 9$ we generated

$$\Pr(smoke_{i,t} = 2 | smoke_{i,1} \neq 1) = \begin{cases} s_{2,t-1} \text{ if } smoke_{i,t-1} = 2 \text{ and} \\ s_{3,t-1} \text{ if } smoke_{i,t-1} = 3. \end{cases}$$

where $b = 1$ (non-smoker - reference category), 2 (ex-smoker), 3 (current smoker)

The remaining steps were the same as data generating mechanism I, except when I generated the values of the other time-dependent variables at time block t, smoking status at time block t, $smoke_{i,t}$ was conditioned on. Townsend deprivation score quintile and the time to CHD were generated conditional on smoking status at time block $t = 0$.

5.1.1.3 Estimating parameters for the data generation mechanism

I derived the parameter values (the 'true' data generation parameter values for the simulation) for each data generation step in the algorithms given above, using appropriate regression, from a cohort of patients extracted from THIN.

The cohort included 321,780 male patients, permanently registered to practices contributing to THIN, aged between 40 and 89 in the first time block $t = 0$ without any CHD events before 1 January 2001. I extracted the time of registration (dichotomised to either in 1999 or other if not in 1999), age in the first time block $t = 0$ and Townsend deprivation score quintile for these patients.

To ensure smoking status recording was complete in the first time block, I identified all smoking status records

between the years 1995 and 2005. For patients without a record when $t = 0$, I replaced the missing value with the smoking status recorded closest to $t = 0$. I assumed patients without a smoking status record between the years 1995 and 2005 were non-smokers in 2000. If patients recorded more than one smoking status in a time block, I selected the highest risk category (i.e. smokers over ex-smokers) because the CHD risk for patients who change from a current smoker to an ex-smoker within a time block (one year) was probably similar to patients who stay a current smoker.

I extracted all weight and systolic blood pressure measurements, identified outliers[4] using methods described in section 4.4, and replaced them with missing values. If patients recorded more than one weight measurement or systolic blood pressure measurement per time block I selected one at random.

Patients received anti-hypertensive drug treatment during a time block if prescribed two or more of the following drugs during the time block; angiotensin-converting-enzyme (ACE) inhibitors, angiotensin receptor blockser (ARB)-2, thiazide and related diuretics, calcium-channel blockers or beta-adrenoceptor blocking drugs.

Using a CHD diagnosis Read code list (Appendix I)[96], I identified all CHD diagnoses records between 1 January 2001 and 31 December 2009. I calculated the time to first CHD diagnosis from 1 January 2001 for each patient with one or more CHD read code.

5.1.1.4 Model of interest

I chose an exponential time-to-event model of interest to investigate the association between CHD events (after the year 2000) with age category, smoking status, anti-hypertensive drug treatment, systolic blood pressure, weight and Townsend deprivation score quintile recorded in 2000, the first time block $t = 0$. I fitted the exponential model to the cohort of patients extracted from THIN to obtain the 'true' coefficients (log hazard ratios) associated with each explanatory variable (Table 5.1).

I generated a hazard function of CHD $\log \lambda_i$ for patients $i = 1, \ldots, N$ in each simulated dataset using the log hazard ratios (column labelled "THIN cohort" in Table 5.1) and the time-to-event X_i drawn from an exponential distribution $X_i \sim exp[\lambda_i]$. The cumulative distribution function of the random variable X_i follows a uniform distribution, $F_X(X_i) \sim U[0, 1]$. Therefore, I generated the survival time X_i for each patient as follows:

$$\begin{aligned} 1 - e^{-\lambda_i X_i} &\sim U \\ e^{-\lambda_i X_i} &\sim U \\ -\lambda_i X_i &\sim \log U \\ X_i &\sim \frac{-\log U}{\lambda_i} \\ i &= 1, \ldots, N \end{aligned}$$

where U was a value selected at random from the uniform distribution to generate event times[97]. Patients with time-to-event less than 9 time blocks were not censored; those with a time-to event of greater than 9 time blocks were censored at 9 time blocks. Approximately 10% of patients in the simulated dataset had a CHD event. All other patients were censored after 9 time blocks; the end of follow-up. For the purpose of this study, transfer out of the practice and death were ignored.

5.1.1.5 Missingness mechanism

I simulated datasets with no missing values for $N = 5,000$ patients with records from time blocks $t = 0, ..., 9$. Of the variables in the model of interest, age, Townsend deprivation score quintile and anti-hypertensive drug treatment were fully observed. For the other variables, I defined a missingness process (or alternatively observation process) from observed recording in THIN (arising from GP consultations), but assuming a MCAR missingness mechanism.

Initially, I considered changing 70% of values for each variable to missing at each time block. However, the imputations did not converge because of the sparsely observed data. Instead, for each patient and time block, I simulated a binary indicator 'consulted', with probability 0.3 of consultation, as approximately 30% of patients attend a consultation each year, which allowed me to reproduce consultation rates in the simulated data (i.e. information recorded for patients who consulted and not recorded for patients who did not consult). Conditional on 'consulted' in a given time block, I (independently) changed weight, systolic blood pressure and smoking status values to missing with probability 0.05. Under this mechanism, the expected years between weight observations is $1/(0.3 \times 0.95) = 3.5$ years, consistent with missing data in electronic health records like THIN.

5.1.1.6 Imputation strategies

I created the Stata command, `twofold`, which implements an extended version of the two-fold FCS algorithm to accommodate multiple imputation of longitudinal, clinical records in large datasets. The previous implementation in SAS was too inflexible for use in this context because it imputed time-dependent variables only, all subjects entered and exited the study at the same time point and only validated using data with 3 time points. This more flexible implementation imputes time-independent variables, increases the time window width and imputes missing data for subjects with different entry and exit time points. The Stata Journal published a paper describing the `twofold` command (Appendix J)[2].

I investigated the following imputation approaches to impute missing data:

1. The FCS imputation model imputed missing values in all time blocks by entering measurements at each time block as separate variables in the imputation model, the CHD event indicator and time-to-event. I refer to this MI model as the 'full imputation' model, which used 10 cycles and 5 imputations.

2. The FCS imputation model imputed missing values in the baseline time block conditional only on other measurements recorded in the same time block, the CHD event indicator and time-to-event. I refer to this

MI model as the 'baseline imputation' model, which used 10 cycles and 5 imputations.

3. The two-fold FCS algorithm[1] imputed the missing data in all time blocks using 5 imputations. I investigated a few different approaches. Firstly, I imputed using $b_w = 5$ within-time iterations, $b_a = 20$ among-time iterations and $\tau = 1$ time window. I compared the estimates and standard errors to the results following imputation with all combinations of among-time iterations $b_a = 3, 10, 30$ and wider time windows $\tau = 2, 3$. I investigated two different approaches to impute missing smoking status:

   - Method 1 - smoking status recorded in the baseline time block was imputed as a time-independent variable.
   - Method 2 - following the investigations in section 5.1.1.2, I imputed smoking status as time-dependent variable recorded in each time block assuming patients were non-smokers in all time blocks if they only had non-smoker status recorded. Next, I used the two-fold FCS algorithm to impute the remaining missing values in each time block to either ex-smokers or current smokers, which includes imputing all patients with no smoking records at any time blocks to either ex-smokers or current smokers.

After generating data using generation mechanism I and changing values to missing, I compared the bias and precision of estimates from fitting the model of interest to data imputed using all imputation approaches and smoking status method 1. After generating data using generation mechanism II and changing values to missing, I compared the bias and precision of estimates from fitting the model of interest to data imputed using all imputation approaches and smoking status method 2.

5.1.1.7 Statistics used in the evaluation

After each simulation $k = 1, \ldots, s$, I fitted the model of interest to the full data, complete records (with missing values) or data imputed using MI. In each case, I used these results to derive parameter estimates and associated standard errors.

Say $\hat{\theta}_k$ denotes a parameter estimate and associated standard error $\hat{\sigma}_k$. For θ, the true parameter value used in the data generation mechanism, I calculated the following statistics:

1. From Rubin's conditions for proper imputation in section 3.3.5, $\hat{\theta}$ is normally distributed with mean θ. Therefore, we calculate bias as the average of the difference between these across the simulations:

$$\frac{1}{s}\sum_{k=1}^{s}\hat{\theta}_k - \theta$$

2. Empirical variance:

$$\frac{1}{s-1}\sum_{k=1}^{s}(\hat{\theta}_k - \bar{\hat{\theta}})^2$$

   where the average imputed mean across simulations is given by:

$$\bar{\hat{\theta}} = \frac{1}{s}\sum_{k=1}^{s}\hat{\theta}_k$$

3. Estimated variance:
$$\bar{\hat{\sigma}}^2 = \frac{1}{s}\sum_{k=1}^{s}\hat{\sigma}_k^2$$

4. Z-score for bias:
$$Bias(\hat{\theta}) = \frac{\hat{\theta} - \theta}{\sqrt{\hat{\sigma}^2/s}}$$
used to calculate a p-value

5. Mean square error:
$$MSE(\hat{\theta}) = Var(\hat{\theta}) + Bias(\hat{\theta}, \theta)^2 = \bar{\hat{\sigma}}^2 + (\hat{\theta} - \theta)^2$$

6. Confidence interval coverage[52], i.e. the proportion of the s confidence intervals
$$\hat{\theta}_k \pm t_{\delta_k, 0.975}\sqrt{\hat{\sigma}_k^2}$$
that include the true value, θ. δ_k is the degrees of freedom calculated using Rubin's Rules.

7. If N_c is the number of patients included in a complete records analysis with variance $\bar{\hat{\sigma}}_c^2$, the effective number of records gained is:
$$\frac{N_c \times \bar{\hat{\sigma}}_c^2}{\bar{\hat{\sigma}}^2}$$

8. The fraction of missing information is defined in section 3.3.5. A smaller the fraction of missing information reported by a method suggests it recovers more information. If the results show that the fraction of missing information (or the fraction of information lost due to missing data) was small, the estimates were more precise (efficient).

To assess if the imputations were unbiased, averaged imputed variance across simulations should approximately equal the empirical variance, according to Rubin's conditions for proper imputation in section 3.3.5: Draw $w_{i,k} \sim s(\mu, \sigma^2)$ with $i = 1, ..., n$ patients and $k = 1, ..., s$ simulations
Calculate sample means
$$\bar{w}_{.,k} = \frac{1}{n}\Sigma w_{i,k}$$
Get $\bar{w}_{.,1}, \bar{w}_{.,2}, ..., \bar{w}_s$ and these should follow $s(\mu, \frac{\sigma^2}{n})$
Calculate:
$$s_k^2 = \frac{1}{n-1}\Sigma(w_{i,k} - \bar{w}_{.,k})^2$$
Sampling distribution of:
$$\frac{\bar{w}_{.,1} - \mu}{\sqrt{\frac{s_1^2}{n}}}, ..., \frac{\bar{w}_{.,s} - \mu}{\sqrt{\frac{s_s^2}{n}}} \sim t_{n-1}$$
$$s_1^2, ..., s_s^2 \sim \frac{\sigma^2\chi_{n-1}^2}{n-1}$$
So average across simulations to give the simulated variance:
$$\frac{1}{s}\Sigma\frac{s_n^2}{n} = \frac{\sigma^2}{n} \tag{5.1}$$

Alternatively, the Monte Carlo standard error is given by:

$$\frac{1}{s}\Sigma(\bar{w}_{i,k} - \mu)^2 \tag{5.2}$$

Over a large number of simulations (5.1) approximately equals (5.2)[18]. (5.1) was the estimate of (5.2) from the data.

Further, I investigated correlations between weight measurements recorded in different time blocks and systolic blood pressure measurements recorded in different time blocks from the data imputed using the two-fold FCS algorithm to investigate if correlations between repeated measurements were maintained in the imputed data[98]. I compared the correlations after imputation to the correlations from the full simulated data. I investigated the correlations with weight measurements in 2000 and correlations with weight measurements in 2005 because I imputed missing values in time blocks from 2000 to 2009 and I wanted to investigate if different results were found at the first time block (when measurements were after 2000) compared to the middle (when measurements were before and after 2005, so possibly more information was available to more accurately estimate correlations compared to 2000).

5.1.2 Results

The 'full imputation' model, which included a variable for measurements at each time block, did not converge on approximately 25% of datasets because repeated measurements of the same health indicator at different time blocks included in the imputation model caused convergence problems. Therefore, it was not considered further.

The log hazard ratios from fitting the time-to-event model to the full data were similar to the THIN cohort log hazard ratios used to generate the data (Table 5.1).

The log hazard ratios from the complete records analysis were similar to the full data log hazard ratio estimates (Table 5.1). However, the standard error (SE) and empirical SE from the complete records analysis were larger than the standard deviations from the full data (Table 5.2). The relative % bias was small for most estimates from the complete records analysis and the coverages were all close to 95% (Table 5.3), suggesting a negligible bias in the estimates from the complete records analysis. Weight and systolic blood pressure had a larger relative % bias, possibly artificially inflated because the weight and systolic blood pressure estimates were small relative to their standard errors (Table 5.3). The p-value for Z score for bias was greater than 0.05 for all but one of the estimates, age group 40-44. In summary, the estimates from a complete records analysis had a small bias and, although larger than for the original full data, it is not more than we would expect by chance. However, I do not expect bias from complete records analysis as data were made MCAR. The inflated SEs occur because of less power due to the reduced sample size.

Table 5.1: Log hazard ratios from fitting an exponential model to predict risk of coronary heart disease to the THIN cohort, full simulated data, complete records analysis after full simulated datasets were changed to missing, imputed data using baseline imputation and imputed data using the two-fold fully conditional specification algorithm Method 1 and Method 2 with 20 among-time iterations, 5 within-time iterations.

Variables		**THIN cohort**	**Full data**	**Complete case**	**Baseline imputation**	**Two-fold FCS**	
						Method 1	**Method 2**
Townsend deprivation score quintile	1	Reference					
	2	0.1520	0.1503	0.1425	0.1497	0.1498	0.1588
	3	0.2377	0.2367	0.2431	0.2366	0.2366	0.2422
	4	0.2433	0.2400	0.2279	0.2391	0.2401	0.2535
	5	0.4034	0.4024	0.3935	0.4020	0.4023	0.4017
Weight (kg)		0.0019	0.0019	0.0015	0.0016	0.0019	0.0017
Systolic blood pressure (mmHg)		0.0048	0.0049	0.0051	0.0048	0.0051	0.0053
Anti-hypertensive drug treatment		0.2935	0.2868	0.2852	0.2897	0.2855	0.2915
Smoking status	Non-smoker	Reference					
	Ex-smoker	0.0679	0.0692	0.0633	0.0672	0.0579	0.0567
	Current smoker	0.2386	0.2385	0.2307	0.2342	0.2325	0.2261
Age group (years)	40 - 44	-1.2820	-1.2872	-1.3167	-1.2869	-1.2880	-1.2890
	45 - 49	-1.0632	-1.0652	-1.0892	-1.0655	-1.0662	-1.0623
	50 - 54	-0.6402	-0.6392	-0.6467	-0.6398	-0.6408	-0.6330
	55 - 59	-0.3589	-0.3597	-0.3700	-0.3598	-0.3605	-0.3536
	60 - 64	-0.2485	-0.2473	-0.2545	-0.2480	-0.2481	-0.2423
	65 - 69	-0.0396	-0.0416	-0.0470	-0.0418	-0.0409	-0.0348
	70 - 74	Reference					
	75 - 79	0.1108	0.1039	0.1116	0.1043	0.1057	0.1129
	80 +	0.1387	0.1421	0.1255	0.1383	0.1414	0.1368
Constant term		-5.1993	-5.2297	-5.2550	-5.2098	-5.2552	-5.2833

Method 1: Only smoking status recorded in the baseline time block is included as a time-independent variable in the imputation model

Method 2: Smoking status is a time-dependent variable in the imputation model

Table 5.2: Standard deviation (SD) from fitting the exponential model to predict risk of coronary heart disease to the full simulated data and standard errors and empirical standard errors found from fitting the exponential model to the complete records analysis after full simulated datasets were changed to missing, imputed data using baseline imputation and imputed data using the two-fold fully conditional specification algorithm Method 1 and Method 2 with 20 among-time iterations, 5 within-time iterations.

Variables		Full data	Complete records		Baseline imputation		Method 1		Method 2	
		SD	Standard error	Empirical standard error	Standard error	Empirical standard error	Standard error	Empirical standard error	Standard error	Empirical standard error
Townsend deprivation score quintile	1	Reference								
	2	0.1286	0.2528	0.2472	0.1290	0.1304	0.1281	0.1295	0.1249	0.1267
	3	0.1326	0.2621	0.2630	0.1349	0.1366	0.1340	0.1363	0.1266	0.1322
	4	0.1419	0.2793	0.2817	0.1442	0.1462	0.1432	0.1473	0.1341	0.1404
	5	0.1541	0.3087	0.3133	0.1600	0.1627	0.1586	0.1606	0.1535	0.1563
Weight (kg)		0.0032	0.0064	0.0062	0.0063	0.0067	0.0041	0.0041	0.0043	0.0041
Systolic blood pressure (mmHg)		0.0026	0.0055	0.0055	0.0054	0.0056	0.0050	0.0049	0.0053	0.0050
Anti-hypertensive drug treatment		0.0957	0.1923	0.1933	0.1113	0.1133	0.1060	0.1033	0.1109	0.1060
Smoking status	Non-smoker	Reference								
	Ex-smoker	0.1074	0.2117	0.2153	0.2104	0.2289	0.2064	0.2180	0.1109	0.1060
	Current smoker	0.1143	0.2260	0.2302	0.2221	0.2410	0.2161	0.2312	0.1538	0.1489
Age group (years)	40 - 44	0.2311	0.4659	0.4936	0.2484	0.2538	0.2425	0.2448	0.2410	0.2409
	45 - 49	0.2137	0.4321	0.4395	0.2287	0.2328	0.2231	0.2236	0.2228	0.2220
	50 - 54	0.1872	0.3673	0.3762	0.1954	0.2021	0.1907	0.1962	0.1871	0.1895
	55 - 59	0.1734	0.3576	0.3657	0.1867	0.1817	0.1832	0.1791	0.1827	0.1825
	60 - 64	0.1783	0.3589	0.3706	0.1846	0.1848	0.1819	0.1831	0.1753	0.1815
	65 - 69	0.1764	0.3545	0.3671	0.1802	0.1815	0.1790	0.1801	0.1727	0.1784
	70 - 74	Reference								
	75 - 79	0.1914	0.3885	0.3879	0.1998	0.1959	0.1966	0.1955	0.1936	0.1957
	80 +	0.2028	0.4122	0.4272	0.2132	0.2106	0.2076	0.2071	0.2139	0.2103
Constant term		0.4554	0.9369	0.9516	0.8885	0.9092	0.7481	0.7406	0.7930	0.7371

Table 5.3: Relative % bias, Z score for bias and coverage from fitting the exponential model to predict risk of coronary heart disease to the complete records after full simulated datasets were changed to missing, imputed data using baseline imputation

Variables		Complete records				Baseline imputation			
		Relative % bias	**p-value for Z score for bias**	**Mean square error**	**Coverage**	**Relative % bias**	**p-value for Z score for bias**	**Mean square error**	**Coverage**
Townsend	1	Reference							
deprivation	2	-6.285	0.221	0.064	95.4	-1.508	0.578	0.017	94.6
score	3	2.269	0.516	0.069	95.1	-0.469	0.796	0.018	94.0
quintile	4	-6.328	0.084	0.078	94.5	-1.726	0.363	0.021	94.7
	5	-2.454	0.318	0.095	95.4	-0.353	0.782	0.026	94.9
Weight (kg)		-19.772	0.054	0.001	95.3	-15.189	0.169	0.001	88.2
Systolic blood pressure (mmHg)		6.866	0.058	0.001	95.2	1.386	0.710	0.001	89.6
Anti-hypertensive drug treatment		-2.816	0.176	0.037	95.2	-1.266	0.300	0.012	94.2
Smoking	Non-smoker	Reference							
status	Ex-smoker	-6.745	0.501	0.045	94.9	-1.057	0.921	0.044	89.5
	Current smoker	-3.336	0.274	0.051	95.1	-1.874	0.557	0.049	87.8
Age	40 - 44	2.707	0.026	0.218	95.0	0.379	0.545	0.062	95.5
group	45 - 49	2.446	0.061	0.187	96.3	0.216	0.754	0.052	94.8
(years)	50 - 54	1.012	0.586	0.135	95.6	-0.063	0.950	0.038	94.5
	55 - 59	3.100	0.336	0.128	94.9	0.273	0.864	0.035	96.2
	60 - 64	2.422	0.607	0.129	93.6	-0.189	0.936	0.034	95.5
	65 - 69	18.669	0.524	0.126	94.6	5.566	0.701	0.032	95.3
	70 - 74	Reference							
	75 - 79	0.707	0.949	0.151	96.0	-5.883	0.293	0.040	95.6
	80 +	-9.525	0.328	0.170	95.5	-0.296	0.951	0.045	95.7
Constant term		1.072	0.064	0.881	94.0	0.203	0.714	0.790	90.9

Table 5.4: Relative % bias, Z score for bias and coverage from fitting the exponential model to predict risk of coronary heart disease to data imputed using the two-fold fully conditional specification algorithm Method 1 and Method 2 with 20 among-time iterations, 5 within-time iterations.

Variables		**Method 1**				**Method 2**			
		Relative % bias	**p-value for Z score for bias**	**Mean square error**	**Coverage**	**Relative % bias**	**p-value for Z score for bias**	**Mean square error**	**Coverage**
Townsend	1	Reference							
deprivation	2	-1.435	0.594	0.016	95.1	4.442	0.088	0.016	94.7
score	3	-0.502	0.782	0.018	94.5	1.884	0.264	0.016	95.9
quintile	4	-1.346	0.482	0.021	94.1	4.191	0.016	0.018	95.9
	5	-0.278	0.825	0.025	95.3	-0.425	0.724	0.024	94.8
Weight (kg)		-0.634	0.925	0.001	94.2	-9.522	0.176	0.001	92.9
Systolic blood pressure (mmHg)		5.810	0.076	0.001	92.7	10.937	0.002	0.001	91.8
Anti-hypertensive drug treatment		-2.729	0.014	0.011	95.0	-0.664	0.579	0.012	93.7
Smoking	Non-smoker	Reference							
status	Ex-smoker	-14.670	0.149	0.043	90.5	-16.449	0.008	0.018	94.6
	Current smoker	-2.555	0.404	0.047	90.9	-5.265	0.010	0.024	93.6
Age	40 - 44	0.467	0.440	0.059	95.2	0.545	0.359	0.058	95.6
group	45 - 49	0.284	0.669	0.050	95.1	-0.081	0.902	0.050	94.9
(years)	50 - 54	0.101	0.917	0.036	94.3	-1.120	0.226	0.035	95.4
	55 - 59	0.457	0.772	0.034	95.6	-1.462	0.364	0.033	94.1
	60 - 64	-0.158	0.946	0.033	95.7	-2.469	0.269	0.031	96.1
	65 - 69	3.177	0.825	0.032	95.3	-12.205	0.376	0.030	96.0
	70 - 74	Reference							
	75 - 79	-4.623	0.407	0.039	94.6	1.918	0.729	0.037	95.4
	80 +	1.903	0.687	0.043	95.3	-1.392	0.775	0.046	95.1
Constant term		1.075	0.017	0.563	94.3	1.616	0.001	0.636	91.0

The log hazard ratios from fitting the time-to-event model to data imputed using the baseline imputation model were similar to the estimates from the complete records analysis (Table 5.1). The relative % bias's were smaller for baseline imputation compared to complete records analysis, suggesting even less bias in the estimates (Table 5.3). At the same time, the precision of estimates of the baseline imputation log hazard ratios improved (smaller SEs) compared to the complete records analysis for fully observed variables but was similar for imputed variables; weight, systolic blood pressure and smoking status (Table 5.2). For these variables, the empirical SE increased slightly, so the SEs were too small compared to empirical SE and, therefore, the coverages for these estimates were underestimated (Table 5.3). The relative % bias MSEs were smaller following baseline imputation compared to the complete records analysis (Table 5.3), suggesting less unexplained information so the imputed data explained more of the variation. Despite the improved SEs, none of the p-values for z-score were significant because of a concomitant reduction in bias.

The log hazard ratios from the analysis of data imputed using the two-fold FCS algorithm Method 1 with $b_w = 5$, $ba = 20$ and $\tau = 1$ were similar to the analysis of data imputed using baseline imputation (Table 5.1). However, SEs and empirical SEs from analysing data imputed using the two-fold FCS algorithm Method 1 were all smaller, and therefore more precise, compared to baseline imputation (Table 5.2). In particular, the continuous variables with missing data (weight and systolic blood pressure) showed the most reduction in bias, giving coverages closer to the nominal 95% coverage level (Table 5.4). However, the reduction in bias was not observed for the categorical variable with missing data, smoking status. The estimates of SEs and empirical SEs for smoking status were slightly smaller compared to analysing data imputed using baseline imputation (Table 5.2).

The results from analysing data imputed using the two-fold FCS algorithm when using Method 2 were similar to Method 1, except the smoking status log hazard ratios were more precise (Table 5.2). The coverage of the coefficient estimates using Method 2 were closer to 95% and had smaller mean square errors compared to imputing missing data using two-fold FCS algorithm Method 1 (Table 5.4). However, this gain in precision comes at an acceptable price in detectable bias for smoking status.

Only a small effective number of records gained were required for the complete records analysis to achieve the same precision as analysing data imputed using baseline imputation (Table 5.5). This increased significantly for the weight coefficient when analysing of data imputed using the two-fold FCS algorithm Method 1, with not as much of an increase for systolic blood pressure coefficient. The effective number of records gained for the smoking status coefficients was similar for data imputed using baseline imputation and using the two-fold FCS algorithm Method 1, when smoking status was time-independent. However, the effective number of records gained increased for the smoking status coefficients when analysing data imputed using the two-fold FCS algorithm Method 2, when smoking status was time-dependent (Table 5.5).

Table 5.5: Improvement in precision over a complete records analysis. Effective number of records gained for the complete records analysis to achieve the same precision as the analysis following imputation (complete case analysis n=1,278)

Variables	**Complete records**	**Baseline imputation**			**Two-fold FCS algorithm**					
					smoking time-independent			**smoking time-dependent**		
	Standard errors	**Standard error**	**Effective number of records gained**	**% additional records**	**Standard error**	**Effective number of records gained**	**% additional records**	**Standard error**	**Effective number of records gained**	**% additional records**
Weight	0.0064	0.0063	1,328	3.2	0.0041	3,136	143.7	0.0042	2,932	127.8
Systolic blood pressure	0.0055	0.0054	1,335	3.7	0.0050	1,557	21.0	0.0050	1,544	20.0
Smoking status:										
Non-smoker	N/A	N/A			N/A			N/A		
Ex-smoker	0.2117	0.2104	1,303	1.2	0.2064	1,354	5.2	0.1285	3,491	171.3
Current smoker	0.2260	0.2221	1,333	3.5	0.2161	1,408	9.4	0.1536	2,787	116.6

The fraction of missing information halves for the weight coefficient when analysing data imputed using the two-fold FCS algorithm compared to analysing data imputed using baseline imputation (Method 1 and 2). However, only when analysing data imputed using the two-fold FCS algorithm Method 2, when smoking status was time-dependent, reduced the fraction of missing information for the smoking status coefficients compared to analysing data imputed using baseline imputation (Table 5.6). The fraction of missing information was similar from analysing data imputed using any of the three imputation methods for the systolic blood pressure coefficient .

After fitting the model of interest to each dataset imputed using Method 1 with $b_a = 3, 10, 20$ and 30 among-time iterations and $\tau = 1, 2$ and 3 window width, and keeping $b_w = 5$ within-time iterations, I found the results had similar bias and precision to the result described with $\tau = 1$ window width, $b_a = 20$ among-time iterations and $b_w = 5$ within-time iterations.

Table 5.6: Fraction of missing information for each covariate with missing data in the imputed data

Variables		Baseline imputation	Two-fold FCS (smoking time-independent)	Two-fold FCS (smoking time-dependent)
Weight		0.6490	0.3040	0.2882
Systolic blood pressure		0.6365	0.6065	0.6114
Smoking status	Non-smoker			
	Ex-smoker	0.6317	0.6339	0.2803
	Current smoker	0.6314	0.6227	0.3347

5.1.2.1 Correlations

From investigating the correlations between weight measurements at different time blocks, I found the two-fold FCS algorithm with $\tau = 1$ time block window produced imputations which accurately estimated correlations between weight measurements with 1 time block difference. However, the correlations were underestimated when the time between the observations was greater than 1 time block (top graphs, Figure 5.4). Increasing the number of among-time iterations resulted in correlations between all time blocks closer to the correlations in the full data. The biggest improvement was from 3 to 10 among-time iterations, but less improvement from 10 to 20 iterations, and less again from 20 to 30 iterations. For example, with 1 within-time iteration, the correlation between weight in 2000 and weight in 2003 was 0.82 when 3 among-time iterations were used, but the correlation increased to 0.90 with 10 among-time iterations, which is closer to the correlation of 0.94 in the full data. However, with 20 among-time iterations the correlation was 0.92, so there was a 0.08 difference in correlations from 3 to 10 among-time iterations but only 0.04 difference in correlations from 10 to 20 among-time iterations. I found the correlations in the middle of the follow-up time (i.e. correlations with weight measured in 2005) were closer to the correlations from the full data compared to correlations at the beginning of follow-up (i.e. correlations with weight measured in 2000) (Figure 5.4).

With a $\tau = 2$ time block window (i.e. imputation at time block t conditioned on measurements at time blocks $t-2$, $t-1$, $t+1$ and $t+2$), correlations between weight measurements 1 or 2 years apart were accurately estimated (middle graphs, Figure 5.4). With a $\tau = 3$ time block window, correlations were accurately estimated when the difference was 1, 2 or 3 time blocks (bottom graphs, Figure 5.4). These results suggest correlations were closer to the correlations from the full data if the number of time blocks between the measurements was less than or equal to the width of the time block window. However, increasing the window width only gave correlations closer to the correlations from the full data for correlations between nearby time blocks but increasing the number of among-time iterations gave correlations closer to the full data for correlations between all time points.

When $b_a = 3$ or $b_a = 10$, the correlations between weight measurements were not close to the correlations from the full data when $\tau = 1$ and increasing the window width gave correlations closer to the correlations from the full data. This may suggest a lack of convergence with 3 or 10 among-time iterations and additional information at other time blocks were required to get correlations closer to the correlations from the full data.

Increasing the number of among-time iterations did not achieve correlations much closer to the correlations from the full data for correlations between repeated systolic blood pressure measurements, suggesting only a few among-time iterations were required to give correlations close to the correlations from the full data for correlations between systolic blood pressure measurements in data imputed using the two-fold FCS algorithm. However, increasing the time window width used by the two-fold FCS algorithm gave correlations closer to the correlations from the full data for correlations between repeated systolic blood pressure measurements when the specified time block width was equal to or less than the difference between time blocks (Figure 5.5), the same as correlations between repeated weight measurements.

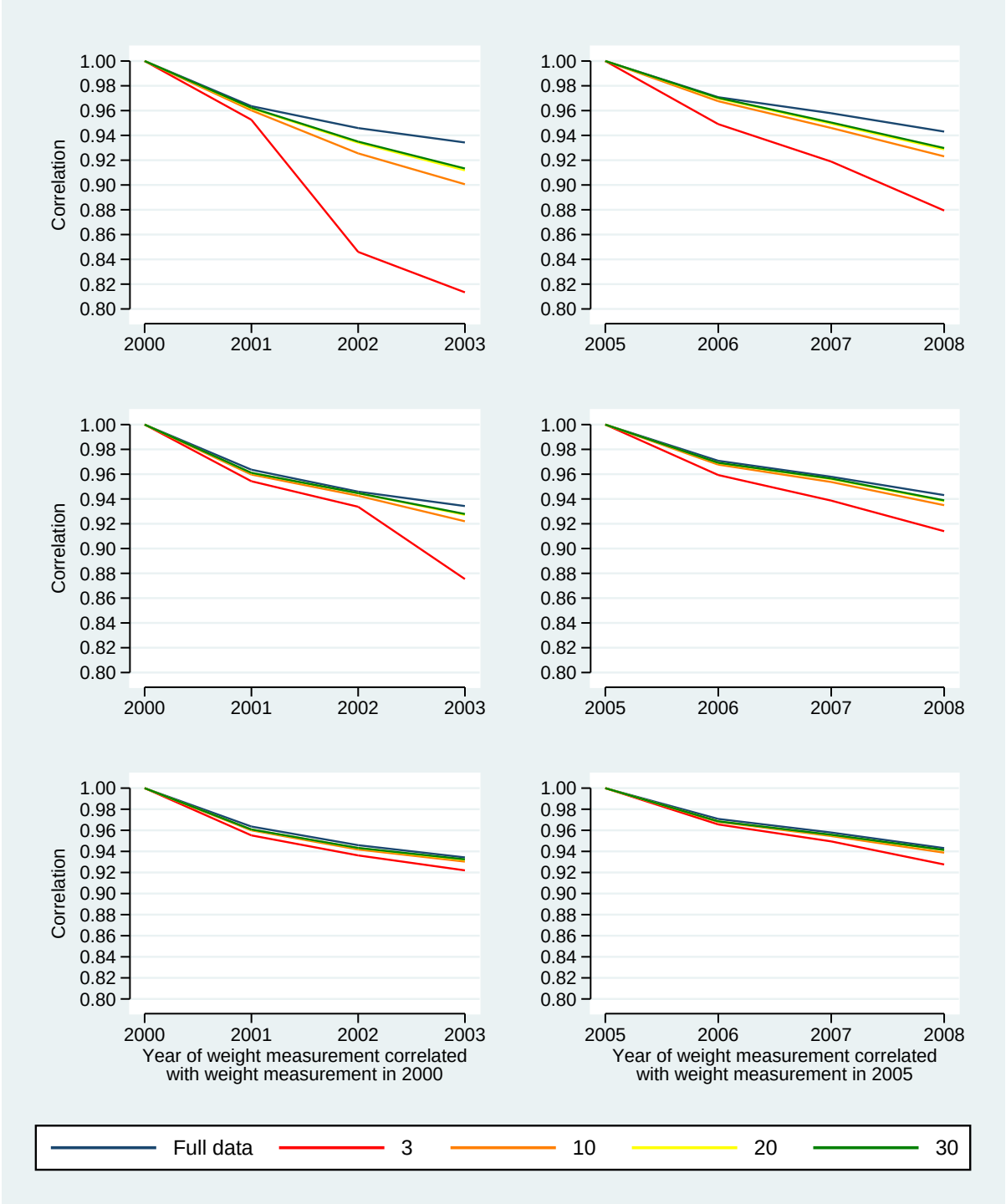

Figure 5.4: Correlations between weight measurements in different time blocks found in the full simulated data and imputed data using the two-fold FCS algorithm with 3, 10, 20 and 30 among-time iterations and 1 (top graphs), 2 (middle graphs) and 3 (bottom graphs) year time block windows.

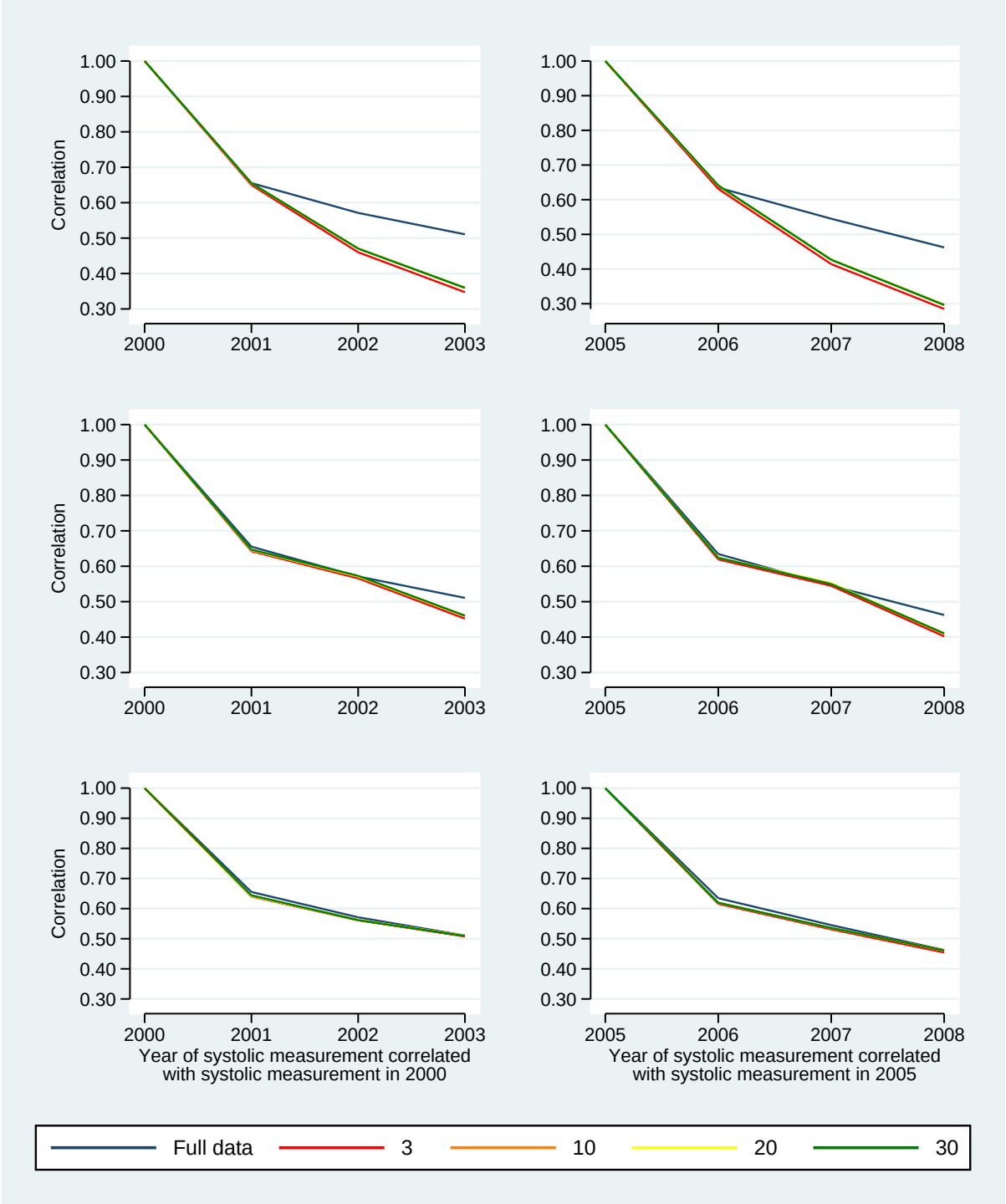

Figure 5.5: Correlations between systolic blood pressure measurements in different time blocks found in the full simulated data and imputed data using the two-fold FCS algorithm with 3, 10, 20 and 30 among-time iterations and 1 (top graphs), 2 (middle graphs) and 3 (bottom graphs) time block windows.

5.1.3 Discussion

This simulation study showed that coefficient estimates were less biased (closer to the 'true' coefficient value) and more precise (smaller standard errors), the effective number of records gained increased significantly for the weight coefficient, with smaller increase for the systolic blood pressure coefficient and the fraction of missing information halves for the weight coefficient from analysing data imputed using the two-fold FCS algorithm compared to data imputed using baseline imputation. This suggests that the two-fold FCS algorithm recovers more of the missing information than baseline imputation, even with 70% missing data in each time block. More information was recovered for the weight variable, with higher correlations between repeated measurements, compared to systolic blood pressure, with lower correlations between repeated measurements.

The bias, precision and number of effective records gained and fraction of missing information from analysing data imputed using the two-fold FCS algorithm with time-independent smoking status were similar to analysing data imputed using baseline imputation. When imputing time-dependent smoking status using the two-fold FCS algorithm, the coefficients of the subsequent analyses were more precise but the smoking status estimates themselves were slightly more biased. The most likely explanation is the approach I used to impute smoking status did not identify all of the non-smokers (because they never had a smoking status recorded, section 5.3) and I may have incorrectly imputed them as either ex-smokers or current smokers. As non-smokers have a lower risk of CVD, these misclassified patients caused the time-to-event model to underestimated the risk associated with ex-smokers and current smokers. Also, the effective number of records gained doubled and the fraction of missing information halved when analysing data imputed using the two-fold FCS algorithm when smoking status was a time-dependent variable compared to a time-independent variable, so the two-fold FCS algorithm recovers more information for time-dependent variables compared to time-independent.

For time-dependent variables, the two-fold FCS algorithm used the information in repeated measurements recorded at different time blocks to inform imputation because they are correlated, but the baseline imputation did not. This enabled the two-fold FCS algorithm to achieve less biased, more precise estimates with increased effective number of records gained and smaller fraction of missing information compared to baseline imputation. This was also true when imputing using the two-fold FCS algorithm with time-dependent smoking status compared to time-independent smoking status.

The estimated correlations between weight measurements recorded in different time blocks were more accurate in data imputed with more among-time iterations. However, correlations estimates between systolic blood pressure measurements in different time blocks were not more accurate with more than 10 among-time iterations. In the full simulated data, the correlations between systolic blood pressure measurements, ranging from 0.45 to 0.65, were lower than the correlations between weight measurements, ranging from 0.95 to 0.98, suggesting accurate correlation estimates were obtained if 'true' correlations were high. If variables had lower correlations between repeated measurements, increasing the number of among-time iterations did not give more accurate correlation estimates because the imputed results can only gain a limited amount of information, implied by the low correlation.

I increased the window width to include variables recorded at more time blocks in the imputation model. This additional information allowed correlations in the imputed data to converge closer to the 'true' correlation for repeated measurements within the specified time window width. This result suggests correlations at greater distance than the window width are systematically underestimated. This can be remedied to a limited extent by increasing the number of among-time iterations, as described in the previous paragraph, but this does not eliminate the bias completely and unable to achieve correlations close to the 'true' correlation if the underlying correlations are lower. In practice, therefore, we should consider a window width the same length as the distance of the correlations we wish to accurately estimate.

5.1.4 Summary

In this section, I used a simulation study to understand and explore the properties of the two-fold FCS algorithm in the context of THIN data, and how (i) the choice of two key parameters (window width and among time iterations) and (ii) the method of imputing smoking (an example of a longitudinal categorical variable) affected bias and precision. Overall, the results showed this method improved bias and precision of time-dependent variables compared to baseline imputation. I investigated the effect of changing the among-time iterations and the window width, which did not effect the bias or precision of the estimates, but more among-time iterations and wider window width achieved correlations closer to the 'true' correlations in the imputed data. Later in the thesis, I impute missing weight, systolic blood pressure and smoking status in THIN using the methods described in this section.

Total cholesterol is another variable I want to include in the model of interest, but my earlier investigation found substantially more missing values in THIN compared to weight, systolic blood pressure and smoking status, as described in section 4. Imputing missing total cholesterol values requires a more detailed investigation so, in the next section, I describe another simulation study to investigate different approaches to impute total cholesterol.

5.2 Simulation study to evaluate imputation strategies for missing total cholesterol

In additional to the health indicators described, I also wish to include total cholesterol in my coronary heart disease (CHD) risk analysis because it strongly predicts CHD events. My investigation in section 4.3 found total cholesterol was recorded most frequently in patients diagnosed with QOF specified diseases, including CHD. But, in this section, my analysis includes predicting a future CHD event in undiagnosed patients with fewer total cholesterol measurements recorded. Also, total cholesterol recording frequency increased for patients prescribed lipid-modifying drugs, compared to patients not prescribed, so GPs can evaluate if the drug lowers total cholesterol effectively.

In this section, I used another simulation study, with missingness patterns and substantial missing values similar to those observed in THIN, to investigate different applications of the two-fold FCS algorithm to impute missing total cholesterol measurements, attempting to optimise the available information.

5.2.1 Methods

To generate the data for this study, I first simulated data using data generating mechanism II described in the previous section, and also simulated variables for lipid-modifying drugs and total cholesterol.

GPs control abnormal cholesterol levels using lipid-modifying drugs. Using THIN data described in section 4.1, I fitted two regression models to the total cholesterol values with age as the dependent variable, one for patients prescribed lipid-modifying drugs and another for patients not prescribed lipid-modifying drugs. The results found, for patients aged 40 years or more, total cholesterol measurements were higher for patients not prescribed lipid-modifying drugs compared to patients prescribed lipid-modifying drugs and total cholesterol decreases for both as age increased (Figure 5.6). These different distributions suggest I should simulate total cholesterol measurements separately for patients prescribed lipid-modifying drugs and patients not prescribed lipid-modifying drugs.

To simulate the data, I first generated baseline age (i.e. age in 2000), denoted age_i, with values $a = 1, ..., 7$ corresponding to the age categories 40-44, 45-49, 50-54, 55-59, 60-64, 65-69 and 70-89.

5.2.1.1 Data generation mechanism

Initially, I simulated total cholesterol measurements separately for patients prescribed any lipid-modifying drugs in the first time block $t = 0$ and not prescribed any lipid-modifying drugs in the first time block $t = 0$, but the associations in the simulated data did not reflect those in the THIN cohort. The problem occurred because only a few patients in the younger age groups were prescribed lipid-modifying drugs and develop CHD (for example, in our sample of 321,831 patients, 9,365 were prescribed lipid-modifying drugs, 524 of these were in age group 40-44 and, of these, 30 had a CHD event) so were not accurately simulated. Instead, I used the following alternative

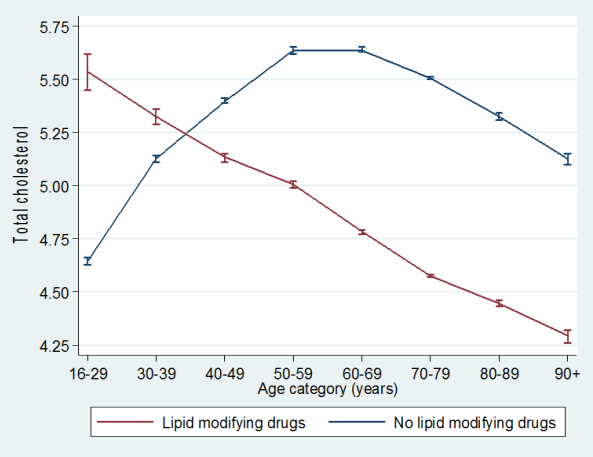

Figure 5.6: Distribution of mean total cholesterol (mmol l^{-1}) by age, for patients prescribed lipid-modifying drugs and not prescribed lipid-modifying drugs

data generating mechanism to generate lipid-modifying drugs and total cholesterol variables:

1. generated binary time-independent lipid-modifying drug treatment variables, denoted $lipid_{i,t}$ for time $t = 1$, from logistic regression models:

$$\begin{aligned} logit(P(lipid_{i,0}=0)) &= \beta_{0,0}^{lipid} + \sum_{a=1}^{7}\beta_{1,0,a}^{lipid}[age_i = a] + \sum_{b=1}^{3}\beta_{2,0,b}^{lipid}[smoke_{i,0} = b] \\ &\quad + \sum_{c=1}^{5}\beta_{3,0,c}^{lipid} townsend_i + \beta_{4,0}^{lipid} systolic_{i,0} \\ &\quad + \beta_{5,0}^{lipid} weight_{i,01} + \beta_{6,0}^{lipid} antihype_{i,0} \end{aligned} \tag{5.3}$$

2. generated time-dependent total cholesterol measurements, denoted $chol_{i,t}$ for calendar years (time blocks) 2000 to 2009 (denoted $t = 0, \ldots 9$). For time block t for patients with $lipid_{i,0} = 0$:

$$\begin{aligned} chol_{i,t} &= \beta_{0,t}^{chol} + \sum_{a=1}^{7}\beta_{1,t,a}^{chol}[age_i = a] + \sum_{b=1}^{3}\beta_{2,t,b}^{chol}[smoke_{i,t} = b] + \sum_{c=1}^{5}\beta_{3,t,c}^{chol}[townsend_i = c] \\ &\quad + \beta_{4,t}^{chol} systolic_{i,t} + \beta_{5,t}^{chol} weight_{i,t} + \beta_{6,t}^{chol} chol_{i,t-1} + \epsilon_{3,t} \end{aligned} \tag{5.4}$$

where $\epsilon_{3,i} \overset{iid}{\sim} N(0, \sigma_{\epsilon_3}^2)$ The difference between the mean total cholesterol values for patients prescribed lipid-modifying drugs and patients not prescribed lipid-modifying drugs was approximately 1 for patients aged 60 years and over (Figure 5.6), so I assumed this difference was true for all ages and drew total cholesterol measurements using the same coefficients for patients with $lipid_{i,0} = 0$ and subtracted 1 from these measurements.

5.2.1.2 Estimating parameters for the data generation mechanisms

Again, as in the previous section, I obtained associated parameters values from the same cohort of patients recorded in THIN. The cohort consisted of male patients, permanently registered with 1 of the 500 general practices contributing to THIN, aged between 40 and 89 in the first time block $t = 0$ without any records of CHD

events before 1 January 2001.

Weight, systolic blood pressure and total cholesterol outliers were identified using the method described previously[4], in section 4.4, and replaced with missing values. Data for continuous variables were extracted using the same methods described in section 5.1. I assumed patients received lipid-modifying drugs during the whole time period if any lipid-regulating drug prescription was recorded during the first time block.

Using a CHD diagnosis Read code list (Appendix I)[96], I identified all CHD diagnosis records between 1 January 2001 and 31 December 2009 and indicated patients with CHD if they had at least one CHD Read code. I also calculated the time-to-event from 1 January 2001 to the first CHD diagnosis date.

The cohort extracted from THIN had missing measurements of weight, systolic blood pressure, smoking status and total cholesterol. I only included patients with complete measurements for all variables in a given time block to estimate parameters in that time block. Again, smoking status was fully observed in the baseline time block, so I estimated parameters using patients with complete measurements of weight, systolic blood pressure and total cholesterol.

5.2.1.3 Model of interest

The model of interest was the same model described in the previous section 5.1, with the additional explanatory variables lipid-modifying drugs and total cholesterol recorded in the first time block $t = 0$.

5.2.1.4 Missingness mechanism

Missingness mechanism 1

After I simulated a completely observed dataset of 5,000 patients with records from 2000 to 2009, I next defined a missingness process (or alternatively observation process) for these simulated data from the observed recordings in THIN (arising from GP consultations). For reasons discussed in section 5.1, I assumed a MCAR missingness mechanism, so the remaining data were a random selection of the full records. Of the variables in the model, age, Townsend deprivation score quintile, anti-hypertensive drug treatment and lipid-modifying drugs were fully observed. Weight, systolic blood pressure, smoking status and total cholesterol missingness was conditional on the patient consulting in the relevant time block. Therefore, for each patient and time block, I simulated a binary indicator 'consulted', with probability 0.3, as approximately 30% of patients attend a consultation each year, of consultation. Conditional on 'consulted' in a given time block, I (independently) changed weight, systolic blood pressure, smoking status and total cholesterol values to missing with probability 0.05. Using this missingnes mechanism, patients have 70% of missing values in for the time-dependent health indicators each year, which approximately agrees with the descriptive analysis findings4.

The only difference between missingness mechanism 1 in this section compared to section 5.1 is total cholesterol with missing values was included in this section.

Missingness mechanism 2

The proportion of total cholesterol measurements recorded in THIN increased following the introduction of QOF, as discussed in section 4.3. Therefore, for missingness mechanism 2, different proportions of total cholesterol measurements were MCAR within each time block for patients who did consult using the following criteria so the percentage missing decreases each time block from 97% in the first time to 70% in the last time block:

$$\begin{aligned} U_i &\sim U(0,1) \\ \text{Total cholesterol is missing for patient } i \text{ if} Pr(chol_{i,t}) = U_i &< (9-t)/10 \\ \textbf{if } t &= 0,\ldots,8 \\ Pr(chol_{i,9}) = U_i &< 0.01 \end{aligned} \tag{5.5}$$

5.2.1.5 Imputation strategies

First, I imputed data made missing by missingness mechanism 1 using imputation method 1. Next, I imputed data made missing by missingness mechanism 2 using each of the following imputation strategies:

Imputation method 1 I imputed missing total cholesterol values using the two-fold FCS algorithm with 20 among-time iterations, 5 within-time iterations, 5 imputations and 1 time block window width, i.e. measurements at time t-1 and t+1 were included in the imputation model (Figure 5.7).

Imputation method 2 I imputed missing total cholesterol values using the two-fold FCS algorithm with a 2 time block window width so measurements at more time blocks included in the imputation model to inform imputation. As well as including measurements at time block t-1 and t+1, measurements at time block t-2 and t+2 were also included in the imputation model to impute missing values at time t (Figure 5.7).

Patient	t-3	t-2	t-1	t	t+1	t+2	t+3
1		5.4	5.5	6.0		6.1	
2	4.5	4.5		4.2	4.1	4.0	
3		5.6			5.0		
4			4.5				
5		3.7	3.5	3.0			4.0
6	4.5			5.1		5.2	4.1
7					5.8		

Figure 5.7: Example of total cholesterol measurements recorded in THIN

Imputation method 3 I imputed total cholesterol with a 1 time block window width. When the two-fold FCS algorithm reached time t, values at adjacent time blocks were temporarily replaced (for this imputation step only) with observed values at the next more extreme time blocks. For example, when imputing at time block t, missing values at t-1 were replaced with observed values at t-2 and missing values at t+1 were replaced with observed values at t+2 (Figure 5.8).

Patient	t-3	t-2	t-1	t	t+1	t+2	t+3
1		5.4	5.5	6.0	6.1 ←	6.1	
2	4.5	4.5 →	4.5	4.2	4.1	4.0	
3		5.6 →	5.6		5.0		
4			4.5				
5		3.7	3.5	3.0			4.0
6	4.5			5.1	5.2 ←	5.2	4.1
7					5.8		

Figure 5.8: Illustration of imputation method 3

Imputation method 4 I imputed total cholesterol with a 1 time block window width. Missing values at time blocks adjacent to the imputed time block were replaced with the closest observed values at more extreme time blocks (Figure 5.9).

Patient	t-3	t-2	t-1	t	t+1	t+2	t+3
1		5.4	5.5	6.0	6.1 ←	6.1	
2	4.5	4.5 →	4.5	4.2	4.1	4.0	
3		5.6 →	5.6		5.0		
4			4.5				
5		3.7	3.5	3.0	4.0 ←		4.0
6	4.5 →		4.5	5.1	5.2 ←	5.2	4.1
7					5.8		

Figure 5.9: Illustration of imputation method 4

For imputation method 4, I included two variables which indicate the time block the measurement in time blocks $t-1$ and $t+1$ were originally observed. The following argument justifies including these variables. The equation $C_t = \beta_0 + \beta_1 t$ models the association between total cholesterol measurements and time. The difference between

total cholesterol values at $t+1$ and $t+2$ was as follows:

$$\begin{aligned} C_{t+1} - C_{t+2} &= \beta_0 + \beta_1(t+1) - \beta_0 - \beta_1(t+2) \\ &= \beta_1((t+1) - (t+2)) \end{aligned} \tag{5.6}$$

and by the same reasoning

$$C_{t-1} - C_{t-2} = \beta_1((t-1) - (t-2)) \tag{5.7}$$

Therefore, we impute missing total cholesterol values at time block t conditional on observed values at $t+1$ and $t-1$ using the result from Equations 5.6 and 5.7 as follows:
Regress C_t on C_{t+1}, C_{t-1} and other variables.
Regress C_t on $C_{t+2} + \beta_1((t+1) - (t+2))$, $C_{t-2} + \beta_1((t-1) - (t-2))$ and other variables.

Where 'other variables' refers to the usual other variables we condition on at the current two-fold FCS algorithm step.

Smoking status was time-dependent so I imputed missing smoking status values using the method for imputing time-dependent smoking status described previously in section 5.1.1.2, patients with only non-smoker records during follow-up were assumed to be non-smokers at all time points and all other missing values were imputed as either ex-smokers or current smokers.

I evaluated the two-fold FCS algorithm using the statistics described in section 5.1.

5.2.2 Results

Table 5.7: Log hazard ratios from fitting the exponential model to the THIN cohort, full simulated data, complete records analysis and data imputed using baseline imputation and imputation method 1 in year 2000 (missingness mechanism 1)

Variables		THIN cohort	Full data	Complete records	Baseline imputation	Method 1
Age group (years)	40 - 44	-1.2305	-1.2671	-1.6946	-1.2797	-1.2673
	45 - 49	-1.0175	-1.0294	-1.1018	-1.0260	-1.0289
	50 - 54	-0.6248	-0.6328	-0.6467	-0.6201	-0.6328
	55 - 59	-0.3236	-0.3256	-0.3278	-0.3107	-0.3264
	60 - 64	-0.2448	-0.2460	-0.2569	-0.2316	-0.2464
	65 - 69	-0.0414	-0.0425	-0.0407	-0.0365	-0.0425
	70 +	Reference				
Townsend deprivation score quintile	1	Reference				
	2	0.0700	0.0783	0.0822	0.0779	0.0777
	3	0.1378	0.1387	0.1416	0.1397	0.1386
	4	0.0843	0.0842	0.0812	0.0857	0.0851
	5	0.3586	0.3686	0.3611	0.3702	0.3686
Weight (kg)		0.0020	0.0019	0.0016	0.0020	0.0019
Systolic blood pressure (mmHg)		0.0041	0.0042	0.0044	0.0042	0.0044
Total cholesterol (mmol l^{-1})		0.0737	0.0728	0.0731	0.0727	0.0719
Smoking status	Non-smoker	Reference				
	Ex-smoker	0.0457	0.0458	0.0501	0.0604	0.0400
	Current smoker	0.1834	0.1819	0.1910	0.1790	0.1743
Lipid-modifying drugs		0.3363	0.3337	0.3313	0.3330	0.3335
Anti-hypertensive drug treatment		0.1655	0.1655	0.1641	0.1650	0.1631
Constant term		-5.4483	-5.4471	-5.4951	-5.4815	-5.4767

From missingness mechanism 1, the total cholesterol coefficients from analysing the complete records (0.0731), after baseline imputation (0.0727) and after imputing using imputation method 1 (0.0719) were similar to the THIN cohort (0.0737) and full data (0.0728) coefficients (Table 5.7). The standard errors from the complete records analysis were larger than the standard errors from the full data analysis because the complete records analysis was based on a smaller sample size (Table 5.8). The results were more precise after baseline imputation and even more precise after imputing using imputation method 1 (the two-fold FCS algorithm with 20 among-time iterations, 5 within-time iterations, 5 imputations and 1 time block window width).

Looking at the total cholesterol coefficients, the relative % biases from all methods were close to zero (Tables 5.9 and 5.10), p-value for Z-score for bias' were greater than 0.05 and the coverage was close to 95%, suggesting negligible bias in the total cholesterol estimates. However, imputation method 1 had the smallest mean square error because the results using imputation method 1 were most precise.

With missingness mechanism 2, the total cholesterol coefficient from the complete records analysis (0.0892) was larger than the THIN cohort (0.0737) and full data (0.0728) estimates (Table 5.11). However, after baseline imputation, the total cholesterol coefficient (0.602) was smaller than the THIN cohort and full data estimates. The

total cholesterol coefficients were more precise after baseline imputation compared to complete records analysis (Table 5.12).

After baseline imputation, the large relative % bias (-18.273) was similar magnitude but opposite signs for the total cholesterol coefficient compared to the complete records analysis relative % bias (21.061) (Table 5.13). The mean square error was smaller for total cholesterol estimate after baseline imputation (0.055) compared to the complete records analysis (0.159) because the estimate was more precise. However, for total cholesterol coefficients from complete records analysis and after baseline imputation, p-value for Z-score for bias' were small and the coverage's were not as close to 95% compared to the full data estimate. The results suggested total cholesterol estimates had some bias when using complete records analysis or baseline imputation with missingness mechanism 2.

With missingness mechanism 2, the total cholesterol coefficient (0.0496) after using imputation method 1 (Table 5.14) was attenuated compared to the total cholesterol estimate after baseline imputation (0.0602) (Table 5.11), but the total cholesterol coefficients from imputation methods 2, 3 and 4 (Table 5.14) were similar to the estimate from baseline imputation (Table 5.11).

The total cholesterol coefficient from imputation method 3 had the smallest standard errors out of the four imputation methods. The standard errors for imputation methods 1, 2 and 4 were similar to baseline imputation (Table 5.15). As results after imputation method 3 were most precise out of baseline imputation and the four imputation methods, it also had the smallest mean square error and coverage closest to 95% (Table 5.16 and 5.17).

Table 5.8: Standard error and empirical standard errors from fitting exponential model to the full data, complete records and data imputed using baseline imputation and imputation method 1 in year 2000 (missingness mechanism 1)

Variables		**Full data**		**Complete records**		**Baseline imputation**		**Method 1**	
		Standard error	**Empirical standard error**	**Standard error**	**Empirical standard error**	**Standard error**	**Empirical standard error**	**Standard error**	**Empirical standard error**
Age group (years)	40 - 44	0.2893	0.2835	2.5836	187.3331	0.3020	0.2995	0.2984	0.2932
	45 - 49	0.2283	0.2163	0.6635	1.0087	0.2383	0.2307	0.2352	0.2269
	50 - 54	0.1536	0.1548	0.3384	0.3204	0.1603	0.1676	0.1634	0.1657
	55 - 59	0.1374	0.1393	0.2783	0.2868	0.1436	0.1476	0.1420	0.1466
	60 - 64	0.1310	0.1287	0.2717	0.2654	0.1307	0.1330	0.1355	0.1339
	65 - 69	0.1232	0.1191	0.2483	0.2448	0.1221	0.1206	0.1253	0.1219
	70 - 74	Reference							
Townsend deprivation score quintile	1	Reference							
	2	0.1184	0.1163	0.2443	0.2388	0.1212	0.1188	0.1187	0.1167
	3	0.1189	0.1211	0.2491	0.2491	0.1205	0.1246	0.1197	0.1219
	4	0.1293	0.1323	0.2726	0.2730	0.1329	0.1370	0.1310	0.1339
	5	0.1446	0.1441	0.3116	0.2984	0.1489	0.1503	0.1476	0.1467
Weight (kg)		0.0030	0.0030	0.0060	0.0061	0.0057	0.0057	0.0037	0.0038
Systolic blood pressure (mmHg)		0.0027	0.0027	0.0053	0.0054	0.0051	0.0051	0.0049	0.0048
Total cholesterol (mmol l^{-1})		0.0379	0.0373	0.0775	0.0763	0.0726	0.0712	0.0682	0.0690
Smoking	Non-smoker	Reference						status	
	Ex-smoker	0.0965	0.0967	0.2089	0.1985	0.1944	0.1841	0.1118	0.1140
	Current smoker	0.1129	0.1126	0.2381	0.2313	0.2223	0.2185	0.1573	0.1528
Lipid-modifying drugs		0.1137	0.1144	0.2474	0.2357	0.1329	0.1327	0.1299	0.1309
Anti-hypertensive drug treatment		0.0908	0.0906	0.1841	0.1855	0.1016	0.1043	0.1029	0.1012
Constant term		0.4930	0.4739	0.9838	0.9692	0.9247	0.8907	0.8061	0.7808

Table 5.9: Relative % bias, Z score for bias, mean square error and coverage from fitting exponential model to the full data and complete records compared to the THIN cohort estimates in year 2000 (missingness mechanism 1)

Variables		**Full data**				**Complete record**			
		Relative % bias	**p-value for Z score for bias**	**Mean square error**	**Coverage**	**Relative % bias**	**p-value for Z score for bias**	**Mean square square**	**Coverage**
Age group (years)	40 - 44	2.9763	<0.0001	0.08501	95.8	37.7240	<0.0001	6.89021	96.5
	45 - 49	1.1724	0.0988	0.05226	93.9	8.2853	0.0001	0.44729	96.2
	50 - 54	1.2780	0.1005	0.02366	95.2	3.5004	0.0412	0.11499	93.5
	55 - 59	0.6180	0.6454	0.01888	95.1	1.2974	0.6334	0.07747	96.1
	60 - 64	0.4866	0.7738	0.01717	94.7	4.9369	0.1598	0.07395	95.7
	65 - 69	2.6152	0.7811	0.01519	94.5	-1.8130	0.9238	0.06164	95.9
	70 +	Reference							
Townsend deprivation score quintile	1	Reference							
	2	11.8566	0.0270	0.01410	94.3	17.4375	0.1145	0.05981	95.0
	3	0.6395	0.8147	0.01413	95.8	2.7406	0.6318	0.06208	94.7
	4	-0.0380	0.9937	0.01673	95.7	-3.6586	0.7207	0.07432	94.3
	5	2.7777	0.0296	0.02101	95.2	0.6829	0.8038	0.09708	95.0
Weight (kg)		-5.1014	0.2704	0.00001	94.7	-22.2600	0.0180	0.00004	95.2
Systolic blood pressure (mmHg)		0.3372	0.8684	0.00001	94.9	6.2387	0.1224	0.00003	95.3
Total cholesterol (mmol l^{-1})		-1.2423	0.4446	0.00143	94.7	-0.8296	0.8029	0.00600	95.1
Smoking status	Non-smoker	Reference							
	Ex-smoker	0.3974	0.9526	0.00932	94.9	9.8234	0.4973	0.04365	94.4
	Current smoker	-0.8162	0.6752	0.01276	95.7	4.1425	0.3132	0.05677	94.6
Lipid-modifying drugs		-0.7819	0.4649	0.01294	96.3	-1.4905	0.5219	0.06125	94.4
Anti-hypertensive drug treatment		-0.0125	0.9943	0.00824	94.4	-0.8644	0.8059	0.03389	95.5
Constant term		-0.0224	0.9377	0.24306	94.7	0.8590	0.1328	0.97005	95.3

Table 5.10: Relative % bias, Z score for bias, mean square error and coverage from fitting exponential model to the data imputed using baseline imputation and imputation method 1 compared to the THIN cohort estimates in year 2000 (missingness mechanism 1)

Variables		**Baseline imputation**				**Method 1**			
		Relative % bias	**p-value for Z score for bias**	**Mean square error**	**Coverage**	**Relative % bias**	**p-value for Z score for bias**	**Mean square square**	**Coverage**
Age group (years)	40 - 44	4.0044	<0.0001	0.09363	96.2	2.9922	0.0001	0.09037	95.8
	45 - 49	0.8433	0.2551	0.05685	94.5	1.1226	0.1250	0.05546	93.9
	50 - 54	-0.7541	0.3528	0.02571	95.8	1.2761	0.1230	0.02675	95.4
	55 - 59	-3.9919	0.0045	0.02079	95.3	0.8779	0.5272	0.02018	95.6
	60 - 64	-5.3787	0.0015	0.01725	95.3	0.6415	0.7141	0.01836	94.8
	65 - 69	-11.8149	0.2053	0.01494	94.6	2.6845	0.7790	0.01569	94.9
	70 +	Reference							
Townsend deprivation score quintile	1	Reference							
	2	11.3102	0.0392	0.01475	93.9	10.9992	0.0406	0.01414	94.3
	3	1.3707	0.6204	0.01453	96.4	0.5836	0.8318	0.01432	95.8
	4	1.7144	0.7312	0.01767	95.7	0.9479	0.8472	0.01717	95.4
	5	3.2389	0.0138	0.02231	95.0	2.7726	0.0333	0.02187	95.1
Weight (kg)		-2.2424	0.8026	0.00003	93.3	-7.6194	0.1928	0.00001	94.5
Systolic blood pressure (mmHg)		0.1703	0.9653	0.00003	93.4	5.8801	0.1193	0.00002	91.3
Total cholesterol (mmol l^{-1})		-1.3916	0.6551	0.00527	93.6	-2.5120	0.3909	0.00466	93.5
Smoking status	Non-smoker	Reference							
	Ex-smoker	32.3589	0.0164	0.03802	93.5	-12.4460	0.1082	0.01252	95.9
	Current smoker	-2.4069	0.5302	0.04945	93.3	-4.9719	0.0670	0.02481	93.7
Lipid-modifying drugs		-0.9652	0.4400	0.01767	94.7	-0.8375	0.4932	0.01689	95.8
Anti-hypertensive drug treatment		-0.3262	0.8666	0.01033	94.9	-1.4306	0.4670	0.01059	94.6
Constant term		0.6091	0.2567	0.85610	91.9	0.5209	0.2658	0.65060	92.5

Table 5.11: Log hazard ratios from fitting the exponential model to the THIN cohort, full simulated data, complete records analysis and data imputed using baseline imputation in year 2000 (missingness mechanism 2)

Variables		**THIN cohort**	**Full data**	**Complete records**	**Baseline imputation**
Age group (years)	40 - 44	-1.2305	-1.2671	-10.9481	-1.2861
	45 - 49	-1.0175	-1.0294	-8.3496	-1.0281
	50 - 54	-0.6248	-0.6328	-3.5515	-0.6205
	55 - 59	-0.3236	-0.3256	-2.2484	-0.3121
	60 - 64	-0.2448	-0.2460	-1.5564	-0.2317
	65 - 69	-0.0414	-0.0425	-0.3912	-0.0373
	70 +	Reference			
Townsend deprivation score quintle	1	Reference			
	2	0.0700	0.0783	-0.2695	0.0772
	3	0.1378	0.1387	-0.4567	0.1362
	4	0.0843	0.0842	-0.8428	0.0867
	5	0.3586	0.3686	-1.6096	0.3665
Weight (kg)		0.0020	0.0019	0.0031	0.0021
Systolic blood pressure (mmHg)		0.0041	0.0042	0.0066	0.0042
Total cholesterol (mmol l^{-1})		0.0737	0.0728	0.0892	0.0602
Smoking status	Non-smoker	Reference			
	Ex-smoker	0.0457	0.0458	0.0991	0.0536
	Current smoker	0.1834	0.1819	-0.2288	0.1856
Lipid-modifying drugs		0.3363	0.3337	-0.7624	0.3216
Anti-hypertensive drug treatment		0.1655	0.1655	0.2110	0.1651
Constant term		-5.4483	-5.4471	-6.9917	-5.4859

Table 5.12: Standard error and empirical standard errors from fitting exponential model to the full data, complete records and data imputed using baseline imputation in year 2000 (missingness mechanism 2)

Variables		**Full data**		**Complete records**		**Baseline imputation**	
		Standard error	**Empirical standard error**	**Standard error**	**Empirical standard error**	**Standard error**	**Empirical standard error**
Age group (years)	40 - 44	0.2893	0.2835	7.843	2650.016	0.3293	0.3416
	45 - 49	0.2283	0.2163	8.3339	1628.772	0.2661	0.2681
	50 - 54	0.1536	0.1548	6.8851	657.7762	0.1877	0.1999
	55 - 59	0.1374	0.1393	6.199	447.0673	0.1588	0.1693
	60 - 64	0.131	0.1287	5.3736	320.8139	0.1409	0.1448
	65 - 69	0.1232	0.1191	3.7199	150.6511	0.1251	0.1252
	70 - 74	Reference					
Townsend deprivation score quintile	1	Reference					
	2	0.1184	0.1163	3.8143	138.9229	0.1398	0.1545
	3	0.1189	0.1211	4.5962	200.4907	0.1446	0.1634
	4	0.1293	0.1323	4.9863	272.6719	0.1576	0.1792
	5	0.1446	0.1441	6.3795	582.9578	0.1726	0.2077
Weight (kg)		0.003	0.003	0.0322	0.0241	0.006	0.0062
Systolic blood pressure (mmHg)		0.0027	0.0027	0.0282	0.0214	0.0056	0.0057
Total cholesterol (mmol l^{-1})		0.0379	0.0373	0.398	0.3008	0.2335	0.2239
] Smoking status	Non-smoker	Reference					
	Ex-smoker	0.0965	0.0967	1.6596	21.6925	0.1869	0.1928
	Current smoker	0.1129	0.1126	3.4183	108.2421	0.2258	0.2363
Lipid-modifying drugs		0.1137	0.1144	4.6431	245.1197	0.2674	0.2649
Anti-hypertensive drug treatment		0.0908	0.0906	1.1306	7.6946	0.1203	0.1344
Constant term		0.493	0.4739	6.0528	96.8013	1.402	1.3547

Table 5.13: Relative % bias, Z score for bias, mean square error and coverage from fitting exponential model to the full data, complete records and data imputed using baseline imputation compared to the THIN cohort estimates in year 2000 (missingness mechanism 2)

Variables		**Full data**				**Complete record**				**Baseline imputation**			
		Relative % bias	**p-value for Z score for bias**	**Mean square error**	**Cover-age**	**Relative % bias**	**p-value for Z score for bias**	**Mean square error**	**Cover-age**	**Relative % bias**	**p-value for Z score for bias**	**Mean square error**	**Cover-age**
Age group (years)	40 - 44	2.9763	<0.0001	0.08501	95.8	789.755	<0.0001	155.945	96.6	4.5241	<0.0001	0.11156	96.4
	45 - 49	1.1724	0.0988	0.05226	93.9	720.628	<0.0001	123.215	97.6	1.0446	0.2069	0.07092	95.6
	50 - 54	1.2780	0.1005	0.02366	95.2	468.428	<0.0001	55.970	96.9	-0.6903	0.4677	0.03526	96.8
	55 - 59	0.6180	0.6454	0.01888	95.1	594.869	<0.0001	42.133	97.0	-3.5352	0.0229	0.02534	95.8
	60 - 64	0.4866	0.7738	0.01717	94.7	535.831	<0.0001	30.596	96.9	-5.3269	0.0035	0.02003	95.8
	65 - 69	2.6152	0.7811	0.01519	94.5	844.327	0.0030	13.960	95.8	-9.9651	0.2970	0.01567	95.1
	70 +	Reference											
Townsend deprivation score quintile	1	Reference											
	2	11.8566	0.0270	0.01410	94.3	-485.210	0.0050	14.664	96.1	10.3583	0.1015	0.01960	96.0
	3	0.6395	0.8147	0.01413	95.8	-431.430	<0.0001	21.479	96.0	-1.1614	0.7264	0.02090	97.6
	4	-0.0380	0.9937	0.01673	95.7	-1100.330	<0.0001	25.723	96.2	2.8886	0.6254	0.02484	97.6
	5	2.7777	0.0296	0.02101	95.2	-548.820	<0.0001	44.573	96.5	2.1980	0.1491	0.02987	98.0
Weight (kg)		-5.1014	0.2704	0.00001	94.7	53.084	0.2924	0.001	92.7	4.2572	0.6500	0.00004	94.8
Systolic blood pressure (mmHg)		0.3372	0.8684	0.00001	94.9	58.708	0.0064	0.001	93.2	2.2933	0.5916	0.00003	94.2
Total cholesterol (mmol $l^{-}1$)		-1.2423	0.4446	0.00143	94.7	21.061	0.2177	0.159	91.9	-18.2734	0.0684	0.05468	90.9
Smoking staus	Non	Reference											
	Ex	0.3974	0.9526	0.00932	94.9	117.092	0.3086	2.757	94.4	17.3383	0.1807	0.03498	94.5
	Current	-0.8162	0.6752	0.01276	95.7	-224.760	0.0001	11.855	95.9	1.2118	0.7556	0.05098	94.3
Lipid-modifying drugs		-0.7819	0.4649	0.01294	96.3	-326.710	<0.0001	22.765	94.0	-4.3781	0.0820	0.07174	94.9
Anti-hypertensive drug treatment		-0.0125	0.9943	0.00824	94.4	27.479	0.2036	1.280	92.8	-0.2475	0.9143	0.01447	97.4
Constant term		-0.0224	0.9377	0.24306	94.7	28.328	<0.0001	39.018	91.6	0.6900	0.3967	1.96712	92.9

Table 5.14: Log hazard ratios from fitting the exponential model to the THIN cohort, full simulated data, and data imputed using imputation methods 1, 2, 3 and 4 in year 2000 (missingness mechanism 2)

Variables		**Method 1**	**Method 2**	**Method 3**	**Method 4**
Age group (years)	40 - 44	-1.2705	-1.2706	-1.2745	-1.2703
	45 - 49	-1.0195	-1.0188	-1.0338	-1.0398
	50 - 54	-0.6284	-0.6320	-0.6339	-0.6353
	55 - 59	-0.3184	-0.3207	-0.3272	-0.3268
	60 - 64	-0.2472	-0.2441	-0.2442	-0.2478
	65 - 69	-0.0429	-0.0466	-0.0411	-0.0425
	70 +	Reference			
Townsend deprivation score quintile	1	Reference			
	2	0.0784	0.0812	0.0799	0.0828
	3	0.1403	0.1357	0.1364	0.1374
	4	0.0900	0.0818	0.0818	0.0825
	5	0.3715	0.3708	0.3676	0.3691
Weight (kg)		0.0018	0.0018	0.0021	0.0021
Systolic blood pressure (mmHg)		0.0043	0.0043	0.0041	0.0039
Total cholesterol (mmol l^{-1})		0.0496	0.0595	0.0599	0.0613
Smoking status	Non-smoker	Reference			
	Ex-smoker	0.0367	0.0350	0.0402	0.0359
	Current smoker	0.1660	0.1657	0.1718	0.1685
Lipid-modifying drugs		0.3149	0.3184	0.3162	0.3185
Anti-hypertensive drug treatment		0.1639	0.1618	0.1657	0.1693
Constant term		-5.3862	-5.4436	-5.4163	-5.4233

Table 5.15: Standard error and empirical standard errors from fitting exponential model to data imputed using the two-fold FCS algorithm in year 2000 (missingness mechanism 2)

Variables		**Method 1**		**Method 2**		**Method 3**		**Method 4**	
		Standard error	**Empirical standard error**	**Standard error**	**Empirical standard error**	**Standard error**	**Empirical standard error**	**Standard error**	**Empirical standard error**
Age group (years)	40 - 44	0.3297	0.3512	0.3252	0.3523	0.3193	0.3338	0.3299	0.3521
	45 - 49	0.2815	0.2909	0.2799	0.2892	0.2702	0.2704	0.2875	0.2857
	50 - 54	0.2159	0.2298	0.2075	0.2341	0.2033	0.2161	0.2145	0.2301
	55 - 59	0.1810	0.2077	0.1782	0.2047	0.1742	0.1878	0.1853	0.2032
	60 - 64	0.1717	0.1880	0.1727	0.1931	0.1605	0.1720	0.1721	0.1892
	65 - 69	0.1540	0.1730	0.1564	0.1766	0.1447	0.1580	0.1594	0.1766
	70 - 74	Reference							
Townsend deprivation score quintile	1	Reference							
	2	0.1195	0.1214	0.1441	0.1581	0.1348	0.1438	0.1449	0.1577
	3	0.1252	0.1357	0.1455	0.1655	0.1320	0.1516	0.1401	0.1626
	4	0.1437	0.1590	0.1569	0.1782	0.1451	0.1638	0.1545	0.1787
	5	0.1686	0.1839	0.1886	0.2069	0.1682	0.1879	0.1856	0.2093
Weight (kg)		0.0045	0.0046	0.0043	0.0047	0.0041	0.0043	0.0042	0.0045
Systolic blood pressure (mmHg)		0.0054	0.0054	0.0053	0.0056	0.0053	0.0053	0.0054	0.0055
Total cholesterol (mmol l^{-1})		0.2360	0.2184	0.2375	0.2280	0.2017	0.2113	0.2254	0.2268
Smoking status	Non-smoker			Reference					
	Ex-smoker	0.1305	0.1443	0.1339	0.1417	0.1249	0.1311	0.1319	0.1407
	Current smoker	0.1725	0.1815	0.1750	0.1861	0.1708	0.1722	0.1745	0.1832
Lipid-modifying drugs		0.2711	0.2630	0.2740	0.2732	0.2368	0.2507	0.2688	0.2712
Anti-hypertensive drug treatment		0.1189	0.1310	0.1165	0.1325	0.1142	0.1236	0.1178	0.1328
Constant term		1.3026	1.2788	1.3729	1.3222	1.2012	1.2155	1.3016	1.3063

Table 5.16: Relative % bias, Z score for bias, mean square error and coverage from fitting exponential model to data imputed using the two-fold FCS algorithm (Methods 1 and 2) compared to the THIN cohort estimates in year 2000 (missingness mechanism 2)

Variables		**Method 1**				**Method 2**			
		Relative % bias	**p-value for Z score for bias**	**Mean square error**	**Coverage**	**Relative % bias**	**p-value for Z score for bias**	**Mean square square**	**Coverage**
Age group (years)	40 - 44	3.254	0.0001	0.11028	96.9	3.263	0.0001	0.10736	96.8
	45 - 49	0.195	0.8237	0.07922	95.5	0.127	0.8843	0.07833	94.7
	50 - 54	0.579	0.5966	0.04662	96.0	1.156	0.2713	0.04310	96.0
	55 - 59	-1.600	0.3661	0.03279	97.3	-0.898	0.6061	0.03177	97.2
	60 - 64	0.979	0.6591	0.02949	96.3	-0.266	0.9053	0.02984	96.4
	65 - 69	3.648	0.7565	0.02373	97.2	12.385	0.2999	0.02450	96.8
	70 +	Reference							
Townsend deprivation score quintile	1	Reference							
	2	12.076	0.0256	0.01436	94.8	16.006	0.0141	0.02089	96.5
	3	1.814	0.5280	0.01569	96.8	-1.483	0.6572	0.02118	97.3
	4	6.844	0.2048	0.02068	97.0	-2.871	0.6261	0.02463	97.5
	5	3.576	0.0163	0.02859	96.7	3.385	0.0420	0.03571	97.4
Weight (kg)		-8.788	0.2095	0.00002	95.1	-9.730	0.1492	0.00002	95.1
Systolic blood pressure (mmHg)		2.588	0.5319	0.00003	93.4	3.710	0.3585	0.00003	93.1
Total cholesterol (mmol l^{-1})		-32.663	0.0013	0.05628	88.1	-19.327	0.0581	0.05661	89.0
Smoking status	Non	Reference							
	Ex	-19.515	0.0310	0.01710	97.1	-23.278	0.0122	0.01804	95.3
	Current	-9.510	0.0014	0.03007	95.4	-9.647	0.0014	0.03094	95.1
Lipid-modifying drugs		-6.346	0.0130	0.07397	91.9	-5.328	0.0389	0.07539	93.8
Anti-hypertensive drug treatment		-0.956	0.6740	0.01414	96.4	-2.250	0.3125	0.01359	97.0
Constant term		-1.141	0.1317	1.70069	91.6	-0.087	0.9132	1.88486	92.1

Table 5.17: Relative % bias, Z score for bias, mean square error and coverage from fitting exponential model to data imputed using the two-fold FCS algorithm (Methods 3 and 4) compared to the THIN cohort estimates in year 2000 (missingness mechanism 2)

Variables		**Method 3**				**Method 4**			
		Relative % bias	**p-value for Z score for bias**	**Mean square error**	**Coverage**	**Relative % bias**	**p-value for Z score for bias**	**Mean square square**	**Coverage**
Age group (years)	40 - 44	3.581	0.0001	0.10389	96.3	3.240	0.0001	0.11043	96.8
	45 - 49	1.603	0.0567	0.07329	95.0	2.192	0.0143	0.08316	95.5
	50 - 54	1.454	0.1578	0.04141	96.0	1.689	0.1201	0.04610	96.1
	55 - 59	1.107	0.5159	0.03037	95.9	1.003	0.5798	0.03434	95.9
	60 - 64	-0.248	0.9048	0.02576	96.0	1.213	0.5853	0.02962	96.6
	65 - 69	-0.837	0.9396	0.02095	97.1	2.524	0.8357	0.02541	97.2
	70 +	Reference							
Townsend deprivation score quintile	1	Reference							
	2	14.236	0.0197	0.01827	95.7	18.270	0.0054	0.02115	95.9
	3	-1.038	0.7320	0.01743	96.8	-0.265	0.9343	0.01962	97.2
	4	-2.903	0.5941	0.02106	97.1	-2.056	0.7229	0.02386	97.5
	5	2.488	0.0936	0.02836	97.2	2.930	0.0737	0.03455	97.8
Weight (kg)		1.978	0.7551	0.00002	94.5	3.224	0.6231	0.00002	94.5
Systolic blood pressure (mmHg)		-2.099	0.6023	0.00003	92.4	-5.011	0.2254	0.00003	93.9
Total cholesterol (mmol l^{-1})		-18.748	0.0305	0.04086	91.8	-16.895	0.0809	0.05097	89.5
Smoking status	Non	Reference							
	Ex	-11.873	0.1701	0.01562	94.9	-21.265	0.0201	0.01749	95.6
	Current	-6.346	0.0314	0.02931	94.4	-8.144	0.0069	0.03069	95.0
Lipid-modifying drugs		-5.970	0.0074	0.05646	93.7	-5.302	0.0362	0.07259	93.0
Anti-hypertensive drug treatment		0.140	0.9490	0.01304	96.0	2.298	0.3076	0.01389	97.6
Constant term		-0.588	0.3996	1.44393	92.3	-0.460	0.5429	1.69467	91.5

5.2.3 Discussion

When approximately 70% of patients were missing total cholesterol measurements at each time block, enough information was available at adjacent time points to inform imputation using imputation method 1 and achieve estimates more precise than baseline imputation. These findings agree with the results from the simulation study investigating imputing time-dependent weight, systolic blood pressure and smoking status using the two-fold FCS algorithm in section 5.1. A small bias occurred in the lowest age group for the full data compared to the THIN cohort. Therefore, maybe it was preferable to compare the estimates from the analysis of imputed data to the full data values, the average of fitting the model to each generated dataset and our best estimate of the true parameter values under the data generating mechanism. However, the practical conclusions would be the same.

When the percentage of patients missing total cholesterol measurements decreased as time blocks increased from 97% in the first time to 70% in the last time block, the total cholesterol coefficients from analysing data imputed using imputation method 1 were more biased than baseline imputation. However, out of the other imputation methods, the total cholesterol coefficients were similar to baseline imputation but most precise after imputation method 3. Therefore, from these results, I recommend using imputation method 1 to impute missing total cholesterol compared to baseline imputation, but using imputation method 3 for a high percentage of missing values, say 95%.

5.2.4 Summary

I investigated different methods of imputing total cholesterol using two-fold FCS algorithm and concluded imputation method 1 was suitable to apply the two-fold FCS algorithm to achieve accurate imputations. Even with a high proportion of missing data when using missingness mechanism 2, although there is a small bias, results are still clearly practically and clinically useful.

Before this chapter, I discovered the recording frequency of health indicators associated with CVD in THIN, why recording changed over time and by age and sex and identified potential auxiliary variables to condition on when applying multiple imputation to achieve a more plausible MAR assumption. However, I did not know how the bias and precision from analysing data imputed using the two-fold FCS algorithm would compare to other methods for handling missing data, such as a complete records analysis and analysing data imputed using 'baseline imputation'. Also, I did not know how varying parameters associated with the two-fold FCS algorithm (such as window width or among-time iterations) affects the results. Finally, I did not know if analysis of data imputed using the two-fold FCS algorithm when high proportions of data are missing would achieve reasonably unbiased results.

I used my contextual knowledge of THIN to simulate data and apply the two-fold FCS algorithm. The simulated data provided a setting similar to THIN, so I could evaluate the best approach to apply the two-fold FCS algorithm to impute missing data in the actual THIN database. In the analysis of data imputed using the two-fold FCS algorithm, I found that time-dependent variables were more precisely estimated compared to complete

records analysis or data imputed using 'baseline imputation'. Also, correlations between repeated measurements in the imputed data were closer to the correlations in the simulated data when I increased the window width or among-time iterations.

Finally, with large proportions of missing data, estimates from analysing data imputed using the two-fold FCS algorithm had small bias. With more missing data the chance of achieving unbiased estimates using multiple imputation increases, even using the two-fold FCS algorithm. To minimise bias, as missing data increases it becomes more and more imperative the underlying assumptions are adhered to; the MAR assumption is plausible and the imputation model is congenial with the model of interest. However, with more missing data it may be more difficult to ensure MAR assumption plausibility.

In the next chapter I use the findings from this chapter to apply the two-fold FCS algorithm to impute missing data in THIN.

Chapter 6

Application of the two-fold fully conditional specification algorithm in substantive analysis of The Health Improvement Network

Now I have explored and understood recording of health indicators associated with cardiovascular disease (CVD) in THIN and evaluated the two-fold FCS algorithm using simulation studies using coronary heart disease (CHD) as the outcome in the models of interest, the next step is to apply the two-fold FCS algorithm to impute missing data in THIN. In this chapter, I perform two analyses of the THIN data and, because of missing health indicator values, I report the results from analysing data imputed using the two-fold FCS algorithm and compare to other methods for handling missing data to evaluate if two-fold FCS algorithm is the preferred method to impute missing values in longitudinal, clinical data. Even though the simulation studies in chapter 5 I investigated health indicators associated with CHD, in this section I chose to investigate associations with cardiovascular disease (CVD) so the results were comparable to other studies, which tend to report associations between health indicators and CVD rather than CHD.

6.1 Imputing missing data in THIN

In this section, I use THIN to investigate the association between established health indicators and subsequent CVD. I included both total and high-density lipoprotein (HDL) cholesterol because HDL cholesterol may be associated with CVD risk after adjusting for total cholesterol. I previously investigated total cholesterol recording in THIN in chapter 4 to understand when data were missing, but not HDL cholesterol. However, I can assume HDL cholesterol had a similar recording frequency to total cholesterol because they were both measured at the same time, but HDL cholesterol was recorded less frequently. From the results in chapter 4, I found substantial missing values, especially for total cholesterol but, given the results from the previous chapter, I can apply the two-fold FCS algorithm with confidence and compare the findings with 'baseline' (imputing missing values at baseline conditional only on other missing values at baseline) and complete records analysis.

6.1.1 Methods

For this part of my studies, I included patients described in section 4.1, with follow-up for each patient including all valid time from 1 January 1995 to 31 December 2011. Outcome was time from the first year of eligible data to first CVD event and I censored patients if a CVD event did not occur before 31 December 2011. I included patients registered with the practice for at least two years and excluded patients if a CVD event occurred before the first year of patients follow-up.

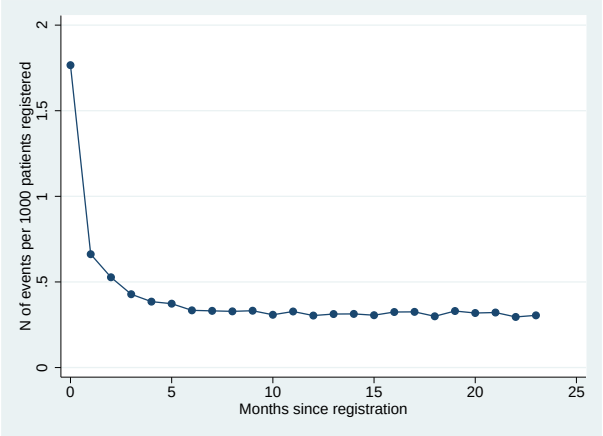

Figure 6.1: Association between number of cardiovascular disease events and the time from registration to first cardiovascular disease event.

I analyse only incident (i.e. new) cases of CVD so, to identify (for subsequent exclusion) likely prevalent (patients with a current rather than new CVD diagnosis) cases, I calculated the time from registration to first CVD event for patients registered during the follow-up period (Figure 6.1). The CVD events peaked at registration, dropped following registration and levelled out after five months. For prevalent CVD cases, the diagnosis was entered onto the patient record when the patient registered with the practice and it appears the patient was diagnosed during or shortly after registration. I excluded patients with a diagnosis of CVD within five months of registration from the analysis because it is likely this peak was due to prevalent CVD[99].

Out of 553 practices, I excluded nine practices because at least one health indicator measurement was missing for more than 95% of patients for most calendar years, or more than 95% of patients were missing more than 95% of measurements for all health indicators for at least three consecutive calendar years.

6.1.1.1 Model of interest

I used an exponential survival model as my model of interest (because I investigated time to an event), to find the association between time from registration to first CVD event and health indicators recorded at baseline, defined as the first year of analysis (i.e. the first year of patient data included in the analysis). When a dataset consists of a large number of patients followed-up over a longer time period, an exponential survival model may be preferable to a Cox model because the assumption of a constant hazard over a longer time period may be violated. The

model comprised of the following variables:

1. Complete variables:
    - Fatal or non-fatal CVD event – either coronary heart disease (Read code list given in Appendix I) or stroke (using QOF Read code list in Appendix B);
    - Time from the end of baseline to the first CVD event;
    - Age at baseline;
    - Sex;
    - Townsend deprivation score quintile - measure of deprivation from 1 (least deprived) to 5 (most deprived).
2. Variables with missing values:
    - BMI - calculated from height and weight (kgm^{-2});
    - Systolic blood pressure (mmHg);
    - Total cholesterol (mmol l^{-1});
    - High-density lipoprotein (HDL) cholesterol (mmol l^{-1}); and
    - Smoking status (current smoker, ex-smoker and non-smoker).

For each calendar year, measurements for each health indicator were extracted. If more than one measurement was recorded during a calendar year, one of the measurements was selected at random for analysis, instead of using the mean of all measurements, to avoid artificially reducing the variability of the data. However, if more than one measurement for smoking status was recorded, the highest risk category was selected (i.e. smokers over ex-smokers) because the risk of CVD during the study duration was likely to remain as high for patients who stop smoking as for those who continue smoking.

I did not include a term for practice in the model of interest as a fixed or random effect because the MI did not converge due to many practices included in the analysis with practice as a fixed effect and currently the two-fold FCS algorithm cannot impute to allow for a random effect. However, I fitted the exponential survival model with robust standard errors which adjusts the SEs to take account of dependence between many patients in the same practice.

6.1.1.2 Analysis

Initially, I fitted the exponential survival model to the observed data at baseline to perform a complete records analysis. Next, I imputed the data by practice and sex stratum using baseline imputation and the two-fold FCS algorithm in each strata separately to allow for possible practice and sex interactions. I imputed missing values at baseline using 5 imputations conditional only on other explanatory variables included in the exponential survival model recorded at baseline and outcome (as CVD event indicator and time-to-event variable) called baseline

multiple imputation (MI). Finally, I imputed longitudinally using the two-fold FCS algorithm for all eligible practices with five imputations, 20 among-time iterations, five within-time iterations and one year window width, conditional on explanatory variables from the model of interest. I also considered including the auxiliary variables listed below, recorded at baseline when included in baseline MI and recorded at each time point when included in the two-fold FCS algorithm:

1. I used QOF Read code lists to identify patients with the following diseases, recorded at any time in the patient record. In section 4.3, I found patients with a diagnosis of these diseases had health indicators recorded more frequently:
    - Diabetes (Appendix C);
    - Chronic obstructive pulmonary disease (COPD, Appendix D);
    - Schizophrenia, bipolar disorder and other psychoses (psychoses, Appendix K);
    - Asthma (Appendix L);
    - Chronic kidney disease (CKD, Appendix M); and
    - Atrial fibrillation (AF, Appendix N).
2. Patients prescribed following drugs were identified:
    - Anti-hypertensive drug treatment - patients received anti-hypertensive drug treatment during a given year if prescribed two-or more of the following drugs during that year; angiotensin-converting-enzyme (ACE) inhibitors, angiotensin receptor blocker (ARB)-2, thiazide and related diuretics, calcium-channel blockers or beta-adrenoceptor blocking drugs.
    - Lipid-modifying drugs - patients received lipid-modifying drugs each calendar year if at least one of the following lipid-regulating drugs was prescribed no more than six months before the total cholesterol measurement in that year; statins, bile acid sequestrants, ezetimibe, fibrate, nicotinic acid or omega 3 polyunsaturated fatty acid ethyl esters.
3. Other auxiliary variables:
    - Respiratory infection – each calendar year, patients had a respiratory infection if at least one Read code from the section of acute respiratory infections was recorded (Appendix O).
    - Alcohol consumption - current heavy drinker if the following was recorded at any time in the patient record: male patients consume more than 5 units of alcohol per day; female patients consume more than 4 units of alcohol per day or at least one Read code which indicated the patient was a heavy drinker (Appendix P).
    - Censored due to non-CVD death – censored patients either transferred out of the practice or died. I created an indicator to identify censored patients who died from non-CVD death.

As explained in section 3.3.6, to obtain more precise estimates I also included auxiliary variables in the imputation models. Therefore, I investigated the auxiliary variables listed above for associations with the value of the

variables with missing data or the probability the value is missing. Before MI, I performed a regression analysis for each health indicator measured at baseline to find which auxiliary variables were associated with the value of the missing health indicators at baseline and a logistic regression analysis for each health indicator to find which auxiliary variables were associated with the probability of the health indicator being missing at baseline. I considered the association significant if p-value$<$0.100 because the auxiliary variables with a weak association can still influence the imputation as well as those variables with a strong association. If these variables were also associated with the probability of the value being missing, including them in the imputation model will increase the MAR assumption plausibility.

I investigated the distributions of the continuous variables in the model of interest and found skewed total cholesterol and HDL cholesterol distributions, so I also entered logged values of total cholesterol and HDL cholesterol in the imputation model.

I also extracted data recorded in the 365 days before the first date patients were included in the analysis to inform imputation of missing values using the two-fold FCS algorithm in the first year of follow-up. In some cases, usually when the baseline year is the year the patients registered with the practice, the two-fold FCS algorithm did not converge because of high proportions of missing data in the year before the first date. When this happened, I re-fitted the imputation model without conditioning on the year before the first date.

Initially, to investigate the computational issues which arose when imputing real data, I imputed missing data using the two-fold FCS algorithm in all practice and sex strata using one imputation, one among-time iteration and one within-time iteration. For any strata that did not converge, I investigated the reasons for non-convergence.

6.1.1.3 Sensitivity analysis

Morris *et al.*[56] recommend imputing logged values of variables included as ratios in the model of interest. Therefore, I repeated the imputation using logged values of height and weight.

6.1.2 Results

In total, 498 practices out of 553 met the inclusion criteria. From investigating potential computational issues, using the two-fold FCS algorithm with one imputation, one among-time iteration and one within-time iteration, I found ten practices had insufficient data recorded in all years and the two-fold FCS algorithm did not converge. I excluded these practices from further analysis.

For 280 practices, some years at the beginning and end of follow-up had insufficient data recorded and the two-fold FCS algorithm did not converge when imputing missing values in these years. Therefore, I changed the start and/or end years for these practices to exclude the years with insufficient data recorded, changing the length of follow-up (Table 6.1). Some of these practices had different first/last years for male and female patients because I investigated them separately.

Table 6.1: Distribution of years of follow-up for the 488 practices before and after excluding years with insufficient data recorded.

Number of years of follow-up	All patients, before exclusions, n (%)	After exclusions, n (%)	
		Male patients	Female patients
3	6 (1.2)	9 (1.8)	9 (1.8)
4	10 (2.1)	15 (3.1)	15 (3.1)
5	12 (2.5)	15 (3.1)	14 (2.9)
6	12 (2.5)	16 (3.3)	19 (3.9)
7	21 (4.3)	30 (6.2)	31 (6.4)
8	35 (7.2)	49 (10.0)	52 (10.7)
9	51 (10.5)	82 (16.8)	77 (15.8)
10	43 (8.8)	70 (14.3)	69 (14.1)
11	49 (10.0)	73 (15.0)	75 (15.4)
12	49 (10.0)	64 (13.1)	61 (12.5)
13	51 (10.5)	29 (5.9)	28 (5.7)
14	28 (5.7)	18 (3.7)	19 (3.9)
15	25 (5.1)	5 (1.0)	6 (1.2)
16	19 (3.9)	2 (0.4)	2 (0.4)
17	77 (15.8)	11 (2.3)	11 (2.3)
Total practice years of follow-up	5,676	4,792	4,783

After I applied these exclusions, 2,772,502 patients (1,347,844 men and 1,424,658 women) remained in the analysis, aged between 30 and 99 years at the baseline year. 50,774 (3.8%) men and 41,836 (2.9%) women had a CVD event during follow-up. Only 45,998 (3.4%) men and 47,371 (3.3%) women had complete records in the baseline year of weight, height, systolic blood pressure, total and HDL cholesterol and smoking status.

From investigating the association between auxiliary variables and missing health indicator or the values themselves at baseline, I found all the variables were associated with the value of at least one health indicator (Tables 6.2 and 6.3). Therefore, I included all auxiliary variables in the baseline MI model and the two-fold FCS algorithm.

In total, 93,369 patients were included in the complete records analysis, which showed CVD risk was greatest for older, male and more deprived patients with higher blood pressure, higher total cholesterol and lower HDL cholesterol. A CVD event was more likely if patients were ex-smokers compared to non-smokers and most likely for current smokers compared to non or ex-smokers (Table 6.4).

Data imputed using baseline MI or two-fold FCS algorithm included 2,772,502 patients. The analysis from fitting the model of interest to the data imputed using baseline MI or the two-fold FCS algorithm found the coefficients for the fully observed variables age, sex and Townsend score quintile were more extreme compared to complete records analysis, suggesting a stronger association between them and CVD risk. The standard errors from analysing data imputed using baseline imputation were all smaller compared to complete records analysis, so these estimates were more precise. This is expected for the fully observed variables because the number of patients included in the analysis increased following imputation compared to the complete records analysis, i.e. more statistical power. From fitting the model of interest to the data imputed using the two-fold FCS algorithm,

Table 6.2: Association between auxiliary variables and the probability of missing health indicator or value of the health indicators height, weight and systolic blood pressure at baseline

Covariates		Height (m)		Weight (kg)		Systolic blood pressure (mmHg)	
		Missing OR (95% CI)	Value Coefficient (95% CI)	Missing OR (95% CI)	Value Coefficient (95% CI)	Missing OR (95% CI)	Value Coefficient (95% CI)
Age at baseline (years)	≥30 and <40	1.09 (1.07, 1.10)	4.79 (4.74, 4.84)	1.07 (1.06, 1.09)	8.51 (8.33, 8.69)	1.62 (1.60, 1.64)	-19.81 (-19.95, -19.68)
	≥40 and <50	0.93 (0.92, 0.94)	3.71 (3.66, 3.76)	1.19 (1.17, 1.20)	10.07 (9.88, 10.25)	1.76 (1.74, 1.78)	-13.52 (-13.66, -13.38)
	≥50 and <60	0.88 (0.87, 0.89)	2.62 (2.57, 2.67)	1.07 (1.05, 1.08)	8.46 (8.28, 8.64)	1.35 (1.33, 1.37)	-7.53 (-7.66, -7.40)
	≥60 and <70	0.80 (0.79, 0.81)	1.67 (1.62, 1.72)	1.01 (1.00, 1.03)	5.30 (5.12, 5.47)	1.17 (1.16, 1.19)	-3.20 (-3.33, -3.07)
	≥70 and <80	Ref	Ref	Ref	Ref	Ref	Ref
	≥80 and <90	1.56 (1.53, 1.58)	-2.18 (-2.25, -2.11)	1.26 (1.24, 1.28)	-6.36 (-6.61, -6.11)	1.10 (1.08, 1.12)	1.40 (1.23, 1.57)
	≥90	2.71 (2.62, 2.80)	-4.17 (-4.36, -3.98)	2.17 (2.08, 2.27)	-11.49 (-12.12, -10.87)	1.58 (1.52, 1.63)	-0.71 (-1.07, -0.34)
Sex		0.76 (0.76, 0.77)	-14.00 (-14.02, -13.97)	0.53 (0.52, 0.53)	-13.53 (-13.63, -13.43)	0.43 (0.43, 0.43)	-4.02 (-4.10, -3.95)
Townsend deprivation score quintile	1	Ref	Ref	Ref	Ref	Ref	Ref
	2	0.95 (0.95, 0.96)	-0.39 (-0.42, -0.36)	0.95 (0.94, 0.96)	0.49 (0.36, 0.62)	1.00 (0.99, 1.01)	0.07 (-0.03, 0.16)
	3	0.95 (0.95, 0.96)	-0.96 (-0.99, -0.93)	0.93 (0.92, 0.94)	0.84 (0.71, 0.98)	1.01 (1.00, 1.02)	0.31 (0.21, 0.41)
	4	0.94 (0.93, 0.94)	-1.58 (-1.61, -1.55)	0.91 (0.90, 0.92)	1.06 (0.92, 1.20)	1.04 (1.03, 1.05)	0.27 (0.17, 0.38)
	5	0.88 (0.88, 0.89)	-2.23 (-2.27, -2.2)	0.87 (0.86, 0.87)	0.64 (0.48, 0.79)	1.07 (1.06, 1.08)	0.37 (0.25, 0.49)
Time to event		0.95 (0.95, 0.95)	-0.10 (-0.10, -0.09)	1.00 (1.00, 1.00)	-0.08 (-0.09, -0.06)	0.96 (0.96, 0.96)	0.56 (0.55, 0.58)
CHD		1.35 (1.32, 1.37)	-0.73 (-0.81, -0.64)	0.94 (0.92, 0.96)	-0.31 (-0.59, -0.03)	0.74 (0.72, 0.75)	3.63 (3.43, 3.84)
Stroke		1.39 (1.36, 1.42)	-0.52 (-0.62, -0.42)	1.02 (0.99, 1.04)	-1.07 (-1.39, -0.75)	0.81 (0.79, 0.83)	3.42 (3.19, 3.65)
Diabetes		0.35 (0.35, 0.36)	-0.37 (-0.41, -0.32)	0.27 (0.27, 0.28)	7.09 (6.95, 7.23)	0.49 (0.48, 0.49)	0.94 (0.82, 1.06)
COPD		0.39 (0.38, 0.36)	-0.79 (-0.85, -0.73)	0.77 (0.76, 0.79)	-3.75 (-3.98, -3.52)	0.86 (0.85, 0.88)	-1.28 (-1.46, -1.10)
Psychosis		0.58 (0.56, 0.61)	-0.25 (-0.39, -0.10)	0.68 (0.65, 0.71)	1.96 (1.40, 2.51)	0.82 (0.79, 0.86)	-2.85 (-3.33, -2.37)
Asthma		0.59 (0.59, 0.59)	-0.41 (-0.44, -0.37)	0.73 (0.72, 0.74)	2.45 (2.32, 2.58)	0.81 (0.80, 0.81)	0.40 (0.29, 0.50)
CKD		0.73 (0.72, 0.74)	0.12 (0.07, 0.17)	0.86 (0.84, 0.87)	1.42 (1.25, 1.59)	0.76 (0.74, 0.77)	1.03 (0.91, 1.16)
Atrial fibrilation		0.91 (0.90, 0.93)	2.02 (1.95, 2.09)	0.98 (0.96, 1.00)	3.59 (3.35, 3.84)	0.81 (0.80, 0.83)	-2.75 (-2.92, -2.58)
Heavy drinker		0.79 (0.78, 0.80)	0.28 (0.25, 0.32)	0.75 (0.74, 0.75)	-0.76 (-0.91, -0.62)	0.78 (0.77, 0.78)	2.14 (2.03, 2.25)
Death		1.60 (1.58, 1.62)	-0.50 (-0.56, -0.43)	1.01 (0.99, 1.02)	-2.04 (-2.25, -1.83)	0.94 (0.92, 0.95)	0.70 (0.54, 0.85)
Anti-hypertensive drug treatment		0.69 (0.68, 0.70)	-0.35 (-0.38, -0.32)	0.48 (0.47, 0.48)	5.82 (5.70, 5.94)	0.14 (0.13, 0.14)	10.53 (10.44, 10.61)
Respiratory infection		0.97 (0.97, 0.98)	-0.12 (-0.15, -0.08)	0.68 (0.68, 0.69)	1.15 (1.02, 1.28)	0.53 (0.53, 0.54)	-0.77 (-0.87, -0.67)
Lipid modifying drug		0.71 (0.70, 0.72)	-0.38 (-0.43, -0.33)	0.50 (0.49, 0.51)	0.30 (0.15, 0.45)	0.45 (0.44, 0.46)	-2.85 (-2.97, -2.73)

OR - odds radio; CI - confidence interval; HDL - high-density liproprotein; CHD - coronary heart disease; COPD - chronic obstructive pulmonary disease; Psychosis - schizophrenia, bipolar disorder and other psychoses; CKD - chronic kidney disease; Ref - reference category

Table 6.3: Association between auxiliary variables and the probability of missing health indicator or value of the health indicators smoking status, total cholesterol and HDL cholesterol at baseline

Covariates		Smoking status		Total cholesterol		HDL cholesterol	
		Missing OR (95% CI)	Value OR (95% CI) *	Missing OR (95% CI)	Value Coefficient (95% CI)	Missing OR (95% CI)	Value Coefficient (95% CI)
Age at baseline (years)	≥30 and <40	1.23 (1.22, 1.25)	4.80 (4.67, 4.93)	3.44 (3.38, 3.50)	-0.43 (-0.45, -0.41)	3.28 (3.22, 3.35)	-0.20 (-0.21, -0.19)
	≥40 and <50	1.25 (1.23, 1.26)	3.95 (3.84, 4.05)	1.70 (1.67, 1.72)	-0.11 (-0.13, -0.10)	1.63 (1.60, 1.65)	-0.16 (-0.16, -0.15)
	≥50 and <60	1.13 (1.12, 1.15)	2.70 (2.64, 2.78)	1.13 (1.12, 1.15)	0.09 (0.08, 0.10)	1.08 (1.06, 1.09)	-0.09 (-0.09, -0.08)
	≥60 and <70	1.01 (1.00, 1.02)	1.64 (1.60, 1.68)	0.91 (0.90, 0.93)	0.07 (0.06, 0.08)	0.88 (0.86, 0.89)	-0.05 (-0.05, -0.05)
	≥70 and <80	Ref	Ref	Ref	Ref	Ref	Ref
	≥80 and <90	1.30 (1.28, 1.32)	0.64 (0.61, 0.66)	1.60 (1.57, 1.63)	-0.03 (-0.05, -0.01)	1.66 (1.62, 1.69)	0.06 (0.05, 0.06)
	≥90	1.97 (1.90, 2.05)	0.51 (0.46, 0.58)	2.76 (2.62, 2.91)	-0.23 (-0.28, -0.18)	2.99 (2.81, 3.17)	0.01 (-0.01, 0.03)
Sex		0.67 (0.67, 0.68)	1.02 (1.01, 1.03)	1.03 (1.02, 1.04)	0.26 (0.26, 0.27)	1.04 (1.03, 1.05)	0.28 (0.28, 0.28)
Townsend deprivation score quintile	1	Ref	Ref	Ref	Ref	Ref	Ref
	2	0.91 (0.90, 0.92)	1.23 (1.21, 1.26)	1.04 (1.03, 1.05)	0.00 (-0.01, 0.01)	1.02 (1.01, 1.03)	-0.01 (-0.01, 0.00)
	3	0.87 (0.86, 0.88)	1.58 (1.55, 1.61)	1.04 (1.03, 1.06)	-0.02 (-0.03, -0.01)	1.03 (1.02, 1.05)	-0.03 (-0.04, -0.03)
	4	0.83 (0.82, 0.84)	2.04 (2.00, 2.08)	1.07 (1.06, 1.09)	-0.02 (-0.03, -0.01)	1.07 (1.06, 1.09)	-0.05 (-0.06, -0.05)
	5	0.78 (0.77, 0.79)	2.70 (2.64, 2.76)	1.06 (1.04, 1.07)	-0.06 (-0.07, -0.04)	1.05 (1.03, 1.07)	-0.07 (-0.08, -0.07)
Time to event		1.03 (1.03, 1.03)	0.99 (0.99, 0.99)	0.99 (0.99, 1.00)	0.05 (0.05, 0.05)	0.99 (0.98, 0.99)	0.00 (0.00, 0.01)
CHD		0.99 (0.97, 1.01)	1.19 (1.14, 1.23)	0.87 (0.85, 0.89)	0.31 (0.29, 0.33)	0.85 (0.83, 0.87)	-0.03 (-0.04, -0.03)
Stroke		1.08 (1.06, 1.10)	1.27 (1.21, 1.33)	1.00 (0.98, 1.03)	0.20 (0.18, 0.22)	0.98 (0.95, 1.01)	0.00 (-0.01, 0.01)
Diabetes		0.53 (0.53, 0.54)	0.82 (0.80, 0.84)	0.31 (0.30, 0.31)	-0.46 (-0.47, -0.46)	0.35 (0.35, 0.36)	-0.16 (-0.16, -0.15)
COPD		0.49 (0.48, 0.50)	2.33 (2.27, 2.39)	1.14 (1.12, 1.17)	-0.06 (-0.08, -0.05)	1.15 (1.13, 1.18)	0.04 (0.04, 0.05)
Psychosis		0.69 (0.66, 0.71)	2.37 (2.18, 2.57)	0.83 (0.79, 0.88)	0.02 (-0.02, 0.07)	0.87 (0.82, 0.92)	-0.09 (-0.11, -0.07)
Asthma		0.54 (0.53, 0.54)	0.71 (0.70, 0.73)	0.88 (0.87, 0.89)	0.02 (0.01, 0.04)	0.89 (0.88, 0.90)	0.02 (0.01, 0.02)
CKD		0.87 (0.86, 0.88)	0.87 (0.85, 0.89)	0.82 (0.81, 0.83)	-0.02 (-0.03, -0.01)	0.86 (0.85, 0.88)	-0.04 (-0.05, -0.04)
Atrial fibrilation		1.02 (1.00, 1.04)	0.72 (0.70, 0.75)	0.98 (0.96, 1.00)	-0.18 (-0.20, -0.17)	1.02 (1.00, 1.04)	-0.01 (-0.02, -0.01)
Heavy drinker		0.69 (0.68, 0.69)	1.45 (1.43, 1.48)	0.79 (0.78, 0.80)	0.16 (0.15, 0.17)	0.82 (0.81, 0.83)	0.14 (0.13, 0.14)
Death		1.14 (1.13, 1.16)	1.39 (1.35, 1.43)	1.19 (1.17, 1.21)	-0.02 (-0.03, 0.00)	1.20 (1.18, 1.22)	-0.01 (-0.02, -0.01)
Anti-hypertensive drug treatment		0.48 (0.47, 0.48)	0.70 (0.69, 0.72)	0.26 (0.26, 0.26)	-0.10 (-0.11, -0.10)	0.30 (0.29, 0.30)	-0.05 (-0.05, -0.05)
Respiratory infection		0.50 (0.50, 0.50)	1.13 (1.11, 1.15)	0.71 (0.70, 0.72)	-0.03 (-0.05, -0.02)	0.78 (0.77, 0.79)	-0.04 (-0.04, -0.03)
Lipid modifying drug		0.55 (0.54, 0.56)	1.03 (1.00, 1.05)	0.15 (0.15, 0.15)	-0.22 (-0.23, -0.21)	0.20 (0.20, 0.21)	-0.03 (-0.03, -0.02)

* From logistic regression model to compare current smokers with ex-smokers (non-smokers excluded because they were not imputed)

OR - odds radio; CI - confidence interval; HDL - high-density liproprotein; CHD - coronary heart disease; COPD - chronic obstructive pulmonary disease; Psychosis - schizophrenia, bipolar disorder and other psychoses; CKD - chronic kidney disease; Ref - reference category

coefficients and standard errors of complete variables were similar to the baseline MI (Table 6.4).

BMI was the only variable with missing values which changed from non-significant for the complete records analysis and analysing data imputed using baseline MI to statistically significant when analysing data imputed using the two-fold FCS algorithm (Table 6.4).

The systolic blood pressure coefficients and standard errors were similar from fitting the model of interest to the data imputed using baseline MI or two-fold FCS algorithm. For total cholesterol, HDL cholesterol and smoking status, the standard errors were smaller from fitting the model of interest to the data imputed using the two-fold FCS algorithm compared to baseline MI, giving the most precise estimates for these variables (Table 6.4).

The complete records analysis found the strongest association between HDL cholesterol and CVD events. This association was weaker when analysing data imputed using baseline MI, and weaker again when analysing data imputed using the two-fold FCS algorithm (although all statistically significant), and the HDL cholesterol IRR from the complete records analysis was outside the HDL cholesterol 95% CI from analysing data imputed using the baseline MI or the two-fold FCS algorithm.

6.1.2.1 Sensitivity analysis

Finally, when I repeated the analysis entering logged values of height, weight, total cholesterol and HDL cholesterol in the imputation model, I found the baseline MI imputed some very extreme values and fitting the model of interest to the imputed data all coefficients were highly significant. I imputed data stratified by practice and sex, and the observed values within some of the strata did not require transformations, so using the transformation possibly introduced bias into the results because the variable became non-normally distributed. However, I do not expect this biased the results found imputing un-logged values of height, weight total cholesterol and HDL cholesterol because I did not include the ratio of cholesterol measurements in the model of interest and Morris *et al.*[56] showed entering height and weight into the imputation model still achieved accurate estimates for BMI. For accurate imputations, Morris *et al.* recommended logging total cholesterol and HDL cholesterol if analysing the ratio of total cholesterol measurements because the denominator (HDL cholesterol) values become very close to zero, which may result in imputing values below zero if un-logged.

Table 6.4: Results from fitting the exponential survival model to predict CVD risk using complete records, data imputed using baseline imputation and using the two-fold fully conditional specification algorithm.

Covariates		Complete records analysis (N=93,369)				Analysis of data imputed using baseline MI (N=2,772,502)				Analysis of data imputed using two-fold fully conditional specification algorithm (N=2,772,502)			
		β	SE	IRR (95% CI)	p-value	β	SE	IRR (95% CI)	p-value	β	SE	IRR (95% CI)	p-value
Age (years)	30 - 39	-2.581	0.1270	0.08 (0.06 - 0.10)	<0.001	-3.566	0.0214	0.03 (0.03 - 0.03)	<0.001	-3.558	0.0219	0.03 (0.03 - 0.03)	<0.001
	40 - 49	-1.415	0.0542	0.24 (0.22 - 0.27)	<0.001	-2.088	0.0131	0.12 (0.12 - 0.13)	<0.001	-2.101	0.0134	0.12 (0.12 - 0.13)	<0.001
	50 - 59	-0.926	0.0378	0.40 (0.37 - 0.43)	<0.001	-1.212	0.0103	0.30 (0.29 - 0.30)	<0.001	-1.242	0.0105	0.29 (0.28 - 0.29)	<0.001
	60 - 69	-0.474	0.0323	0.62 (0.58 - 0.66)	<0.001	-0.547	0.0093	0.58 (0.57 - 0.59)	<0.001	-0.570	0.0095	0.57 (0.56 - 0.58)	<0.001
	70 - 79	0		1		0		1		0		1	
	80 - 89	0.354	0.0458	1.42 (1.30 - 1.56)	<0.001	0.363	0.0112	1.44 (1.41 - 1.47)	<0.001	0.393	0.0115	1.48 (1.45 - 1.51)	<0.001
	90 +	0.781	0.1481	2.18 (1.63 - 2.92)	<0.001	0.578	0.0240	1.78 (1.70 - 1.87)	<0.001	0.604	0.0246	1.83 (1.74 - 1.92)	<0.001
Sex	Male	0		1		0		1		0		1	
	Female	-0.312	0.0278	0.73 (0.69 - 0.77)	<0.001	-0.401	0.0099	0.67 (0.66 - 0.68)	<0.001	-0.402	0.0083	0.67 (0.66 - 0.68)	<0.001
Townsend deprivation score quintile	1 (least)	0		1	<0.001	0		1	<0.001	0		1	<0.001
	2	0.053	0.0374	1.05 (0.98 - 1.14)	0.154	0.058	0.0096	1.06 (1.04 - 1.08)	<0.001	0.055	0.0095	1.06 (1.04 - 1.08)	<0.001
	3	0.069	0.0384	1.07 (0.99 - 1.16)	0.074	0.117	0.0099	1.12 (1.10 - 1.15)	<0.001	0.109	0.0098	1.12 (1.09 - 1.14)	<0.001
	4	0.142	0.0394	1.15 (1.07 - 1.25)	<0.001	0.184	0.0104	1.20 (1.18 - 1.23)	<0.001	0.170	0.0103	1.19 (1.16 - 1.21)	<0.001
	5 (most)	0.252	0.0424	1.29 (1.18 - 1.40)	<0.001	0.276	0.0127	1.32 (1.28 - 1.35)	<0.001	0.260	0.0117	1.30 (1.27 - 1.33)	<0.001
BMI (per 10kg m^{-2})		0.029	0.0026	1.03 (0.98 - 1.08)	0.255	-0.008	0.0013	0.99 (0.96 - 1.02)	0.546	0.080	0.0012	1.08 (1.05 - 1.12)	<0.001
Systolic blood pressure (per 10mmHg)		0.019	0.0006	1.02 (1.01 - 1.03)	0.002	0.068	0.0002	1.07 (1.07 - 1.07)	<0.001	0.066	0.0002	1.07 (1.06 - 1.07)	<0.001
Total cholesterol (mmol l^{-1})		0.064	0.0111	1.07 (1.04 - 1.09)	<0.001	0.062	0.0047	1.06 (1.05 - 1.08)	<0.001	0.058	0.0032	1.06 (1.05 - 1.07)	<0.001
HDL cholesterol (mmol l^{-1})		-0.355	0.0369	0.70 (0.65 - 0.75)	<0.001	-0.284	0.0191	0.75 (0.72 - 0.79)	<0.001	-0.233	0.0150	0.79 (0.77 - 0.82)	<0.001
Smoking status	Non	0		1		0		1		0		1	
	Ex	0.136	0.0298	1.15 (1.08 - 1.21)	<0.001	0.138	0.0177	1.15 (1.10 - 1.20)	<0.001	0.195	0.0091	1.22 (1.19 - 1.24)	<0.001
	Current	0.339	0.0344	1.40 (1.31 - 1.50)	<0.001	0.376	0.0252	1.46 (1.36 - 1.56)	<0.001	0.474	0.0100	1.61 (1.57 - 1.64)	<0.001
Constant		-3.800	0.1363	-	<0.001	-4.414	0.0498	-	<0.001	-4.702	0.0574	-	<0.001

β: coefficient; SE: standard error; IRR: incidence risk ratio; CI: confidence interval; MI: multiple imputation; BMI: body mass index; HDL: high-density lipoprotein

6.1.3 Discussion

From the full data, the complete records analysis found older, male and more deprived patients with higher blood pressure, higher total cholesterol and lower HDL cholesterol and either ex-smokers or current smokers were associated with increased CVD events. Following imputation using the two-fold FCS algorithm, higher BMI was also associated with increased CVD risk. In chapter 5 I showed the two-fold FCS algorithm worked reliably in this setting and, in this section, I successfully used it to impute for the whole of THIN and applied a CVD risk model to the imputed data. I also found that, as in chapter 5, the two-fold FCS algorithm successfully used longitudinal data to recover information on sparsely observed variables.

These results showed time-dependent variables had smaller SEs from analysing data imputed using the two-fold FCS algorithm compared to complete case analysis and analysing data imputed using baseline MI. The estimates themselves following imputation using the two-fold FCS algorithm were either similar or slightly stronger association (more precise) than estimates following baseline MI, which agrees with the findings from the simulation studies in chapter 5.

I compared the imputed coefficients to those from other studies, but only broad comparisons were appropriate because the models were probably adjusted differently. These studies fitted Cox proportional hazards models, rather than exponential survival models, so reported hazard ratios (HRs) instead of IRR. Even though the underlying model assumptions were different, the coefficients are comparable because they both measure association with time to an event.

QRISK2[10] is a widely used CVD risk prediction model developed using the UK primary care database QRESEARCH[25]. QRISK2 found an association between BMI and CVD of HR 1.17 (95% CI 1.21 - 1.12) for women and 1.19 (95% CI 1.23 - 1.14) for men, both for a 10kg m^{-2} increase in BMI. A meta analysis found a similar association between BMI and CVD of HR 1.16 (95% CI 1.07 - 1.26) for a 10kg m^{-2} increase in BMI[84]. The IRR I found from data imputed using two-fold FCS algorithm of 1.08 (95% CI 1.05 - 1.12) was similar to the IRR from analysing data imputed using baseline MI of 0.99 (95% CI 0.96 - 1.02), for a 10kg m^{-2} increase in BMI, but the IRR from data imputed using the two-fold FCS algorithm was statistically significant and closer to the HR found from other studies compared to the IRR from analysing data imputed using baseline MI.

QRISK2 found an association between systolic blood pressure and CVD of HR 1.10 (95% CI 1.10 - 1.09) for women and 1.09 (95% CI 1.10 - 1.08) for men, both for a 10mmHg increase in systolic blood pressure, which were similar to the IRR I found from analysing data imputed using baseline MI 1.07 (95% CI 1.07 - 1.07) and from data imputed using the two-fold FCS algorithm of 1.07 (95% CI 1.06 - 1.07) for a 10mmHg increase in systolic blood pressure. However, another study study reported larger (1.20 to 1.67) HR per 10mmHg increase in systolic blood pressure (across different ages and different cardiovascular diseases[100]) compared to this study and QRISK2.

HRs per mmol l^{-1} varied from 1.18 to 2.27 for the association between CHD risk and total cholesterol (across different age groups[101]) which had a stronger association compared to the IRRs found for the association between total cholesterol and CVD risk from analysing data imputed using baseline MI of 1.06 (95% CI 1.05 - 1.08) and data imputed using the two-fold FCS algorithm of 1.06 (95% CI 1.05 - 1.07). Finally, HRs per mmol l^{-1} were 0.53 (95% CI 0.46 - 0.60) for the association between CHD risk and HDL cholesterol (across different age groups[102]) which had a stronger association compared to the IRRs found for the association between total cholesterol and CVD risk from analysing data imputed using baseline MI of 0.75 (95% CI 0.72 - 0.79) and data imputed using the two-fold FCS algorithm of 0.79 (95% CI 0.77 - 0.82).

Therefore, from investigating other studies, we might expect stronger association between BMI, systolic blood pressure, total cholesterol and HDL cholesterol and CVD compared to those found in this section, but they were close and in the same direction. I could not compare total cholesterol and HDL cholesterol results to QRISK2 because they reported the ratio of total cholesterol to HDL cholesterol. The ultimate test would be to compare the discrimination and calibration for the model of interest to other CVD risk prediction models. However, this is beyond the scope of work for this thesis, but could be considered in future work.

6.1.4 Summary

In this section, I investigated imputing missing THIN data using the two-fold FCS algorithm. However, usually epidemiological studies using electronic health records do not analyse the whole database. Therefore, in the next section, I describe a epidemiological study using THIN which analyses patients with type 2 diabetes in THIN. Again, the analysis includes variables with missing values, so I investigate using different imputation methods including two-fold FCS imputation.

6.2 Total cholesterol reduction after initiating statin treatment for patients diagnosed with type 2 diabetes

The aim of this epidemiological study is to determine the characteristics of patients with type 2 diabetes who had a greater than average total cholesterol reduction after initiating statin treatment. To achieve this, I investigate different imputation methods to impute missing health indicator values. I impute missing values using the two-fold fully conditional specification (FCS) algorithm and compare this method to alternative approaches, complete records analysis and applying multiple imputation to impute missing values 6 months before initiating statin treatment, to determine if the two-fold FCS algorithm is the preferred method for handling missing data in this context. This study received approval from the THIN Scientific Review Committee (Appendix Q).

6.2.1 Introduction and clinical motivation

Patients diagnosed with type 2 diabetes are at high risk of developing cardiovascular disease (CVD)[103]. To reduce CVD risk, the National Institute for Health Care Excellence (NICE) recommends prescribing statins to all patients diagnosed with diabetes to control lipids and CVD risk[103]. However, a recent review suggests initiating statin therapy in patients with diabetes and atherogenic dyslipidaemia did not effectively reduce total cholesterol and subsequent CVD risk[104]. Atherogenic dyslipidaemia is a blood fat disorder which commonly occurs in patients with type II diabetes, causing artery walls to thicken and characterised by low levels of high-density lipoprotein (HDL) cholesterol, high triglycerides and high low-density lipoprotein (LDL) cholesterol, resulting in high total cholesterol.

In this study, I identify the characteristics of patients with type 2 diabetes and greater total cholesterol reduction after initiating statin treatment to discover if statins effectively reduce total cholesterol in patients with type II diabetes and if the associations found support the NICE guidance or the review. Clinicians can use these results as guidance to identify the patients with type II diabetes most likely to respond to statin treatment i.e. the greatest total cholesterol reduction.

I investigated GP records of patients with type 2 diabetes who initiated statin treatment to examine the association between sociodemographic variables and health indicators measured before initiating statin treatment and the greater total cholesterol reduction from before initiating statin treatment to within the first 6 months after first statin treatment. In particular, I explored if patients with type 2 diabetes and lipid measurements which indicate artherogenic dyslipidaemia before initiating statin treatment had smallest total cholesterol reduction. Many studies used CVD events to assess the performance of statins. However, CVD events may not occur for some time, if at all, during the study period. Therefore, as total cholesterol is an independent predictor for CVD, it may act as a surrogate measure for CVD risk if we assume a patients CVD risk reduces if total cholesterol reduces. Another advantage is the study should have more power when analysing a continuous outcome. I was interested in analysing LDL cholesterol levels because (out of the lipid measurements) LDL cholesterol is the strongest risk factor for CVD[104], but an exploratory analysis of recording in the 15 month period before initiating

statin treatment found 35% of patients missing LDL cholesterol measurements and 4% of patients missing total cholesterol measurements. Therefore, I chose to analyse total cholesterol instead of LDL cholesterol, not only because of the improved recording but also because one of the components used to calculate total cholesterol is LDL cholesterol[106], so I expect to find similar results.

The Quality Outcomes Framework (QOF)[107] was introduced in 2004 to encourage GPs to regularly record health indicators for people with chronic diseases like diabetes, discussed in section 3.1.4. Health indicators associated with increased risk of the QOF specified diseases were more frequently recorded following the introduction of QOF[8, 13]. For example, GPs recorded body mass index (BMI), smoking status, glycosylated haemoglobin (HbA_{1c}), blood pressure and total cholesterol measurements for patients diagnosed with type 1 or type 2 diabetes, as shown in section 4.3 (except for HbA_{1c}). Therefore, I restricted the analysis to after the introduction of QOF. Also, NICE identifies the following health indicators associated with CVD risk in patients diagnosed with diabetes: overweight, high blood pressure, high serum albumin, smoking and high-risk lipid profile (low HDL cholesterol and high LDL cholesterol)[103].

I investigated different methods for handling missing data. First, I performed a complete records analysis and then I investigated three different approaches of imputing using multiple imputation (MI), described below, and compared to the results of analysis using each imputed data to each other, and the complete records analysis, to determine if the two-fold FCS algorithm is an appropriate imputation methods in this context.

- Impute missing values in the 6 months before initiating statin treatment conditional only on other measurements recorded in the 6 months before initiating statin treatment;
- Impute missing values in the 6 months before initiating statin treatment using the two-fold FCS algorithm conditional on measurements recorded every 6 month time block before initiating statin treatment; and
- Impute missing values in the 6 months before initiating statin treatment using the two-fold FCS algorithm conditional on measurements recorded every 6 month time block before and after the 6 months before initiating statin treatment.

6.2.2 Methods

6.2.2.1 Study population

For this study, in addition to the criteria to select patients from The Health Improvement Network (THIN) described in section 4.1, I included patients:

1. diagnosed with type 2 diabetes. This included patients with at least one of the following:
    - a type 2 diabetes Read code in the patient record. I assumed patients with non-specific diabetes were type 2 diabetes and included them in the analysis;
    - an additional health data (AHD) code for annual diabetes check, current diabetes status or insulin dosage;

- the data type of diabetic register, diabetic consultation or diabetes concerns; or
- prescribed type 2 diabetes specific medication.

2. with no previous CVD event before initiating statin treatment; and
3. permanently registered with the practice at some time from 1 January 2004 to 31 December 2012.

I included patients aged 30 years or over when they enter the study (study start date), which occurs at the latest date of (i) first type 2 diabetes diagnosis; (ii) ACU date; (iii) AMR date; (iv) when the patient registered with the practice or (v) 1 January 2004. After the start date, the patient initiated statin treatment when first prescribed one of the statin drug codes from the relevant BNF chapter. I included patients with at least two statin prescriptions within the first 6 month after initiating statin treatment because receiving repeated statin prescriptions suggests these patients adhered to statin treatment. I included patient measurements in the analysis up until the earliest date of (i) death; (ii) when the patient transferred out of the practice; (iii) the last date the practice contributes data to THIN or (iv) 31 December 2012. I excluded patients if the time between first statin prescription and last date was less than 6 months.

To exclude prevalent cases, I excluded patients prescribed statins before or within 6 months of the study start date.

6.2.2.2 Study variables

I analysed following explanatory variables measured within 6 months before initiating statin treatment, called 'baseline':

- Sociodemographic: age at first statin prescription, sex, deprivation (quintiles of Townsend deprivation scores) and ethnicity (white, black, South Asian or other).
- Health indicators: body mass index (BMI), systolic blood pressure, total cholesterol, HDL cholesterol, LDL cholesterol, glycosylated haemoglobin (HbA_{1c}), glomerular filtration rate (GFR) and smoking status.
- Statin therapy: statin type and dose. I created categories using simvastatin dose and found equivalent doses for statins other than simvastatin[108].
- Other variables: Calendar year of first statin prescription and time from diagnosis of diabetes to first statin prescription.

Atherogenic dyslipidaemia is characterised by low HDL cholesterol and high LDL cholesterol so it was accounted for in the analysis by adjusting for HDL cholesterol and LDL cholesterol. Also, I did not adjust for baseline total cholesterol because I used baseline total cholesterol to calculate the outcome.

Typically, for most continuous health indicators, a low value indicates good health. However, for HDL cholesterol and GFR, a high value indicates good health. If more than one measurement was recorded at baseline, I included the measurement closest to initiating statin treatment in the analysis. When the data were extracted, I identified and excluded outliers using the method described previously in section 4.4[4].

6.2.2.3 Analysis

To model total cholesterol reduction after initiating statin treatment, I calculated the difference between the total cholesterol measurements recorded at baseline and within the first 6 months after initiating statin treatment. I investigated if this difference was normally distributed and produced a summary of the mean total cholesterol at baseline and the mean total cholesterol reduction, stratified by the baseline characteristics described above. I explored the proportions of missing data at baseline and in each 6 month time block before and after baseline.

I analysed the observed data (Figure 6.2, part 1) using a linear model of interest where the outcome was the difference between total cholesterol measurements recorded at baseline and within the first 6 months after first statin prescription.

Before imputation, I evaluated the plausibility of the MAR assumption to assess if MI was appropriate. In section 3.3.6, I explained that auxiliary variables associated with the health indicator being missing increase the plausibility of the MAR assumption when included in the imputation model, but only influences the imputed values if also associated with the value of the health indicator with missing data. For example, when investigating auxiliary variables to impute missing values for weight, if an auxiliary variable, say asthma, is associated with the weight values I include it in the imputation model to impute missing weight values, regardless of its associated with weight values being missing, increasing efficiency. However, if it is also associated with weight values being missing, imputing missing weight values conditional on asthma increases the MAR assumption plausibility and, hence, reduce bias.

To investigate the association between auxiliary variables and the values of the variables with missing values, I fitted regression models with each variable with missing values as the predictor. I also fitted logistic regression models for each variable with missing data, with the predictor indicating when variable with missing data was missing. These regression analyses conditioned on all variables in the linear model of interest, including the outcome, and all auxiliary variables. I included auxiliary variables in the imputation model if they were associated with the value of the health indicator with missing values with p-value$<$0.100 because the auxiliary variables with a weak association can still influence the imputation as well as those variables with a strong association.

If the distribution of the continuous variables with missing data was skewed, I transformed the data before imputation[47]. Next, I imputed missing values at baseline using fully conditional specification (FCS)[21] MI, described in section 3.3.4 (Figure 6.2, part 2). In addition to the variables measured at baseline in the model of interest (including the outcome variable), I included the following auxiliary variables in the imputation model if associated with the value of the variable with missing data:

1. Patients diagnosed the following before initiating statin treatment, identified using QOF Read code lists: chronic kidney disease (CKD, Appendix M); chronic obstructive pulmonary disease (COPD, Appendix D); schizophrenia, bipolar disorder and other psychoses (psychoses, Appendix K); asthma (Appendix L); and atrial fibrillation (AF, Appendix N). I assumed patients without a Read code did not have the disease.

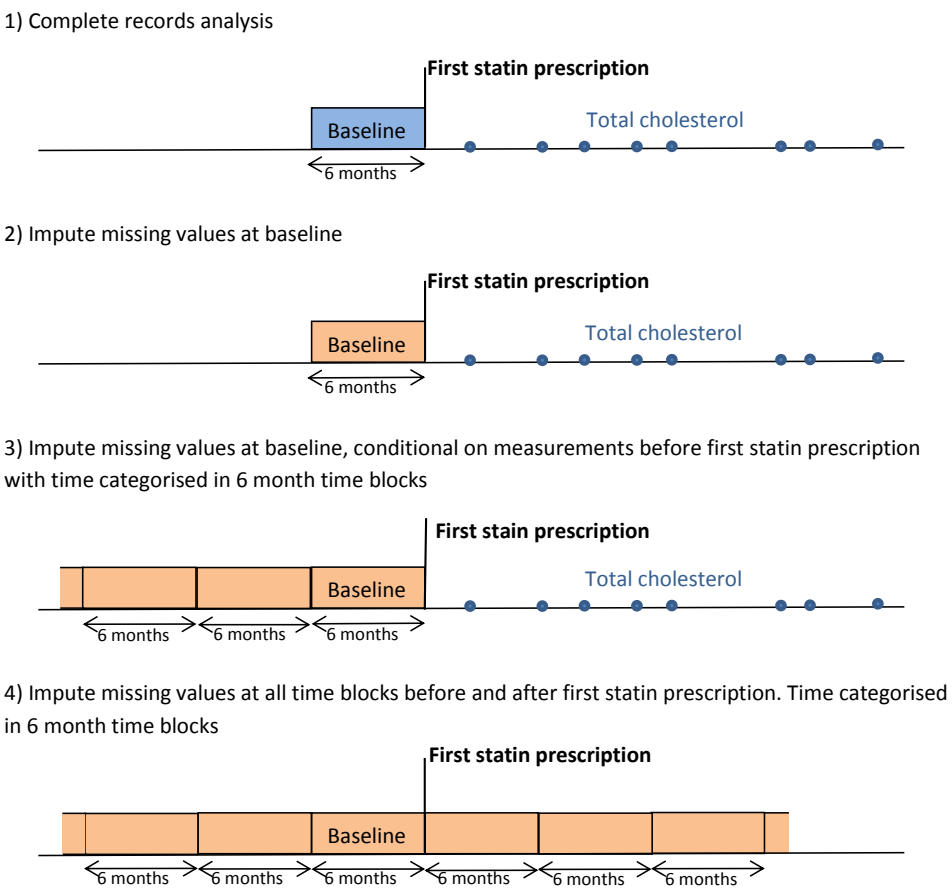

Figure 6.2: Figure to illustrate different imputation methods

2. Patients prescribed following drugs during baseline:

   - Anti-hypertensive drug treatment.
   - Other lipid-modifying drugs bile acid sequestrants; ezetimibe; fibrate and omega 3 polyunsaturated fatty acid ethyl esters. I excluded nicotinic acid from the imputation model for the binary model because patients included in the analysis only had a few prescriptions of nicotinc acid.

3. Respiratory infection during the baseline period (Read code list in Appendix O).

4. Recognised alcohol problem (i.e. current heavy drinker) – the following was recorded at any time in the patient record: male patient consumed more than 5 units of alcohol per day; female patient consumed more than 4 units of alcohol per day or at least one Read code which indicated the patient was a heavy drinker (Appendix P).

5. Death - patients died at the end of follow-up.

6. CVD event - I identified patients with a CVD event at any time after the first statin prescription (separate indicators for CHD and stroke) using Read code lists (Appendix I and Appendix B).

I did not include a variable measuring the time to death or CVD event because the models of interest were not survival models.

I also imputed missing values at baseline again using the two-fold fully conditional specification (FCS)[1] algorithm, but now conditional on heath indicators recorded before baseline (Figure 6.2, part 3). The time before baseline was divided into 6 month time blocks. I excluded the earliest time block if it was less than 6 months and included up to 4 time blocks before baseline. If health indicators were recorded more than once during each time block before baseline, I selected a measurement at random to include in the analysis, except smoking status. If smoking status was recorded more than once, I selected the most frequently occurring category. In addition to the sociodemographic variables, health indicators (measured during each time block) and outcome from the model of interest. I also included the following auxiliary variables (described above): diseases diagnosed before first statin prescription, recognised alcohol problem, death and CVD event. Additionally, I included in the imputation model anti-hypertensive drug treatment, other lipid-modifying drugs and respiratory infection recorded at each time block.

Initially, I was going to condition on smoking status at each time point when using the two-fold FCS algorithm, but there were not enough observed smoking status records at these time blocks and not enough patients with each smoking status category recorded so perfect prediction errors occurred. Therefore, I condition on smoking status at baseline only in the imputation model.

Next, I imputed missing values at baseline this time using the two-fold FCS algorithm, conditional on heath indicators recorded before and after baseline (Figure 6.2, part 4). The time before and after baseline was divided into 6 month time blocks, excluding the earliest or latest time block if it was less than 6 months, and selected measurements during each time block using the method described earlier. I included up to 4 time blocks before and after baseline. I investigated imputing using the two-fold FCS algorithm conditional on health indicators recorded before and after baseline to determine if the additional information recorded after initiating statin treatment included in the imputation model would affect the analysis of the imputed data. Patients who initiate statin treatment may have more information recorded after they initiate statin treatment because they were monitored by the GP, including the same auxiliary variables as before.

Each imputation approach generated 5 imputed datasets. The two-fold FCS algorithm used 5 within-time iterations and 20 among-time iterations. After each imputation, I analysed the imputed data as follows for the linear model of interest and compared to the results from the complete records analysis and the results of analysis using the data imputed with different MI approaches:

- fitted unadjusted models for each variable in the model of interest;
- adjusted models for each variable in the model of interest adjusted for age, sex and ethnicity; and
- the fully adjusted model of interest.

Finally, I used a maximum likelihood, a numerical method which appropriately incorporates missing data by finding the parameter values which make the likelihood function as large as possible. I compared these results with the results of analysis using data imputed with MI to investigate if they agree. I used structural equation modelling with full-information maximum likelihood (FIML) to obtain full dataset estimates[44], as described in section 3.2.3.1. FIML obtains estimates and standard errors by directly maximizing the likelihood for the specified model so I could compare them to the estimates and standard errors obtained from the results of analysing data imputed using MI. Using structural equation modelling allowed me to include the same auxiliary variables as the imputation model.

I found estimates and standard errors for the association between explanatory variables in the model of interest (fitted to measurements recorded at baseline) with the difference between total cholesterol measurements recorded at baseline and within the first 6 months following first statin prescription.

6.2.3 Results

I identified 35,621 patients with type II diabetes; 21,242 (60.0%) had a total cholesterol measurement recorded 6 months before the first statin treatment (baseline) and 6 months after first statin treatment. After I excluded 5 additional patients because the type of statin prescribed was missing, 21,237 patients remained in the analysis. The distribution of the difference between the total cholesterol measurements at baseline and the first 6 months following first statin treatment was normally distributed (Figure 6.3).

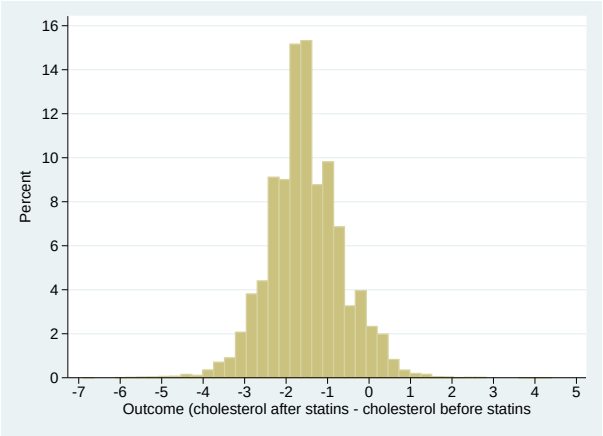

Figure 6.3: Histogram of total cholesterol difference between measures recorded 6 months before first statin treatment and 6 months after first statin treatment

I investigated the percentage of missing values at baseline and each 6 month time blocks before and after baseline for each health indicator with missing values. I found the percentage of missing values varied from 6.5% to 53% at baseline, but more than 85% of health indicators were missing at other time blocks (Table 6.5).

Table 6.5: Percentage missing health indicators at baseline (6 months before initiating statin treatment) and each 6 month time block before (negative values) and after (positive values) baseline

Time block	**Weight** n (%)	**Systolic blood pressure** n (%)	**HDL cholesterol** n (%)	**LDL cholesterol** n (%)	**GFR** n (%)	**HbA$_{1c}$** n (%)	**Smoking status** n (%)
-4	20,783 (97.9)	20,770 (97.8)	20,881 (98.3)	20,921 (98.5)	21,008 (98.9)	20,872 (98.3)	20,969 (98.7)
-3	20,601 (97.0)	20,564 (96.8)	20,817 (98.0)	20,810 (98.0)	20,943 (98.6)	20,732 (97.6)	20,880 (98.3)
-2	20,085 (94.6)	20,216 (95.2)	20,406 (96.1)	20,503 (96.5)	20,694 (97.4)	20,364 (95.9)	20,501 (96.5)
-1	19,668 (92.6)	19,590 (92.2)	20,207 (95.1)	20,325 (95.7)	20,444 (96.3)	20,014 (94.2)	20,249 (95.3)
Baseline	3,969 (18.7)	1,382 (6.5)	4,109 (19.3)	8,158 (38.4)	11,382 (53.6)	1,687 (7.9)	3,077 (14.5)
1	19,630 (92.4)	19,462 (91.6)	18,330 (86.3)	18,660 (87.9)	19,732 (92.9)	19,418 (91.4)	20,290 (95.5)
2	19,585 (92.2)	19,465 (91.7)	19,775 (93.1)	19,940 (93.9)	19,852 (93.5)	19,753 (93.0)	20,285 (95.5)
3	19,812 (93.3)	19,789 (93.2)	19,913 (93.8)	20,081 (94.6)	19,839 (93.4)	19,805 (93.3)	20,279 (95.5)
4	19,917 (93.8)	19,947 (93.9)	20,048 (94.4)	20,141 (94.8)	19,741 (93.0)	19,954 (94.0)	20,372 (95.9)

HDL: high-density lipoprotein; LDL: low-density ; HbA$_{1c}$: glycosylated haemoglobin; GFR:glomerular filtration rate

6.2.3.1 Descriptive analysis of total cholesterol reduction between before initiating statin treatment and in the first six months after first statin treatment

Patients at different ages had similar mean total cholesterol values at baseline but older patients had a greater total cholesterol reduction following first statin treatment (Table 6.6). Female patients started with higher total cholesterol and it reduced more after first statin treatment compared to male patients. I observed similar total cholesterol mean and reduction between the Townsend deprivation score quintiles. Mean total cholesterol was similar for each ethnicity at baseline, but patients of white ethnicity had greater total cholesterol reduction and patients of black ethnicity the smallest (Table 6.6).

Table 6.6: Mean total cholesterol at baseline (6 months before first statin treatment) and mean total cholesterol reduction between baseline and the first 6 months after first statin treatment by each baseline characteristic

Baseline characteristic		**N (%)**	**Mean total cholesterol at baseline (95% CI)**	**Mean total cholesterol reduction (95% CI)**
Total		**21,237 (100.0)**	**5.60 (5.59, 5.61)**	**1.53 (1.54, 1.52)**
Age (years) at first statin prescription	≥30 and <40	1,118 (5.3)	5.79 (5.73, 5.84)	1.43 (1.37, 1.49)
	≥40 and <50	3,579 (16.9)	5.67 (5.64, 5.70)	1.45 (1.42, 1.48)
	≥50 and <60	5,223 (24.6)	5.65 (5.62, 5.67)	1.48 (1.46, 1.51)
	≥60 and <70	5,553 (26.1)	5.54 (5.52, 5.57)	1.56 (1.53, 1.58)
	≥70 and <80	4,178 (19.7)	5.48 (5.46, 5.51)	1.60 (1.57, 1.62)
	≥80	1,586 (7.5)	5.65 (5.60, 5.69)	1.65 (1.61, 1.69)
Sex	Male	11,307 (53.2)	5.44 (5.43, 5.46)	1.50 (1.48, 1.51)
	Female	9,930 (46.8)	5.77 (5.76, 5.79)	1.57 (1.55, 1.59)
Townsend deprivation score quintile	1 (least)	5,044 (23.8)	5.59 (5.56, 5.61)	1.54 (1.52, 1.56)
	2	4,597 (21.6)	5.59 (5.57, 5.62)	1.56 (1.53, 1.58)
	3	4,618 (21.7)	5.62 (5.59, 5.65)	1.53 (1.50, 1.55)
	4	4,192 (19.7)	5.59 (5.57, 5.62)	1.52 (1.50, 1.55)
	5 (most)	2,786 (13.1)	5.60 (5.56, 5.63)	1.47 (1.44, 1.51)
Ethnicity	White	20,340 (95.8)	5.60 (5.59, 5.61)	1.53 (1.52, 1.55)
	Black	244 (1.1)	5.47 (5.36, 5.57)	1.25 (1.13, 1.37)
	South Asian	420 (2.0)	5.49 (5.41, 5.57)	1.43 (1.34, 1.52)
	Other	233 (1.1)	5.62 (5.51, 5.74)	1.54 (1.42, 1.66)
BMI (kg m^{-2})	Underweight	60 (0.3)	5.53 (5.30, 5.75)	1.37 (1.17, 1.57)
	Normal	2,585 (12.2)	5.52 (5.49, 5.56)	1.46 (1.43, 1.50)
	Overweight	6,075 (28.6)	5.56 (5.54, 5.58)	1.54 (1.52, 1.57)
	Obese	8,517 (40.1)	5.58 (5.56, 5.60)	1.53 (1.51, 1.55)
	Missing	4,000 (18.8)	-	-
Systolic blood pressure (mmHg)	≥0 and <120	1,809 (8.5)	5.52 (5.48, 5.57)	1.47 (1.43, 1.52)
	≥120 and <130	3,343 (15.7)	5.53 (5.50, 5.56)	1.46 (1.43, 1.49)
	≥130 and <140	4,959 (23.4)	5.55 (5.53, 5.58)	1.52 (1.49, 1.54)
	≥140 and <150	4,752 (22.4)	5.58 (5.55, 5.60)	1.54 (1.51, 1.56)
	≥150 and <160	2,217 (10.4)	5.63 (5.59, 5.67)	1.58 (1.54, 1.62)
	≥160	2,775 (13.1)	5.74 (5.70, 5.78)	1.60 (1.57, 1.63)
	Missing	1,382 (6.5)	-	-
HDL cholesterol (mmol l^{-1})	≥0.00 and <1.00	2,861 (13.5)	5.27 (5.24, 5.30)	1.45 (1.42, 1.49)
	≥1.00 and <1.25	5,841 (27.5)	5.55 (5.53, 5.57)	1.55 (1.53, 1.57)
	≥1.25 and <1.50	3,837 (18.1)	5.67 (5.64, 5.70)	1.59 (1.56, 1.62)
	≥1.50 and <1.75	2,621 (12.3)	5.76 (5.73, 5.80)	1.55 (1.51, 1.58)
	≥1.75 and <2.00	1,007 (4.7)	5.81 (5.76, 5.87)	1.52 (1.46, 1.57)
	≥2.00	961 (4.5)	5.92 (5.87, 5.98)	1.44 (1.39, 1.49)

Table 6.6: (continued)

Baseline characteristic		**N (%)**	**Mean total cholesterol at baseline (95% CI)**	**Mean total cholesterol reduction (95% CI)**
	Missing	4,109 (19.3)	-	-
LDL cholesterol (mmol l^{-1})	≥0.0 and <2.0	355 (1.7)	4.11 (4.04, 4.18)	0.76 (0.68, 0.83)
	≥2.0 and <2.5	959 (4.5)	4.52 (4.49, 4.56)	0.95 (0.90, 1.00)
	≥2.5 and <3.0	2,364 (11.1)	4.97 (4.95, 4.99)	1.21 (1.18, 1.24)
	≥3.0 and <3.5	3,654 (17.2)	5.35 (5.34, 5.37)	1.42 (1.40, 1.44)
	≥3.5 and <4.0	2,951 (13.9)	5.79 (5.77, 5.81)	1.61 (1.58, 1.64)
	≥4.0 and <4.5	1,576 (7.4)	6.31 (6.29, 6.34)	1.89 (1.85, 1.93)
	≥4.5	1,220 (5.7)	7.17 (7.13, 7.20)	2.28 (2.22, 2.33)
	Missing	8,158 (38.4)	-	-
HbA_{1c} (%)	≥0.0 and <6.0	2,249 (10.6)	5.58 (5.54, 5.62)	1.52 (1.48, 1.55)
	≥6.0 and <6.5	2,940 (13.8)	5.56 (5.53, 5.59)	1.53 (1.50, 1.56)
	≥6.5 and <7.0	3,333 (15.7)	5.53 (5.50, 5.56)	1.51 (1.48, 1.54)
	≥7.0 and <7.5	2,852 (13.4)	5.48 (5.45, 5.51)	1.52 (1.49, 1.55)
	≥7.5 and <8.0	2,214 (10.4)	5.45 (5.41, 5.48)	1.48 (1.45, 1.52)
	≥8.0 and <9.0	2,803 (13.2)	5.51 (5.48, 5.55)	1.48 (1.45, 1.51)
	≥9.0	3,158 (14.9)	5.72 (5.68, 5.75)	1.52 (1.48, 1.55)
	Missing	1,688 (8.0)	-	-
GFR (ml min^{-1})	≥0 and <50.0	648 (3.1)	5.57 (5.49, 5.65)	1.65 (1.58, 1.73)
	≥50.0 and <59.9	953 (4.5)	5.57 (5.51, 5.63)	1.61 (1.55, 1.66)
	60	1,900 (8.9)	5.59 (5.55, 5.63)	1.54 (1.50, 1.58)
	≥60.1 and <69.9	1,206 (5.7)	5.54 (5.49, 5.59)	1.52 (1.47, 1.57)
	≥70.0 and <79.9	1,647 (7.8)	5.57 (5.53, 5.62)	1.52 (1.48, 1.56)
	≥80.0 and <89.9	1,511 (7.1)	5.58 (5.53, 5.62)	1.45 (1.41, 1.50)
	90	1,479 (7.0)	5.57 (5.52, 5.62)	1.44 (1.39, 1.48)
	≥90.1	511 (2.4)	5.52 (5.44, 5.60)	1.37 (1.29, 1.45)
	Missing	11,382 (53.6)	-	-
Smoking status	Never smoked	9,294 (43.8)	5.61 (5.59, 5.63)	1.54 (1.52, 1.56)
	Ex-smoker	5,815 (27.4)	5.53 (5.51, 5.56)	1.54 (1.52, 1.56)
	Current smoker	3,051 (14.4)	5.66 (5.63, 5.70)	1.47 (1.43, 1.50)
	Missing	3,077 (14.5)	-	-
Statin dose (simvastatin or equivalent) (mg)	≤10	2,331 (11.0)	5.49 (5.46, 5.53)	1.24 (1.20, 1.27)
	20	9,662 (45.5)	5.58 (5.56, 5.60)	1.47 (1.46, 1.49)
	40	9,114 (42.9)	5.64 (5.62, 5.66)	1.66 (1.64, 1.68)
	≥80	130 (0.6)	6.02 (5.80, 6.25)	1.94 (1.73, 2.16)
Time from diabetes diagnosis to first statin treatment (years)	≥0 and <1	3,694 (17.4)	5.83 (5.81, 5.86)	1.60 (1.57, 1.63)
	≥1 and <2	4,335 (20.4)	5.71 (5.69, 5.74)	1.58 (1.55, 1.61)
	≥2 and <3	2,360 (11.1)	5.64 (5.60, 5.68)	1.57 (1.53, 1.61)
	≥3 and <4	1,753 (8.3)	5.59 (5.54, 5.63)	1.56 (1.52, 1.61)
	≥4 and <6	2,422 (11.4)	5.53 (5.49, 5.57)	1.51 (1.48, 1.55)
	≥6 and <8	1,649 (7.8)	5.45 (5.41, 5.50)	1.48 (1.44, 1.52)
	≥8 and <10	1,247 (5.9)	5.39 (5.34, 5.44)	1.46 (1.41, 1.50)
	≥10 and <15	1,808 (8.5)	5.40 (5.36, 5.44)	1.43 (1.39, 1.47)
	≥15 and <20	885 (4.2)	5.37 (5.30, 5.43)	1.44 (1.39, 1.50)
	≥20	1,084 (5.1)	5.40 (5.34, 5.45)	1.36 (1.31, 1.41)
Calendar year of first statin prescription	2004	3,026 (14.2)	5.75 (5.72, 5.78)	1.58 (1.55, 1.61)
	2005	4,336 (20.4)	5.66 (5.64, 5.69)	1.57 (1.54, 1.60)
	2006	3,829 (18.0)	5.46 (5.43, 5.49)	1.50 (1.47, 1.52)
	2007	2,428 (11.4)	5.48 (5.44, 5.52)	1.45 (1.41, 1.49)
	2008	2,270 (10.7)	5.49 (5.45, 5.53)	1.46 (1.43, 1.50)
	2009	1,858 (8.7)	5.59 (5.54, 5.63)	1.55 (1.51, 1.59)
	2010	1,384 (6.5)	5.64 (5.59, 5.69)	1.54 (1.50, 1.59)
	2011	1,396 (6.6)	5.74 (5.69, 5.80)	1.54 (1.49, 1.59)
	2012	710 (3.3)	5.73 (5.66, 5.81)	1.59 (1.52, 1.66)

BMI: body mass index; HDL: high-density lipoprotein; LDL: low-density ; HbA_{1c}: glycosylated haemoglobin; GFR: glomerular filtration rate

Patients with higher systolic blood pressure, HDL cholesterol, LDL cholesterol, HbA1c (>9%), statin dose, short time between diabetes diagnosis and statin prescription and current smokers had higher total cholesterol measurements at baseline. Patients with higher BMI, systolic blood pressure, LDL cholesterol, lower GFR, being never or ex-smoker, higher statin dose and short time between diabetes diagnosis and statin prescription had the greatest total cholesterol reduction after initiating statin treatment (Table 6.6).

6.2.3.2 Association with greater total cholesterol reduction between before initiating statin treatment and in the first six months after first statin treatment

In total, 4,496 (19.0%) patients were included in the complete records analysis (Table 6.7). For the model of interest with total cholesterol reduction outcome reported here, the coefficients are reductions in total cholesterol on the total cholesterol scale (mmol l^{-1}) per relevant change in the explanatory variable.

Older patients with higher systolic blood pressure, high LDL cholesterol, low HbA_{1c}, high simvastatin dose, shorter time from diabetes diagnosis to first statin prescription and calender year were associated with greater total cholesterol reduction in the unadjusted analysis. The associations in the adjusted analysis (for age, sex and ethnicity) was similar to the unadjusted, except HbA_{1c}, which was no longer associated with total cholesterol reduction.

These associations attenuated even more in the fully adjusted analysis and systolic blood pressure became non-significant. The fully adjusted analysis showed older patients with higher LDL cholesterol (-0.46, 95% CI -0.49,0.43), lower GFR (0.24, 95% CI 0.10, 0.38) and higher statin dose at baseline were associated with greater total cholesterol reduction. Calendar year was also associated with greater total cholesterol reduction but the coefficients do not suggest a direct relationship (Table 6.7).

Before performing MI, I considered the MAR assumption plausibility for the variables with missing data. From chapter 4, I concluded a plausible MAR assumption for height, weight, systolic blood pressure and smoking status in all patients in THIN. I concluded the MAR assumption was plausible for this cohort because QOF recommends these health indicators were measured regularly for patients with type II diabetes. I did not investigate this assumption for HbA_{1c}, but QOF also recommends HbA_{1c} was measured regularly for patients with type II diabetes, so the MAR assumption was plausible for this variable in this cohort.

The QOF guidance for patients with type 2 diabetes does not include recording HDL and LDL cholesterol or GFR. HDL and LDL cholesterol can be measured the same time as total cholesterol, but they were recorded less frequently in THIN (Table 6.6). GFR is a risk factor for CKD so it is important to measure this for patients with type II diabetes, with high CKD risk. However, this cohort had many missing GFR values at baseline (Table 6.6). Therefore, I increased the MAR assumption plausibility for these health indicators by including auxiliary variables listed in the methods section in the imputation model.

After investigating which auxiliary variables were associated with missing health indicator values or the values themselves at baseline, the variables from the model of interest age, Townsend deprivation score quintile, ethnicity, time from first diabetes diagnosis and calendar year were associated with both the value of HDL cholesterol and HDL cholesterol being missing (Table 6.9). I observed the same associations for GFR (Table 6.11). All variables from the model of interest were associated with both the value of LDL cholesterol and LDL cholesterol being missing (Table 6.10).

The auxiliary variables anti-hypertensive drug treatment and omega 3 were associated with both the value of HDL cholesterol and HDL cholesterol being missing at baseline (Table 6.9), so conditioning on these variables in the imputation model increased the MAR assumption plausibility for HDL cholesterol. The auxiliary variables CKD and anti-hypertensive drug treatment were associated with both the value of LDL cholesterol and LDL cholesterol being missing at baseline (Table 6.10), and the same for GFR (Table 6.11), so all imputation models condition on these auxiliary variables to increase the MAR assumption plausibility for these variables with missing data.

Table 6.7: Complete records regression analysis to investigate the association between the baseline variables and the difference between total cholesterol measurements records before initiating statin treatment and in the first six months after first statin treatment for patients with type II diabetes (N=4,496)

		Unadjusted analysis		**Adjusted analysis***		**Fully adjusted analysis**	
		Coefficient (95% CI)	**p-value**	**Coefficient (95% CI)**	**p-value**	**Coefficient (95% CI)**	**p-value**
Age (years) at first statin prescription	≥30 and <40	0.02 (-0.10,0.15)	0.011	0.02 (-0.10,0.14)	0.002	0.05 (-0.06,0.16)	<0.001
	≥40 and <50	0.07 (-0.01,0.14)		0.07 (-0.01,0.14)		0.11 (0.04,0.19)	
	≥50 and <60	0.02 (-0.04,0.09)		0.03 (-0.04,0.09)		0.06 (0.00,0.13)	
	≥60 and <70	Ref		Ref		Ref	
	≥70 and <80	-0.07 (-0.14,0.01)		-0.06 (-0.14,0.01)		-0.08 (-0.15,-0.02)	
	≥80	-0.14 (-0.24,-0.04)		-0.14 (-0.24,-0.03)		-0.11 (-0.21,-0.02)	
Female		-0.04 (-0.08,0.01)	0.161	-0.03 (-0.08,0.02)	0.186	0.04 (-0.01,0.09)	0.119
Townsend score quintile	1 (least)	Ref	0.252	Ref	0.301	Ref	0.456
	2	-0.04 (-0.12,0.03)		-0.05 (-0.12,0.02)		-0.03 (-0.10,0.03)	
	3	0.04 (-0.03,0.11)		0.04 (-0.04,0.11)		0.03 (-0.03,0.10)	
	4	0.02 (-0.06,0.09)		0.01 (-0.07,0.08)		0.00 (-0.07,0.06)	
	5 (most)	0.01 (-0.07,0.10)		0.00 (-0.09,0.09)		0.00 (-0.08,0.08)	
Ethnic group	White	Ref	0.263	Ref	0.320	Ref	0.188
	Black	0.15 (-0.06,0.35)		0.13 (-0.07,0.34)		0.17 (-0.02,0.36)	
	South Asian	0.09 (-0.07,0.25)		0.06 (-0.10,0.23)		-0.01 (-0.15,0.14)	
	Other	-0.10 (-0.32,0.11)		-0.12 (-0.34,0.10)		-0.12 (-0.32,0.08)	
BMI ($100kgm^{-2}$)		0.30 (-0.11,0.71)	0.156	0.15 (-0.29,0.59)	0.509	0.01 (-0.41,0.43)	0.952
Systolic blood pressure (100mmHg)		-0.23 (-0.37,-0.08)	0.002	-0.19 (-0.34,-0.04)	0.012	-0.04 (-0.17,0.10)	0.569
HDL cholesterol ($mmolL^{-1}$)		0.00 (-0.07,0.06)	0.888	0.03 (-0.05,0.10)	0.480	-0.03 (-0.10,0.03)	0.345
LDL cholesterol ($mmolL^{-1}$)		-0.46 (-0.49,-0.43)	<0.001	-0.47 (-0.50,-0.44)	<0.001	-0.46 (-0.49,-0.43)	<0.001
Glycosylated haemoglobin (100%)		2.14 (0.50,3.78)	0.010	1.50 (-0.18,3.19)	0.081	0.94 (-0.64,2.51)	0.244
Glomerular filtration rate ($100mLmin^{-1}$)		0.39 (0.25,0.54)	<0.001	0.31 (0.16,0.47)	<0.001	0.24 (0.10,0.38)	0.001
Smoking status	Non-smoker	Ref	0.188	Ref	0.467	Ref	0.194
	Ex-smoker	0.00 (-0.06,0.05)		0.00 (-0.05,0.06)		-0.02 (-0.07,0.03)	
	Current smoker	0.06 (-0.01,0.13)		0.04 (-0.03,0.11)		0.05 (-0.02,0.11)	

Table 6.7: (continued)

		Unadjusted analysis		Adjusted analysis*		Fully adjusted analysis	
		Coefficient (95% CI)	**p-value**	**Coefficient (95% CI)**	**p-value**	**Coefficient (95% CI)**	**p-value**
Simvastatin dose or equivalent (mg)	≤10	0.42 (0.34,0.51)	<0.001	0.45 (0.36,0.53)	<0.001	0.40 (0.32,0.48)	<0.001
	20	0.14 (0.09,0.19)		0.15 (0.10,0.20)		0.15 (0.10,0.20)	
	40	Ref		Ref		Ref	
	≥80	0.23 (-0.24,0.70)		0.21 (-0.26,0.69)		0.16 (-0.27,0.59)	
Time from diabetes diagnosis to first statin treatment (years)	≥0 and <1	Ref	<0.001	Ref	<0.001	Ref	0.329
	≥1 and <2	0.04 (-0.04,0.11)		0.04 (-0.03,0.11)		-0.04 (-0.11,0.03)	
	≥2 and <3	0.06 (-0.03,0.15)		0.07 (-0.02,0.16)		-0.03 (-0.11,0.06)	
	≥3 and <4	0.05 (-0.06,0.16)		0.06 (-0.05,0.16)		-0.08 (-0.18,0.02)	
	≥4 and <6	0.15 (0.05,0.24)		0.15 (0.06,0.24)		-0.05 (-0.13,0.04)	
	≥6 and <8	0.13 (0.02,0.23)		0.13 (0.03,0.24)		-0.11 (-0.20,-0.01)	
	≥8 and <10	0.15 (0.03,0.27)		0.16 (0.04,0.28)		-0.07 (-0.18,0.04)	
	≥10 and <15	0.28 (0.18,0.39)		0.30 (0.19,0.40)		0.02 (-0.08,0.11)	
	≥15 and <20	0.21 (0.06,0.36)		0.23 (0.08,0.37)		0.04 (-0.10,0.17)	
	≥20	0.28 (0.15,0.41)		0.28 (0.15,0.41)		0.01 (-0.11,0.13)	
Calender year of first statin prescription	2004	0.13 (-0.11,0.37)	<0.001	0.13 (-0.11,0.38)	<0.001	0.14 (-0.08,0.36)	0.005
	2005	-0.14 (-0.28,-0.01)		-0.13 (-0.26,0.00)		-0.14 (-0.27,-0.02)	
	2006	0.04 (-0.05,0.14)		0.05 (-0.05,0.15)		-0.03 (-0.12,0.06)	
	2007	0.12 (0.03,0.21)		0.12 (0.02,0.21)		0.05 (-0.03,0.14)	
	2008	0.13 (0.04,0.22)		0.14 (0.04,0.23)		0.07 (-0.02,0.15)	
	2009	0.01 (-0.08,0.11)		0.01 (-0.08,0.10)		-0.01 (-0.10,0.07)	
	2010	Ref		Ref		Ref	
	2011	0.03 (-0.07,0.13)		0.02 (-0.08,0.13)		0.05 (-0.04,0.15)	
	2012	-0.07 (-0.20,0.06)		-0.07 (-0.20,0.06)		-0.02 (-0.13,0.10)	

*Regression analysis for age, sex and ethnicity adjusted for each other
CI: confidence interval; BMI: body mass index; HDL: high-density lipoprotein; LDL: low-density lipoprotein; HbA_{1c}: glycosylated haemoglobin; GFR: glomerular filtration rate; Ref: Reference category

All of the auxiliary variables were associated with the value of at least one of the variables with missing data, so I included all of them in the imputation model.

I observed some differences when comparing the results from analysing data imputed using baseline MI (to impute missing values at baseline) compared to the complete records analysis. Female patients were not associated with greater total cholesterol reduction in the complete records analysis, but from analysing data imputed using baseline MI, they were associated with greater total cholesterol reduction in the unadjusted analysis and adjusted analysis (for age, sex and ethnicity). In the fully adjusted analysis, male patients were associated with greater total cholesterol reduction(Table 6.13). As well as female patients, I found other completely observed variables were also significant in the analysis of data imputed using baseline MI but not complete records analysis; Townsend deprivation score quintile (not fully adjusted analysis), ethnicity and time from diabetes diagnosis to first statin treatment (Table 6.13).

I observed similar findings from analysing data imputed using baseline MI and complete records analysis for imputed variables, except for HDL cholesterol and HbA_{1c}. HDL cholesterol was not associated with greater total cholesterol reduction in the complete case analysis, but it was for the adjusted baseline MI (possibly because this analysis is also adjusted for LDL cholesterol) (Table 6.13). In the complete records analysis, lower HbA_{1c} was associated with greater total cholesterol reduction, opposite to the expected direction of association, in the unadjusted analysis, but not associated the adjusted or fully adjusted analysis. However, from analysing data imputed using baseline MI, HbA_{1c} was not significantly associated with greater total cholesterol reduction, but greater HbA_{1c} was associated with greater total cholesterol reduction in the adjusted and fully adjusted analysis (Table 6.13). The direction of association changed to the expected direction in the fully adjusted analysis of data imputed using baseline MI.

The fully adjusted analysis showed older, male patients of white ethnicity with higher LDL cholesterol, higher HbA_{1c}, lower GFR, non or ex-smokers, higher statin dose, shorter time from first diabetes diagnosis and earlier calendar years were associated with greater total cholesterol reduction. These findings agreed more closely with the descriptive analysis findings compared to the complete records analysis, except for sex. The descriptive analysis suggested female patients had greater total cholesterol reduction.

The analysis using FIML produced similar results as the fully adjusted analysis of data imputed using baseline MI (Table 6.13) and from analysing data imputed using the two-fold FCS algorithm, conditional on measurements recorded before baseline imputation (Table 6.14). In the fully adjusted analysis of data imputed using baseline MI, Townsend deprivation score quintile, BMI and systolic blood pressure were not significantly associated with greater total cholesterol reduction. However,there was some association (p-values between 0.060 and 0.064) from the fully adjusted analysis of data imputed using the two-fold FCS algorithm conditional on measurements recorded before and after baseline (Table 6.15).

Table 6.8: Association between the outcome variable (difference between total cholesterol 6 months before initiating statin treatment and 6 months after first statin treatment), covariates and auxiliary variables and the chance of observing height and weight and their respective values at baseline

Variable		Height (cm)		Weight (kg)	
		Missing OR (95% CI)	Value Coefficient (95% CI)	Missing OR (95% CI)	Value Coefficient (95% CI)
Outcome		0.95 (0.72,1.25)	**0.18 (0.07,0.29)**	**0.96 (0.93,1.00)**	0.14 (-0.15,0.43)
Age (years) at first statin prescription	≥30 and <40	1.93 (0.72,5.20)	**2.39 (1.93,2.84)**	1.09 (0.92,1.29)	**8.64 (7.42,9.86)**
	≥40 and <50	0.51 (0.18,1.44)	**1.69 (1.39,1.99)**	1.07 (0.96,1.20)	**7.73 (6.92,8.53)**
	≥50 and <60	1.22 (0.60,2.51)	**0.84 (0.58,1.1)**	0.99 (0.90,1.10)	**5.03 (4.32,5.74)**
	≥60 and <70	Ref	**Ref**	Ref	**Ref**
	≥70 and <80	1.28 (0.59,2.74)	**-1.23 (-1.51,0.95)**	1.03 (0.92,1.14)	**-6.86 (-7.62,-6.10)**
	≥80	2.24 (0.95,5.27)	**-3.07 (-3.47,2.66)**	1.19 (1.03,1.38)	**-14.59 (-15.69,-13.48)**
Sex		1.36 (0.83,2.23)	**-14.18 (-14.37,13.99)**	1.06 (0.99,1.14)	**-9.46 (-9.98,-8.94)**
Townsend deprivation score quintile	1	Ref	**Ref**	Ref	**Ref**
	2	1.11 (0.55,2.24)	**-0.46 (-0.73,0.18)**	0.93 (0.84,1.03)	**0.16 (-0.58,0.90)**
	3	1.03 (0.51,2.11)	**-1.26 (-1.52,0.98)**	0.98 (0.89,1.09)	**0.72 (-0.02,1.46)**
	4	1.18 (0.59,2.40)	**-1.52 (-1.8,1.24)**	1.02 (0.92,1.13)	**1.00 (0.23,1.76)**
	5	0.55 (0.20,1.52)	**-2.20 (-2.51,1.87)**	1.05 (0.94,1.19)	**1.18 (0.31,2.05)**
Ethnic group	White	Ref	**Ref**	**Ref**	**Ref**
	Black	1.54 (0.21,11.56)	**0.03 (-0.84,0.9)**	**1.32 (0.97,1.78)**	**-5.50 (-7.91,-3.08)**
	South Asian	-	**-4.99 (-5.66,4.33)**	**1.08 (0.84,1.38)**	**-13.81 (-15.62,-11.99)**
	Other	-	**-4.29 (-5.17,3.4)**	**1.57 (1.17,2.11)**	**-11.31 (-13.84,-8.79)**
CKD		0.78 (0.39,1.58)	0.12 (-0.14,0.38)	0.98 (0.89,1.08)	**1.67 (0.97,2.38)**
Statin dose (simvastatin or equivalent) (mg)	≤10	0.87 (0.35,2.17)	-0.20 (-0.52,0.13)	0.92 (0.81,1.04)	**-1.71 (-2.59,-0.84)**
	20	1.04 (0.61,1.79)	-0.13 (-0.34,0.08)	0.98 (0.91,1.06)	**-0.57 (-1.13,-0.01)**
	40	Ref	Ref	Ref	**Ref**
	≥80	-	-0.40 (-1.59,0.79)	1.35 (0.90,2.03)	**1.19 (-2.15,4.52)**
Time from diabetes diagnosis to first statin treatment (years)	≥0 and <1	Ref	**Ref**	Ref	**Ref**
	≥1 and <2	0.87 (0.42,1.82)	**-0.07 (-0.37,0.23)**	1.10 (0.98,1.24)	**0.07 (-0.74,0.88)**
	≥2 and <3	0.65 (0.25,1.72)	**0.21 (-0.14,0.56)**	1.17 (1.02,1.33)	**0.70 (-0.26,1.66)**
	≥3 and <4	0.60 (0.20,1.86)	**0.09 (-0.3,0.48)**	1.11 (0.96,1.29)	**-0.04 (-1.09,1.02)**
	≥4 and <6	0.73 (0.28,1.92)	**0.39 (0.03,0.74)**	1.15 (1.01,1.32)	**-0.29 (-1.24,0.67)**
	≥6 and <8	1.13 (0.43,2.99)	**0.52 (0.12,0.93)**	1.15 (0.99,1.34)	**0.47 (-0.61,1.56)**
	≥8 and <10	0.91 (0.29,2.82)	**0.26 (-0.18,0.71)**	1.20 (1.02,1.42)	**-1.95 (-3.15,-0.75)**
	≥10 and <15	-	**0.28 (-0.1,0.67)**	1.10 (0.95,1.28)	**-2.08 (-3.13,-1.02)**
	≥15 and <20	2.12 (0.82,5.46)	**0.07 (-0.44,0.58)**	1.14 (0.94,1.38)	**-5.19 (-6.56,-3.82)**
	≥20	1.56 (0.58,4.19)	**-0.23 (-0.7,0.24)**	1.05 (0.88,1.26)	**-10.8 (-12.06,-9.54)**
Calender year of first statin prescription	2004	**0.44 (0.16,1.20)**	**-0.17 (-0.61,0.28)**	**0.93 (0.79,1.10)**	**-1.84 (-3.05,-0.62)**
	2005	**0.19 (0.06,0.58)**	**-0.12 (-0.54,0.3)**	**0.93 (0.79,1.09)**	**-1.70 (-2.85,-0.55)**
	2006	**0.40 (0.15,1.06)**	**0.15 (-0.27,0.58)**	**0.87 (0.74,1.02)**	**-1.53 (-2.68,-0.37)**
	2007	**0.48 (0.18,1.31)**	**0.24 (-0.21,0.69)**	**0.88 (0.74,1.04)**	**-0.83 (-2.06,0.39)**
	2008	**0.36 (0.12,1.11)**	**0.13 (-0.32,0.59)**	**1.01 (0.86,1.20)**	**-0.65 (-1.90,0.59)**
	2009	**0.27 (0.07,1.01)**	**0.31 (-0.17,0.78)**	**1.03 (0.86,1.23)**	**0.45 (-0.85,1.74)**
	2010	**Ref**	**Ref**	**Ref**	**Ref**
	2011	**1.58 (0.64,3.91)**	**0.23 (-0.27,0.74)**	**1.17 (0.98,1.41)**	**1.37 (-0.02,2.76)**
	2012	**1.78 (0.64,4.99)**	**0.91 (0.29,1.53)**	**1.26 (1.01,1.58)**	**2.01 (0.30,3.71)**
COPD		0.83 (0.25,2.80)	0.04 (-0.41,0.49)	0.87 (0.74,1.04)	0.87 (-0.34,2.08)
Psychosis		1.33 (0.17,10.16)	**1.28 (0.3,2.26)**	0.91 (0.63,1.32)	1.73 (-0.91,4.36)
Asthma		0.39 (0.14,1.09)	**-0.33 (-0.6,0.05)**	1.04 (0.93,1.15)	**1.92 (1.16,2.67)**
Atrial fibrilation		1.22 (0.46,3.22)	**1.89 (1.49,2.3)**	1.03 (0.89,1.20)	**3.94 (2.84,5.04)**
Anti-hypertensive drug treatment		0.62 (0.37,1.04)	-0.06 (-0.25,0.14)	1.06 (0.99,1.15)	**5.08 (4.54,5.62)**
Bile acid		-	-1.35 (-5.22,2.52)	**3.37 (1.06,10.67)**	-8.35 (-20.71,4.02)
Ezetimibe		7.99 (1.75,36.43)	-0.34 (-1.97,1.29)	1.23 (0.70,2.15)	**-4.85 (-9.32,-0.39)**
Fibrate		-	-0.01 (-0.76,0.75)	**1.34 (1.03,1.74)**	0.24 (-1.87,2.35)
Nicotinic acid		-	0.50 (-4.58,5.6)	-	2.71 (-9.72,15.14)
Omega 3		-	0.86 (-1.53,3.25)	1.58 (0.72,3.48)	2.09 (-4.78,8.96)
Respiratory infection		1.56 (0.70,3.47)	-0.21 (-0.55,0.15)	1.07 (0.94,1.21)	0.76 (-0.20,1.72)
Heavy drinker		0.46 (0.16,1.30)	0.08 (-0.19,0.35)	1.02 (0.92,1.14)	-0.42 (-1.16,0.32)
CHD		0.33 (0.05,2.43)	-0.20 (-0.63,0.23)	1.01 (0.86,1.19)	-0.29 (-1.47,0.88)
Stroke		0.40 (0.05,2.95)	-0.22 (-0.76,0.32)	**1.30 (1.08,1.57)**	-1.22 (-2.71,0.28)
Death		**4.19 (2.06,8.52)**	**-0.32 (-0.69,0.06)**	**1.27 (1.11,1.45)**	0.01 (-1.03,1.04)

OR: odds ratio; CI: confidence interval; CKD: chronic kidney disease; COPD: chronic obstructive pulmonary disease; Psychoses: schizophrenia, bipolar disorder and other psychoses; CHD: coronary heart disease; Ref: reference category
In bold if variable is associated (p-value<0.100) with the chance of observing height and weight and their respective values at baseline

Table 6.9: Association between the outcome (difference between total cholesterol 6 months before initiating statin treatment and 6 months after first statin treatment), covariates and auxiliary variables and the chance of observing systolic blood pressure and HDL cholesterol and their respective values at baseline

Variable		Systolic blood pressure (mmHg)		HDL cholesterol (mmol l^{-1})	
		Missing	Value	Missing	Value
		OR (95% CI)	Coefficient (95% CI)	OR (95% CI)	Coefficient (95% CI)
Outcome		**0.95 (0.89,1.01)**	**-0.72 (-0.99,-0.44)**	**1.04 (1.00,1.08)**	0.00 (-0.00,0.01)
Age (years) at first statin prescription	≥30 and <40	1.30 (1.03,1.65)	**-7.23 (-8.39,-6.07)**	**1.24 (1.05,1.47)**	-0.10 (-0.13,-0.08)
	≥40 and <50	1.11 (0.93,1.32)	**-4.22 (-4.98,-3.46)**	**1.15 (1.03,1.29)**	-0.09 (-0.11,-0.08)
	≥50 and <60	1.10 (0.94,1.29)	**-1.42 (-2.09,-0.75)**	**1.10 (0.99,1.21)**	-0.04 (-0.05,-0.02)
	≥60 and <70	Ref	**Ref**	**Ref**	Ref
	≥70 and <80	0.94 (0.78,1.12)	**0.51 (-0.20,1.22)**	**1.03 (0.93,1.14)**	0.05 (0.03,0.06)
	≥80	0.96 (0.74,1.24)	**2.32 (1.31,3.34)**	**1.07 (0.92,1.24)**	0.14 (0.12,0.16)
Sex		**1.19 (1.06,1.34)**	-0.22 (-0.70,0.27)	0.96 (0.89,1.03)	**0.21 (0.20,0.22)**
Townsend deprivation score quintile	1	Ref	Ref	Ref	Ref
	2	1.08 (0.91,1.28)	0.11 (-0.58,0.81)	0.99 (0.89,1.09)	0.01 (-0.00,0.03)
	3	1.09 (0.93,1.29)	0.00 (-0.69,0.70)	0.97 (0.88,1.08)	-0.02 (-0.03,-0.00)
	4	1.09 (0.92,1.29)	-0.64 (-1.35,0.08)	1.05 (0.94,1.16)	-0.04 (-0.06,-0.02)
	5	1.18 (0.98,1.43)	-0.22 (-1.03,0.60)	1.11 (0.99,1.25)	-0.06 (-0.08,-0.04)
Ethnic group	White	Ref	**Ref**	**Ref**	Ref
	Black	1.21 (0.75,1.96)	**2.92 (0.69,5.15)**	**0.72 (0.50,1.03)**	0.10 (0.05,0.14)
	South Asian	0.87 (0.58,1.30)	**-3.60 (-5.30,-1.91)**	**0.52 (0.38,0.71)**	-0.05 (-0.08,-0.01)
	Other	1.54 (1.01,2.35)	**-1.09 (-3.41,1.22)**	**0.52 (0.34,0.80)**	-0.01 (-0.06,0.04)
CKD		0.99 (0.84,1.18)	**1.03 (0.37,1.69)**	1.04 (0.95,1.15)	**-0.05 (-0.07,-0.04)**
Statin dose (simvastatin or equivalent) (mg)	≤10	0.93 (0.76,1.14)	-0.19 (-1.02,0.63)	**0.81 (0.72,0.91)**	0.01 (-0.01,0.03)
	20	1.09 (0.96,1.24)	-0.50 (-1.03,0.03)	**0.89 (0.83,0.97)**	0.01 (-0.01,0.02)
	40	Ref	Ref	**Ref**	Ref
	≥80	1.50 (0.81,2.76)	-0.83 (-3.90,2.24)	**0.96 (0.62,1.47)**	-0.02 (-0.09,0.04)
Time from diabetes diagnosis to first statin treatment (years)	≥0 and <1	Ref	Ref	**Ref**	Ref
	≥1 and <2	1.01 (0.85,1.21)	-0.15 (-0.92,0.61)	**1.08 (0.96,1.21)**	0.02 (0.01,0.04)
	≥2 and <3	1.08 (0.87,1.33)	-0.43 (-1.34,0.47)	**1.04 (0.90,1.19)**	0.02 (-0.00,0.04)
	≥3 and <4	1.05 (0.82,1.33)	-0.77 (-1.77,0.22)	**1.09 (0.94,1.26)**	0.02 (0.00,0.05)
	≥4 and <6	1.20 (0.98,1.48)	-0.13 (-1.03,0.77)	**1.11 (0.97,1.27)**	0.01 (-0.01,0.03)
	≥6 and <8	1.15 (0.90,1.46)	-0.29 (-1.31,0.73)	**1.20 (1.03,1.40)**	0.04 (0.01,0.06)
	≥8 and <10	1.19 (0.92,1.55)	-0.08 (-1.21,1.04)	**1.32 (1.13,1.56)**	0.05 (0.02,0.07)
	≥10 and <15	1.08 (0.85,1.38)	-0.22 (-1.21,0.77)	**1.28 (1.11,1.48)**	0.09 (0.07,0.11)
	≥15 and <20	1.36 (1.01,1.81)	-0.22 (-1.51,1.08)	**1.67 (1.40,2.00)**	0.15 (0.12,0.18)
	≥20	1.25 (0.96,1.64)	-1.56 (-2.76,-0.37)	**1.67 (1.42,1.97)**	0.29 (0.27,0.32)
Calender year of first statin prescription	2004	**0.84 (0.64,1.09)**	**4.11 (2.97,5.25)**	**2.28 (1.89,2.73)**	0.08 (0.05,0.10)
	2005	**0.81 (0.63,1.04)**	**3.12 (2.04,4.21)**	**2.10 (1.76,2.51)**	0.08 (0.06,0.10)
	2006	**0.88 (0.69,1.13)**	**1.50 (0.41,2.59)**	**1.64 (1.37,1.96)**	0.06 (0.04,0.09)
	2007	**0.75 (0.57,0.99)**	**0.44 (-0.71,1.60)**	**1.68 (1.39,2.03)**	0.02 (-0.00,0.05)
	2008	**1.03 (0.79,1.34)**	**0.95 (-0.23,2.12)**	**1.36 (1.12,1.65)**	0.02 (-0.01,0.04)
	2009	**1.11 (0.85,1.45)**	**0.70 (-0.51,1.92)**	**1.25 (1.02,1.53)**	0.02 (-0.01,0.05)
	2010	**Ref**	**Ref**	**Ref**	Ref
	2011	**1.32 (1.00,1.74)**	**0.65 (-0.66,1.96)**	**0.92 (0.73,1.16)**	0.02 (-0.00,0.05)
	2012	**1.34 (0.96,1.85)**	**0.62 (-0.98,2.21)**	**0.95 (0.72,1.26)**	0.02 (-0.02,0.05)
COPD		1.03 (0.79,1.36)	-0.85 (-1.99,0.29)	**1.18 (1.01,1.38)**	0.02 (-0.01,0.04)
Psychosis		1.01 (0.61,1.68)	**-3.35 (-5.87,-0.83)**	1.22 (0.85,1.74)	**-0.10 (-0.16,-0.04)**
Asthma		1.14 (0.97,1.33)	-0.57 (-1.28,0.14)	1.02 (0.92,1.13)	**0.03 (0.01,0.04)**
Atrial fibrilation		0.88 (0.66,1.17)	**-3.52 (-4.53,-2.50)**	1.03 (0.89,1.19)	**-0.03 (-0.05,-0.01)**
Anti-hypertensive drug treatment		**0.42 (0.37,0.47)**	**8.41 (7.90,8.92)**	**1.11 (1.03,1.20)**	**-0.05 (-0.06,-0.04)**
Bile acid		3.00 (0.63,14.28)	-0.75 (-11.19,9.69)	0.39 (0.05,3.07)	**0.23 (0.03,0.44)**
Ezetimibe		0.84 (0.30,2.34)	**-4.40 (-8.48,-0.32)**	0.88 (0.46,1.68)	-0.02 (-0.11,0.06)
Fibrate		1.00 (0.63,1.60)	-1.31 (-3.24,0.61)	0.99 (0.75,1.31)	**-0.15 (-0.19,-0.10)**
Nicotinic acid		1.00 (1.00,1.00)	-5.33 (-17.87,7.21)	0.71 (0.08,6.06)	**-0.39 (-0.67,-0.12)**
Omega 3		**3.12 (1.16,8.40)**	0.37 (-6.02,6.76)	**0.14 (0.02,1.04)**	**-0.11 (-0.23,0.01)**
Respiratory infection		**0.75 (0.60,0.94)**	**-0.83 (-1.72,0.06)**	1.06 (0.93,1.20)	**-0.03 (-0.05,-0.01)**
Heavy drinker		1.06 (0.89,1.24)	**1.40 (0.70,2.10)**	0.99 (0.89,1.10)	**0.13 (0.11,0.14)**
CHD		1.03 (0.79,1.36)	**1.40 (0.30,2.49)**	1.05 (0.90,1.23)	**-0.04 (-0.07,-0.02)**
Stroke		1.10 (0.79,1.53)	**1.69 (0.32,3.06)**	1.00 (0.83,1.22)	0.00 (-0.03,0.03)
Death		**1.28 (1.02,1.60)**	-0.07 (-1.02,0.88)	1.10 (0.96,1.25)	-0.02 (-0.04,0.01)

OR: odds ratio; CI: confidence interval; HDL: high-density lipoprotein; CKD: chronic kidney disease; COPD: chronic obstructive pulmonary disease; Psychoses: schizophrenia, bipolar disorder and other psychoses; CHD: coronary heart disease; Ref: reference category
In bold if variable is associated (p-value<0.100) with the chance of observing systolic blood pressure and HDL cholesterol and their respective values at baseline

Table 6.10: Association between the outcome (difference between total cholesterol 6 months before initiating statin treatment and 6 months after first statin treatment), covariates and auxiliary variables and the chance of observing LDL cholesterol and HbA_{1c} and their respective values at baseline

Variable		**LDL cholesterol (mmol l^{-1})**		**HbA_{1c} (%)**	
		Missing	**Value**	**Missing**	**Value**
		OR (95% CI)	**Coefficient (95% CI)**	**OR (95% CI)**	**Coefficient (95% CI)**
Outcome		**0.95 (0.92,0.98)**	**-0.39 (-0.41,-0.38)**	**0.75 (0.70,0.79)**	**-0.05 (-0.07,-0.03)**
Age (years) at first statin prescription	≥30 and <40	**1.47 (1.28,1.68)**	**0.13 (0.07,0.19)**	**0.96 (0.75,1.22)**	**0.81 (0.71,0.91)**
	≥40 and <50	**1.36 (1.24,1.49)**	**0.09 (0.06,0.13)**	**0.73 (0.62,0.86)**	**0.56 (0.50,0.63)**
	≥50 and <60	**1.14 (1.06,1.24)**	**0.06 (0.03,0.10)**	**0.86 (0.75,0.99)**	**0.35 (0.29,0.41)**
	≥60 and <70	**Ref**	**Ref**	**Ref**	**Ref**
	≥70 and <80	**1.02 (0.93,1.11)**	**-0.02 (-0.06,0.01)**	**0.93 (0.80,1.08)**	**-0.30 (-0.36,-0.23)**
	≥80	**1.13 (1.00,1.28)**	**0.06 (0.01,0.12)**	**0.82 (0.65,1.02)**	**-0.45 (-0.54,-0.35)**
Sex		**0.88 (0.83,0.94)**	**0.16 (0.13,0.18)**	**1.46 (1.31,1.62)**	-0.03 (-0.07,0.01)
Townsend deprivation score quintile	1	**Ref**	**Ref**	Ref	**Ref**
	2	**1.14 (1.05,1.24)**	**-0.03 (-0.06,0.01)**	0.89 (0.76,1.03)	**0.06 (-0.01,0.12)**
	3	**1.07 (0.98,1.16)**	**-0.02 (-0.05,0.02)**	0.92 (0.80,1.07)	**0.08 (0.02,0.15)**
	4	**1.16 (1.07,1.27)**	**-0.05 (-0.09,-0.02)**	0.85 (0.73,0.99)	**0.15 (0.09,0.22)**
	5	**1.31 (1.19,1.45)**	**-0.05 (-0.09,-0.01)**	0.90 (0.76,1.07)	**0.27 (0.19,0.34)**
Ethnic group	White	**Ref**	**Ref**	Ref	Ref
	Black	**0.55 (0.41,0.73)**	**0.29 (0.19,0.40)**	1.17 (0.74,1.85)	0.07 (-0.12,0.27)
	South Asian	**0.51 (0.41,0.64)**	**-0.04 (-0.12,0.04)**	1.07 (0.74,1.53)	0.07 (-0.08,0.23)
	Other	**0.63 (0.47,0.85)**	**0.05 (-0.06,0.15)**	0.90 (0.55,1.47)	0.08 (-0.13,0.28)
CKD		**1.20 (1.11,1.30)**	**-0.06 (-0.10,-0.03)**	**0.87 (0.76,1.01)**	**0.07 (0.01,0.13)**
Statin dose (simvastatin or equivalent) (mg)	≤10	**0.77 (0.70,0.85)**	**0.07 (0.03,0.12)**	**1.03 (0.85,1.24)**	**-0.24 (-0.31,-0.17)**
	20	**0.76 (0.71,0.81)**	**0.05 (0.03,0.08)**	**1.15 (1.03,1.29)**	**-0.18 (-0.23,-0.14)**
	40	**Ref**	**Ref**	**Ref**	**Ref**
	≥80	**0.92 (0.64,1.31)**	**0.10 (-0.07,0.26)**	**1.39 (0.78,2.49)**	**0.26 (-0.02,0.53)**
Time from diabetes diagnosis to first statin treatment (years)	≥0 and <1	**Ref**	**Ref**	**Ref**	**Ref**
	≥1 and <2	**1.06 (0.96,1.16)**	**-0.10 (-0.14,-0.07)**	**1.27 (1.08,1.49)**	**-0.12 (-0.19,-0.06)**
	≥2 and <3	**1.07 (0.96,1.19)**	**-0.13 (-0.18,-0.09)**	**1.41 (1.17,1.71)**	**0.05 (-0.03,0.13)**
	≥3 and <4	**1.18 (1.04,1.32)**	**-0.22 (-0.27,-0.17)**	**1.40 (1.14,1.72)**	**0.24 (0.15,0.33)**
	≥4 and <6	**1.10 (0.99,1.23)**	**-0.23 (-0.28,-0.19)**	**1.14 (0.94,1.39)**	**0.39 (0.31,0.47)**
	≥6 and <8	**1.15 (1.02,1.30)**	**-0.28 (-0.33,-0.22)**	**1.12 (0.89,1.41)**	**0.67 (0.58,0.76)**
	≥8 and <10	**1.09 (0.96,1.25)**	**-0.31 (-0.37,-0.25)**	**1.17 (0.91,1.50)**	**0.72 (0.62,0.82)**
	≥10 and <15	**1.24 (1.10,1.39)**	**-0.33 (-0.38,-0.28)**	**1.09 (0.87,1.36)**	**0.93 (0.84,1.02)**
	≥15 and <20	**1.36 (1.17,1.58)**	**-0.28 (-0.35,-0.21)**	**0.89 (0.65,1.23)**	**1.03 (0.91,1.14)**
	≥20	**1.55 (1.35,1.79)**	**-0.33 (-0.39,-0.26)**	**1.09 (0.82,1.43)**	**0.93 (0.83,1.04)**
Calender year of first statin prescription	2004	**2.08 (1.81,2.39)**	**0.08 (0.02,0.14)**	**0.40 (0.32,0.51)**	**0.23 (0.12,0.33)**
	2005	**1.81 (1.59,2.07)**	**0.01 (-0.04,0.07)**	**0.45 (0.36,0.55)**	**0.25 (0.15,0.35)**
	2006	**1.43 (1.25,1.63)**	**-0.10 (-0.15,-0.04)**	**0.41 (0.33,0.51)**	**0.15 (0.05,0.25)**
	2007	**1.17 (1.02,1.35)**	**-0.03 (-0.09,0.02)**	**0.65 (0.52,0.81)**	**0.11 (0.00,0.21)**
	2008	**1.04 (0.90,1.20)**	**-0.03 (-0.09,0.03)**	**0.65 (0.52,0.81)**	**0.08 (-0.02,0.19)**
	2009	**1.07 (0.92,1.25)**	**-0.04 (-0.10,0.02)**	**0.85 (0.68,1.06)**	**0.12 (0.00,0.23)**
	2010	**Ref**	**Ref**	**Ref**	**Ref**
	2011	**1.05 (0.89,1.23)**	**0.08 (0.02,0.14)**	**1.05 (0.84,1.32)**	**-0.05 (-0.17,0.07)**
	2012	**1.14 (0.94,1.38)**	**0.03 (-0.05,0.11)**	**1.22 (0.93,1.59)**	**0.06 (-0.08,0.21)**
COPD		1.09 (0.95,1.25)	**-0.07 (-0.13,-0.01)**	0.95 (0.74,1.21)	**0.16 (0.05,0.26)**
Psychosis		1.18 (0.88,1.58)	0.03 (-0.10,0.16)	**1.64 (1.06,2.54)**	**-0.27 (-0.50,-0.05)**
Asthma		0.98 (0.90,1.07)	0.03 (-0.01,0.06)	1.08 (0.93,1.25)	-0.04 (-0.10,0.02)
Atrial fibrilation		1.08 (0.96,1.22)	**-0.06 (-0.11,-0.01)**	**1.32 (1.07,1.62)**	0.04 (-0.06,0.13)
Anti-hypertensive drug treatment		**1.06 (0.99,1.12)**	**-0.13 (-0.16,-0.11)**	**1.11 (1.00,1.24)**	**-0.17 (-0.22,-0.12)**
Bile acid		0.59 (0.16,2.19)	0.30 (-0.16,0.75)	**6.48 (1.89,22.22)**	**-1.26 (-2.30,-0.22)**
Ezetimibe		0.89 (0.54,1.48)	0.03 (-0.17,0.23)	1.47 (0.72,3.01)	0.10 (-0.28,0.47)
Fibrate		1.04 (0.83,1.31)	0.00 (-0.10,0.10)	1.22 (0.82,1.82)	0.11 (-0.07,0.28)
Nicotinic acid		1.13 (0.25,5.21)	0.34 (-0.34,1.02)	1.00 (1.00,1.00)	-0.42 (-1.53,0.69)
Omega 3		**0.44 (0.19,1.03)**	-0.06 (-0.33,0.21)	1.72 (0.63,4.67)	**-0.54 (-1.11,0.03)**
Respiratory infection		0.99 (0.89,1.11)	-0.02 (-0.07,0.02)	1.13 (0.94,1.36)	-0.02 (-0.10,0.06)
Heavy drinker		1.03 (0.94,1.12)	**0.04 (0.00,0.07)**	1.11 (0.95,1.29)	**-0.22 (-0.29,-0.16)**
CHD		1.01 (0.89,1.16)	**0.07 (0.01,0.13)**	**1.34 (1.07,1.68)**	**0.11 (0.01,0.21)**
Stroke		0.92 (0.78,1.09)	0.05 (-0.02,0.12)	1.15 (0.85,1.54)	**0.26 (0.13,0.38)**
Death		**1.10 (0.98,1.23)**	-0.02 (-0.07,0.03)	0.94 (0.75,1.17)	**0.09 (0.00,0.17)**

OR: odds ratio; CI: confidence interval; LDL: low-density lipoprotein; HbA_{1c}: glycosylated haemoglobin; CKD: chronic kidney disease; COPD: chronic obstructive pulmonary disease; Psychoses: schizophrenia, bipolar disorder and other psychoses; CHD: coronary heart disease; Ref: reference category
In bold if variable is associated (p-value<0.100) with the chance of observing LDL cholesterol and HbA_{1c} and their respective values at baseline

Table 6.11: Association between the outcome (difference between total cholesterol 6 months before initiating statin treatment and 6 months after first statin treatment), covariates and auxiliary variables and the chance of observing GFR and its values at baseline

Variable		**GFR (ml min^{-1})**	
		Missing	**Value**
		OR (95% CI)	**Coefficient (95% CI)**
Outcome		0.99 (0.96,1.03)	**0.83 (0.50,1.15)**
Age (years) at first statin prescription	≥30 and <40	**1.29 (1.10,1.53)**	**8.13 (6.72,9.53)**
	≥40 and <50	**1.09 (0.98,1.22)**	**5.13 (4.22,6.03)**
	≥50 and <60	**1.07 (0.97,1.18)**	**2.60 (1.80,3.41)**
	≥60 and <70	**Ref**	**Ref**
	≥70 and <80	**1.03 (0.93,1.15)**	**-2.93 (-3.81,-2.04)**
	≥80	**0.87 (0.75,1.02)**	**-6.21 (-7.46,-4.97)**
Sex		1.02 (0.95,1.10)	**-1.51 (-2.10,-0.92)**
Townsend deprivation score quintile	1	**Ref**	Ref
	2	**1.16 (1.04,1.28)**	-0.38 (-1.22,0.46)
	3	**1.23 (1.11,1.36)**	-0.57 (-1.41,0.27)
	4	**1.27 (1.15,1.42)**	-0.54 (-1.40,0.33)
	5	**1.24 (1.10,1.40)**	-0.95 (-1.94,0.03)
Ethnic group	White	Ref	Ref
	Black	1.16 (0.85,1.58)	-1.73 (-4.31,0.85)
	South Asian	1.07 (0.84,1.37)	0.82 (-1.18,2.82)
	Other	0.80 (0.58,1.11)	-0.28 (-2.72,2.15)
CKD		**0.89 (0.81,0.99)**	**-17.68 (-18.52,-16.84)**
Statin dose (simvastatin or equivalent) (mg)	≤10	**0.80 (0.71,0.91)**	0.67 (-0.36,1.71)
	20	**0.93 (0.86,1.00)**	0.51 (-0.13,1.14)
	40	**Ref**	Ref
	≥80	**1.47 (0.91,2.38)**	0.18 (-4.57,4.94)
Time from diabetes diagnosis to first statin treatment (years)	≥0 and <1	**Ref**	Ref
	≥1 and <2	**0.80 (0.71,0.89)**	-0.45 (-1.33,0.43)
	≥2 and <3	**0.84 (0.73,0.95)**	0.20 (-0.87,1.27)
	≥3 and <4	**0.82 (0.71,0.95)**	-0.55 (-1.76,0.67)
	≥4 and <6	**0.84 (0.74,0.96)**	0.10 (-1.00,1.19)
	≥6 and <8	**0.91 (0.78,1.05)**	0.30 (-0.94,1.55)
	≥8 and <10	**0.93 (0.79,1.10)**	0.03 (-1.36,1.43)
	≥10 and <15	**0.90 (0.77,1.04)**	-0.66 (-1.91,0.59)
	≥15 and <20	**0.96 (0.79,1.15)**	-0.85 (-2.48,0.79)
	≥20	**0.99 (0.83,1.18)**	-2.18 (-3.70,-0.66)
Calender year of first statin prescription	2004	**71.19 (58.04,87.33)**	**0.49 (-1.75,2.73)**
	2005	**30.55 (25.88,36.05)**	**3.55 (2.08,5.03)**
	2006	**8.75 (7.50,10.20)**	**1.36 (0.19,2.52)**
	2007	**1.92 (1.63,2.26)**	**-0.66 (-1.75,0.44)**
	2008	**1.26 (1.07,1.50)**	**-0.11 (-1.19,0.97)**
	2009	**1.08 (0.91,1.30)**	**-0.49 (-1.61,0.62)**
	2010	**Ref**	**Ref**
	2011	**0.98 (0.81,1.19)**	**-1.18 (-2.37,0.01)**
	2012	**1.07 (0.85,1.35)**	**-1.43 (-2.88,0.03)**
COPD		1.05 (0.89,1.25)	-0.07 (-1.49,1.36)
Psychosis		0.77 (0.53,1.11)	-0.58 (-3.31,2.16)
Asthma		**0.91 (0.82,1.02)**	0.24 (-0.61,1.09)
Atrial fibrilation		1.02 (0.88,1.19)	-0.14 (-1.44,1.16)
Anti-hypertensive drug treatment		**0.91 (0.84,0.98)**	**-1.00 (-1.61,-0.39)**
Bile acid		1.03 (0.24,4.41)	0.66 (-10.86,12.19)
Ezetimibe		1.26 (0.72,2.18)	0.04 (-4.43,4.51)
Fibrate		1.24 (0.93,1.66)	**-3.77 (-6.22,-1.31)**
Nicotinic acid		1.00 (1.00,1.00)	-
Omega 3		0.76 (0.31,1.90)	-0.29 (-6.84,6.27)
Respiratory infection		0.94 (0.82,1.07)	**-1.12 (-2.19,-0.05)**
Heavy drinker		0.95 (0.86,1.06)	**1.65 (0.82,2.49)**
CHD		0.94 (0.79,1.10)	0.70 (-0.77,2.17)
Stroke		1.07 (0.87,1.31)	-0.01 (-1.91,1.89)
Death		1.08 (0.93,1.25)	-0.04 (-1.41,1.34)

OR: odds ratio; CI: confidence interval; GFR: glomerular filtration rate; CKD: chronic kidney disease; COPD: chronic obstructive pulmonary disease; Psychoses: schizophrenia, bipolar disorder and other psychoses; CHD: coronary heart disease; Ref: reference category
In bold if variable is associated (p-value<0.100) with the chance of observing GFR and its values at baseline

Table 6.12: Association between the outcome (difference between total cholesterol 6 months before initiating statin treatment and 6 months after first statin treatment), covariates and auxiliary variables and the chance of observing smoking status and its values at baseline

Variable		**Smoking status**		
		Missing	**Ex-smokers**	**Current smokers**
		OR (95% CI)	**Value Coefficient (95% CI)**	**Value Coefficient (95% CI)**
Outcome		-0.01 (-0.05,0.03)	-0.01 (-0.05,0.03)	**0.05 (0.00,0.10)**
Age (years) at first statin prescription	≥30 and <40	**-0.78 (-0.97,-0.59)**	**-0.78 (-0.97,-0.59)**	**0.39 (0.21,0.57)**
	≥40 and <50	**-0.71 (-0.83,-0.60)**	**-0.71 (-0.83,-0.60)**	**0.21 (0.08,0.34)**
	≥50 and <60	**-0.40 (-0.50,-0.30)**	**-0.40 (-0.50,-0.30)**	**0.04 (-0.08,0.16)**
	≥60 and <70	**Ref**	**Ref**	**Ref**
	≥70 and <80	**-0.07 (-0.17,0.03)**	**-0.07 (-0.17,0.03)**	**-0.54 (-0.69,-0.39)**
	≥80	**-0.02 (-0.16,0.12)**	**-0.02 (-0.16,0.12)**	**-1.02 (-1.28,-0.76)**
Sex		**-0.80 (-0.87,-0.73)**	**-0.80 (-0.87,-0.73)**	**-0.40 (-0.49,-0.31)**
Townsend deprivation score quintile	1	**Ref**	**Ref**	**Ref**
	2	**0.00 (-0.10,0.10)**	**0.00 (-0.10,0.10)**	**0.10 (-0.04,0.24)**
	3	**0.05 (-0.05,0.15)**	**0.05 (-0.05,0.15)**	**0.51 (0.38,0.64)**
	4	**0.10 (-0.00,0.20)**	**0.10 (-0.00,0.20)**	**0.69 (0.56,0.83)**
	5	**0.20 (0.07,0.32)**	**0.20 (0.07,0.32)**	**0.92 (0.78,1.07)**
Ethnic group	White	**Ref**	**Ref**	**Ref**
	Black	**-0.84 (-1.20,-0.48)**	**-0.84 (-1.20,-0.48)**	**-0.72 (-1.13,-0.31)**
	South Asian	**-1.38 (-1.71,-1.06)**	**-1.38 (-1.71,-1.06)**	**-1.41 (-1.79,-1.04)**
	Other	**-0.96 (-1.34,-0.58)**	**-0.96 (-1.34,-0.58)**	**-0.85 (-1.28,-0.42)**
CKD		0.02 (-0.08,0.11)	0.02 (-0.08,0.11)	**-0.17 (-0.30,-0.03)**
Statin dose (simvastatin or equivalent) (mg)	≤10	**0.08 (-0.04,0.20)**	0.08 (-0.04,0.20)	-0.03 (-0.19,0.13)
	20	**0.04 (-0.04,0.12)**	0.04 (-0.04,0.12)	0.01 (-0.09,0.10)
	40	**Ref**	Ref	Ref
	≥80	**0.29 (-0.17,0.75)**	0.29 (-0.17,0.75)	0.35 (-0.16,0.87)
Time from diabetes diagnosis to first statin treatment (years)	≥0 and <1	**Ref**	**Ref**	Ref
	≥1 and <2	**0.02 (-0.09,0.14)**	**0.02 (-0.09,0.14)**	-0.04 (-0.18,0.10)
	≥2 and <3	**0.03 (-0.10,0.16)**	**0.03 (-0.10,0.16)**	0.02 (-0.14,0.18)
	≥3 and <4	**0.15 (0.00,0.29)**	**0.15 (0.00,0.29)**	-0.05 (-0.24,0.13)
	≥4 and <6	**0.11 (-0.02,0.24)**	**0.11 (-0.02,0.24)**	-0.04 (-0.20,0.13)
	≥6 and <8	**0.05 (-0.09,0.20)**	**0.05 (-0.09,0.20)**	-0.08 (-0.27,0.10)
	≥8 and <10	**-0.06 (-0.22,0.11)**	**-0.06 (-0.22,0.11)**	-0.17 (-0.38,0.04)
	≥10 and <15	**0.06 (-0.09,0.20)**	**0.06 (-0.09,0.20)**	-0.01 (-0.19,0.18)
	≥15 and <20	**-0.15 (-0.34,0.04)**	**-0.15 (-0.34,0.04)**	-0.18 (-0.43,0.06)
	≥20	**-0.44 (-0.63,-0.25)**	**-0.44 (-0.63,-0.25)**	-0.08 (-0.29,0.13)
Calender year of first statin prescription	2004	-0.25 (-0.42,-0.08)	**-0.25 (-0.42,-0.08)**	-0.10 (-0.31,0.10)
	2005	-0.10 (-0.25,0.06)	**-0.10 (-0.25,0.06)**	-0.07 (-0.27,0.12)
	2006	0.00 (-0.16,0.15)	**0.00 (-0.16,0.15)**	-0.13 (-0.33,0.06)
	2007	0.04 (-0.13,0.21)	**0.04 (-0.13,0.21)**	-0.01 (-0.22,0.20)
	2008	0.02 (-0.15,0.19)	**0.02 (-0.15,0.19)**	-0.03 (-0.24,0.18)
	2009	0.10 (-0.07,0.28)	**0.10 (-0.07,0.28)**	0.01 (-0.20,0.23)
	2010	Ref	**Ref**	Ref
	2011	-0.03 (-0.22,0.16)	**-0.03 (-0.22,0.16)**	0.07 (-0.16,0.30)
	2012	-0.15 (-0.39,0.08)	**-0.15 (-0.39,0.08)**	-0.08 (-0.36,0.20)
COPD		1.68 (1.47,1.90)	**1.68 (1.47,1.90)**	**2.35 (2.11,2.58)**
Psychosis		0.45 (0.03,0.87)	**0.45 (0.03,0.87)**	**1.22 (0.84,1.60)**
Asthma		**0.12 (0.02,0.23)**	**0.12 (0.02,0.23)**	**-0.20 (-0.33,-0.07)**
Atrial fibrilation		0.05 (-0.10,0.19)	0.05 (-0.10,0.19)	**-0.33 (-0.56,-0.10)**
Anti-hypertensive drug treatment		**0.03 (-0.05,0.10)**	0.03 (-0.05,0.10)	**-0.40 (-0.49,-0.31)**
Bile acid		-0.70 (-2.32,0.92)	-	-
Ezetimibe		-0.44 (-1.09,0.21)	-0.44 (-1.09,0.21)	-0.10 (-0.85,0.65)
Fibrate		0.21 (-0.06,0.49)	0.21 (-0.06,0.49)	-0.03 (-0.39,0.33)
Nicotinic acid		-0.31 (-2.15,1.54)	-0.31 (-2.15,1.54)	0.20 (-2.16,2.56)
Omega 3		0.05 (-0.90,0.99)	0.05 (-0.90,0.99)	0.44 (-0.59,1.48)
Respiratory infection		0.12 (-0.01,0.26)	**0.12 (-0.01,0.26)**	**0.15 (-0.00,0.31)**
Heavy drinker		0.43 (0.32,0.53)	**0.43 (0.32,0.53)**	**0.70 (0.58,0.82)**
CHD		0.20 (0.04,0.36)	**0.20 (0.04,0.36)**	**0.31 (0.11,0.51)**
Stroke		-0.04 (-0.24,0.17)	-0.04 (-0.24,0.17)	**0.41 (0.17,0.66)**
Death		0.10 (-0.04,0.24)	0.10 (-0.04,0.24)	**0.48 (0.30,0.66)**

OR: odds ratio; CI: confidence interval; CKD: chronic kidney disease; COPD: chronic obstructive pulmonary disease; Psychoses: schizophrenia, bipolar disorder and other psychoses; CHD: coronary heart disease; Ref: reference category
In bold if variable is associated (p-value<0.100) with the chance of observing smoking status and its values at baseline

Table 6.13: Regression analysis following baseline multiple imputation with 5 imputations and using full information maximum likelihood (FIML) to investigate the association between the baseline variables and the difference between total cholesterol measurements records before initiating statin treatment and in the first six months after first statin treatment for patients with type II diabetes (N=21,237)

		Baseline multiple imputation						**FIML analysis**	
		Unadjusted analysis		**Adjusted analysis***		**Fully adjusted analysis**			
		Coefficient (95% CI)	**p-value**	**Coefficient (95% CI)**	**p-value**	**Coefficient (95% CI)**	**p-value**	**Coefficient (95% CI)**	**p-value**
Age (years) at first statin prescription	≥30 and <40	0.13 (0.07,0.18)	<0.001	0.12 (0.07,0.18)	<0.001	0.17 (0.11,0.23)	<0.001	0.17 (0.12,0.23)	<0.001
	≥40 and <50	0.10 (0.06,0.14)		0.10 (0.06,0.14)		0.14 (0.10,0.18)		0.14 (0.10,0.18)	
	≥50 and <60	0.07 (0.04,0.11)		0.07 (0.04,0.11)		0.11 (0.07,0.14)		0.10 (0.07,0.14)	
	≥60 and <70	Ref		Ref		Ref		Ref	
	≥70 and <80	-0.04 (-0.08,-0.00)		-0.04 (-0.07,-0.00)		-0.05 (-0.09,-0.02)		-0.06 (-0.09,-0.02)	
	≥80	-0.10 (-0.15,-0.05)		-0.09 (-0.14,-0.04)		-0.08 (-0.13,-0.03)		-0.08 (-0.13,-0.03)	
Female		-0.07 (-0.10,-0.05)	<0.001	-0.07 (-0.09,-0.04)	<0.001	0.03 (0.00,0.05)	0.026	0.03 (0.01,0.06)	0.015
Townsend deprivation score quintile	1 (least)	Ref	0.003	Ref	0.039	Ref	0.219	Ref	0.091
	2	-0.01 (-0.05,0.02)		-0.01 (-0.05,0.02)		-0.01 (-0.05,0.02)		-0.02 (-0.05,0.01)	
	3	0.01 (-0.02,0.05)		0.01 (-0.03,0.05)		0.01 (-0.03,0.05)		0.00 (-0.03,0.04)	
	4	0.02 (-0.02,0.05)		0.01 (-0.03,0.05)		-0.01 (-0.04,0.03)		-0.01 (-0.04,0.02)	
	5 (most)	0.07 (0.03,0.11)		0.05 (0.01,0.09)		0.03 (-0.01,0.08)		0.04 (-0.00,0.07)	
Ethnic group	White	Ref	<0.001	Ref	<0.001	Ref	<0.001	Ref	<0.001
	Black	0.28 (0.17,0.40)		0.27 (0.16,0.38)		0.32 (0.21,0.42)		0.32 (0.21,0.42)	
	South Asian	0.11 (0.02,0.19)		0.07 (-0.01,0.16)		0.00 (-0.08,0.08)		0.00 (-0.08,0.08)	
	Other	0.00 (-0.12,0.11)		-0.03 (-0.15,0.08)		-0.03 (-0.14,0.07)		-0.03 (-0.14,0.07)	
BMI (per 100kg m^{-2})		0.01 (-0.23,0.24)	0.943	-0.18 (-0.43,0.07)	0.150	-0.18 (-0.42,0.05)	0.111	-0.22 (-0.45,0.01)	0.061
Systolic blood pressure (per 100mmHg)		-0.26 (-0.33,-0.19)	<0.001	-0.20 (-0.27,-0.13)	<0.001	-0.05 (-0.12,0.02)	0.130	-0.06 (-0.12,0.01)	0.092
HDL cholesterol (mmol l^{-1})		0.00 (-0.03,0.04)	0.923	0.05 (0.01,0.09)	0.007	0.01 (-0.03,0.04)	0.718	-0.01 (-0.04,0.03)	0.755
LDL cholesterol (mmol l^{-1})		-0.49 (-0.51,-0.48)	<0.001	-0.50 (-0.52,-0.49)	<0.001	-0.49 (-0.51,-0.48)	<0.001	-0.49 (-0.51,-0.48)	<0.001
HbA_{1c} (per 100%)		-0.21 (-0.99,0.56)	0.585	-1.14 (-1.92,-0.35)	0.005	-1.17 (-1.92,-0.42)	0.002	-1.03 (-1.8,-0.26)	0.009
GFR (per 100ml min^{-1})		0.42 (0.34,0.50)	<0.001	0.31 (0.22,0.40)	<0.001	0.24 (0.15,0.34)	<0.001	0.23 (0.14,0.33)	<0.001

Smoking status	Non-smoker	Ref	<0.001	Ref	0.003	Ref	<0.001	Ref	0.002
	Ex-smoker	0.00 (-0.03,0.02)		0.00 (-0.03,0.03)		0.00 (-0.03,0.03)		0.00 (-0.03,0.03)	
	Current smoker	0.08 (0.05,0.12)		0.06 (0.03,0.10)		0.07 (0.03,0.11)		0.06 (0.03,0.09)	
Statin dose (simvastatin or equivalent) (mg)	≤10	0.42 (0.38,0.46)	<0.001	0.45 (0.41,0.49)	<0.001	0.41 (0.38,0.45)	<0.001	0.42 (0.38,0.46)	<0.001
	20	0.18 (0.16,0.21)		0.20 (0.17,0.22)		0.21 (0.18,0.23)		0.21 (0.18,0.23)	
	40	Ref		Ref		Ref		Ref	
	≥80	-0.28 (-0.44,-0.13)		-0.30 (-0.45,-0.15)		-0.16 (-0.31,-0.00)		-0.16 (-0.30,-0.02)	
Time from diabetes diagnosis to first statin treatment (years)	≥0 and <1	Ref	<0.001	Ref	<0.001	Ref	0.029	Ref	0.021
	≥1 and <2	0.02 (-0.02,0.06)		0.02 (-0.02,0.06)		-0.04 (-0.08,-0.00)		-0.04 (-0.08,-0.01)	
	≥2 and <3	0.03 (-0.02,0.08)		0.03 (-0.01,0.08)		-0.04 (-0.09,0.00)		-0.04 (-0.08,0.00)	
	≥3 and <4	0.04 (-0.01,0.09)		0.04 (-0.01,0.09)		-0.07 (-0.12,-0.02)		-0.07 (-0.12,-0.03)	
	≥4 and <6	0.09 (0.04,0.13)		0.09 (0.04,0.14)		-0.05 (-0.09,-0.00)		-0.05 (-0.09,-0.00)	
	≥6 and <8	0.12 (0.07,0.17)		0.12 (0.07,0.17)		-0.04 (-0.09,0.01)		-0.04 (-0.09,0.01)	
	≥8 and <10	0.14 (0.09,0.20)		0.15 (0.09,0.21)		-0.04 (-0.09,0.02)		-0.04 (-0.09,0.02)	
	≥10 and <15	0.16 (0.11,0.21)		0.18 (0.13,0.23)		-0.02 (-0.07,0.03)		-0.02 (-0.07,0.03)	
	≥15 and <20	0.16 (0.09,0.22)		0.17 (0.11,0.24)		0.01 (-0.05,0.08)		0.01 (-0.05,0.07)	
	≥20	0.24 (0.18,0.30)		0.23 (0.17,0.29)		0.02 (-0.03,0.08)		0.03 (-0.03,0.09)	
Calender year of first statin prescription	2004	-0.04 (-0.09,0.02)	<0.001	-0.02 (-0.07,0.04)	<0.001	-0.08 (-0.14,-0.03)	<0.001	-0.08 (-0.14,-0.03)	<0.001
	2005	-0.03 (-0.08,0.03)		-0.01 (-0.07,0.04)		-0.11 (-0.16,-0.06)		-0.11 (-0.16,-0.06)	
	2006	0.05 (-0.01,0.10)		0.06 (0.00,0.11)		-0.08 (-0.14,-0.03)		-0.08 (-0.13,-0.03)	
	2007	0.09 (0.04,0.15)		0.10 (0.04,0.16)		0.01 (-0.05,0.06)		0.01 (-0.05,0.06)	
	2008	0.08 (0.02,0.14)		0.09 (0.03,0.15)		0.02 (-0.04,0.08)		0.02 (-0.03,0.07)	
	2009	0.00 (-0.07,0.06)		0.00 (-0.06,0.06)		-0.03 (-0.08,0.03)		-0.03 (-0.08,0.03)	
	2010	Ref		Ref		Ref		Ref	
	2011	0.00 (-0.07,0.07)		0.00 (-0.07,0.07)		0.06 (-0.00,0.12)		0.05 (-0.01,0.11)	
	2012	-0.05 (-0.13,0.03)		-0.05 (-0.13,0.03)		0.01 (-0.07,0.08)		0.01 (-0.07,0.08)	

*Regression analysis for age, sex and ethnicity adjusted for each other
CI: confidence interval; BMI: body mass index; HDL: high-density lipoprotein; LDL: low-density lipoprotein;
HbA_{1c}: glycosylated haemoglobin; GFR: glomerular filtration rate; Ref: Reference category

Table 6.14: Regression analysis of data imputed using two-fold FCS algorithm with 5 imputations, 20 among-time iterations and 5 within-time iterations (conditional on time points before first statin treatment) to investigate the association between the baseline variables and the difference between total cholesterol measurements records before initiating statin treatment and in the first six months after first statin treatment for patients with type II diabetes (N=21,237)

		Unadjusted analysis		**Adjusted analysis***		**Fully adjusted analysis**	
		Coefficient (95% CI)	**p-value**	**Coefficient (95% CI)**	**p-value**	**Coefficient (95% CI)**	**p-value**
Age (years) at first statin prescription	≥30 and <40	0.13 (0.07,0.18)	<0.001	0.12 (0.07,0.18)	<0.001	0.16 (0.10,0.21)	<0.001
	≥40 and <50	0.10 (0.06,0.14)		0.10 (0.06,0.14)		0.13 (0.09,0.17)	
	≥50 and <60	0.07 (0.04,0.11)		0.07 (0.04,0.11)		0.10 (0.06,0.13)	
	≥60 and <70	Ref		Ref		Ref	
	≥70 and <80	-0.04 (-0.08,-0.00)		-0.04 (-0.07,-0.00)		-0.05 (-0.09,-0.02)	
	≥80	-0.10 (-0.15,-0.05)		-0.09 (-0.14,-0.04)		-0.10 (-0.14,-0.05)	
Female		-0.07 (-0.10,-0.05)	<0.001	-0.07 (-0.09,-0.04)	<0.001	0.03 (0.00,0.05)	0.019
Townsend score quintile	1 (least)	Ref	0.003	Ref	0.039	Ref	0.090
	2	-0.01 (-0.05,0.02)		-0.01 (-0.05,0.02)		-0.01 (-0.05,0.02)	
	3	0.01 (-0.02,0.05)		0.01 (-0.03,0.05)		0.00 (-0.03,0.04)	
	4	0.02 (-0.02,0.05)		0.01 (-0.03,0.05)		0.00 (-0.04,0.04)	
	5 (most)	0.07 (0.03,0.11)		0.05 (0.01,0.09)		0.04 (0.00,0.08)	
Ethnic group	White	Ref	<0.001	Ref	<0.001	Ref	<0.001
	Black	0.28 (0.17,0.40)		0.27 (0.16,0.38)		0.32 (0.21,0.42)	
	South Asian	0.11 (0.02,0.19)		0.07 (-0.01,0.16)		0.01 (-0.07,0.09)	
	Other	0.00 (-0.12,0.11)		-0.03 (-0.15,0.08)		-0.03 (-0.14,0.08)	
BMI ($100kgm^{-2}$)		-0.01 (-0.24,0.22)	0.929	-0.19 (-0.43,0.05)	0.111	-0.22 (-0.53,0.09)	0.132
Systolic blood pressure (100mmHg)		-0.25 (-0.33,-0.18)	<0.001	-0.20 (-0.27,-0.13)	<0.001	-0.06 (-0.13,0.01)	0.104
HDL cholesterol ($mmolL^{-1}$)		0.00 (-0.04,0.03)	0.887	0.05 (0.01,0.08)	0.009	0.01 (-0.03,0.04)	0.745
LDL cholesterol ($mmolL^{-1}$)		-0.49 (-0.51,-0.47)	<0.001	-0.50 (-0.52,-0.48)	<0.001	-0.49 (-0.51,-0.47)	<0.001
Glycosylated haemoglobin (100%)		-0.13 (-0.90,0.65)	0.745	-1.05 (-1.84,-0.26)	0.009	-1.05 (-1.82,-0.28)	0.007
Glomerular filtration rate ($100mLmin^{-1}$)		0.41 (0.29,0.53)	<0.001	0.30 (0.17,0.43)	<0.001	0.26 (0.16,0.36)	<0.001

Table 6.14: (continued)

		Unadjusted analysis		Adjusted analysis*		Fully adjusted analysis	
		Coefficient (95% CI)	**p-value**	**Coefficient (95% CI)**	**p-value**	**Coefficient (95% CI)**	**p-value**
Smoking status	Non-smoker	Ref	<0.001	Ref	0.008	Ref	<0.001
	Ex-smoker	0.00 (-0.03,0.03)		0.01 (-0.02,0.03)		0.01 (-0.02,0.04)	
	Current smoker	0.08 (0.04,0.11)		0.06 (0.02,0.09)		0.07 (0.03,0.10)	
Simvastatin dose or equivalent (mg)	≤10	0.42 (0.38,0.46)	<0.001	0.45 (0.41,0.49)	<0.001	0.41 (0.37,0.46)	<0.001
	20	0.18 (0.16,0.21)		0.20 (0.17,0.22)		0.21 (0.18,0.23)	
	40	Ref		Ref		Ref	
	≥80	-0.28 (-0.44,-0.13)		-0.30 (-0.45,-0.15)		-0.18 (-0.33,-0.04)	
Time from diabetes diagnosis to first statin treatment (years)	≥0 and <1	Ref	<0.001	Ref	<0.001	Ref	0.014
	≥1 and <2	0.02 (-0.02,0.06)		0.02 (-0.02,0.06)		-0.04 (-0.08,-0.00)	
	≥2 and <3	0.03 (-0.02,0.08)		0.03 (-0.01,0.08)		-0.04 (-0.08,0.01)	
	≥3 and <4	0.04 (-0.01,0.09)		0.04 (-0.01,0.09)		-0.07 (-0.12,-0.02)	
	≥4 and <6	0.09 (0.04,0.13)		0.09 (0.04,0.14)		-0.05 (-0.09,-0.00)	
	≥6 and <8	0.12 (0.07,0.17)		0.12 (0.07,0.17)		-0.03 (-0.08,0.02)	
	≥8 and <10	0.14 (0.09,0.20)		0.15 (0.09,0.21)		-0.03 (-0.08,0.02)	
	≥10 and <15	0.16 (0.11,0.21)		0.18 (0.13,0.23)		-0.02 (-0.06,0.03)	
	≥15 and <20	0.16 (0.09,0.22)		0.17 (0.11,0.24)		0.01 (-0.05,0.08)	
	≥20	0.24 (0.18,0.30)		0.23 (0.17,0.29)		0.03 (-0.03,0.09)	
Calender year of first statin prescription	2004	-0.04 (-0.09,0.02)	<0.001	-0.02 (-0.07,0.04)	<0.001	-0.08 (-0.14,-0.03)	<0.001
	2005	-0.03 (-0.08,0.03)		-0.01 (-0.07,0.04)		-0.11 (-0.17,-0.06)	
	2006	0.05 (-0.01,0.10)		0.06 (0.00,0.11)		-0.08 (-0.13,-0.03)	
	2007	0.09 (0.04,0.15)		0.10 (0.04,0.16)		0.01 (-0.04,0.06)	
	2008	0.08 (0.02,0.14)		0.09 (0.03,0.15)		0.02 (-0.04,0.07)	
	2009	0.00 (-0.07,0.06)		0.00 (-0.06,0.06)		-0.03 (-0.08,0.03)	
	2010	Ref		Ref		Ref	
	2011	0.00 (-0.07,0.07)		0.00 (-0.07,0.07)		0.05 (-0.01,0.10)	
	2012	-0.05 (-0.13,0.03)		-0.05 (-0.13,0.03)		0.00 (-0.07,0.07)	

*Regression analysis for age, sex and ethnicity adjusted for each other
CI: confidence interval; BMI: body mass index; HDL: high-density lipoprotein; LDL: low-density lipoprotein; HbA_{1c}: glycosylated haemoglobin; GFR: glomerular filtration rate; Ref: Reference category

Table 6.15: Regression analysis of data imputed using two-fold FCS algorithm with 5 imputations, 20 among-time iterations and 5 within-time iterations (conditional on time points before and after first statin treatment) to investigate the association between the baseline variables and the difference between total cholesterol measurements records before initiating statin treatment and in the first six months after first statin treatment for patients with type II diabetes (N=21,237)

		Unadjusted analysis		**Adjusted analysis***		**Fully adjusted analysis**	
		Coefficient (95% CI)	**p-value**	**Coefficient (95% CI)**	**p-value**	**Coefficient (95% CI)**	**p-value**
Age (years) at first statin prescription	≥30 and <40	0.13 (0.07,0.18)	<0.001	0.12 (0.07,0.18)	<0.001	0.16 (0.11,0.22)	<0.001
	≥40 and <50	0.10 (0.06,0.14)		0.10 (0.06,0.14)		0.12 (0.08,0.16)	
	≥50 and <60	0.07 (0.04,0.11)		0.07 (0.04,0.11)		0.11 (0.07,0.14)	
	≥60 and <70	Ref		Ref		Ref	
	≥70 and <80	-0.04 (-0.08,-0.00)		-0.04 (-0.07,-0.00)		-0.05 (-0.09,-0.01)	
	≥80	-0.10 (-0.15,-0.05)		-0.09 (-0.14,-0.04)		-0.11 (-0.17,-0.06)	
Female		-0.07 (-0.10,-0.05)	<0.001	-0.07 (-0.09,-0.04)	<0.001	0.03 (0.00,0.06)	0.044
Townsend score quintile	1 (least)	Ref	0.003	Ref	0.039	Ref	0.064
	2	-0.01 (-0.05,0.02)		-0.01 (-0.05,0.02)		-0.03 (-0.07,0.00)	
	3	0.01 (-0.02,0.05)		0.01 (-0.03,0.05)		-0.02 (-0.05,0.02)	
	4	0.02 (-0.02,0.05)		0.01 (-0.03,0.05)		-0.02 (-0.06,0.02)	
	5 (most)	0.07 (0.03,0.11)		0.05 (0.01,0.09)		0.03 (-0.01,0.07)	
Ethnic group	White	Ref	<0.001	Ref	<0.001	Ref	0.001
	Black	0.28 (0.17,0.40)		0.27 (0.16,0.38)		0.23 (0.11,0.34)	
	South Asian	0.11 (0.02,0.19)		0.07 (-0.01,0.16)		0.03 (-0.06,0.11)	
	Other	0.00 (-0.12,0.11)		-0.03 (-0.15,0.08)		-0.04 (-0.16,0.07)	
BMI ($100kgm^{-2}$)		-0.05 (-0.26,0.17)	0.679	-0.22 (-0.45,0.00)	0.054	-0.21 (-0.43,0.01)	0.060
Systolic blood pressure (100mmHg)		-0.26 (-0.33,-0.18)	<0.001	-0.20 (-0.27,-0.13)	<0.001	-0.07 (-0.14,0.01)	0.066
HDL cholesterol ($mmolL^{-1}$)		0.00 (-0.04,0.03)	0.851	0.05 (0.01,0.08)	0.011	0.01 (-0.04,0.05)	0.810
LDL cholesterol ($mmolL^{-1}$)		-0.50 (-0.51,-0.48)	<0.001	-0.50 (-0.52,-0.49)	<0.001	-0.49 (-0.51,-0.48)	<0.001
Glycosylated haemoglobin (100%)		-0.15 (-0.94,0.64)	0.706	-1.07 (-1.87,-0.26)	0.009	-1.18 (-2.02,-0.33)	0.006
Glomerular filtration rate ($100mLmin^{-1}$)		0.40 (0.27,0.53)	<0.001	0.29 (0.14,0.43)	<0.001	0.21 (0.02,0.40)	0.009

Table 6.15: (continued)

		Unadjusted analysis		Adjusted analysis*		Fully adjusted analysis	
		Coefficient (95% CI)	**p-value**	**Coefficient (95% CI)**	**p-value**	**Coefficient (95% CI)**	**p-value**
Smoking status	Non-smoker	Ref	<0.001	Ref	0.018	Ref	<0.001
	Ex-smoker	0.00 (-0.03,0.03)		0.01 (-0.02,0.04)		0.01 (-0.01,0.04)	
	Current smoker	0.07 (0.04,0.11)		0.05 (0.02,0.08)		0.07 (0.03,0.10)	
Simvastatin dose or equivalent (mg)	≤10	0.42 (0.38,0.46)	<0.001	0.45 (0.41,0.49)	<0.001	0.42 (0.37,0.46)	<0.001
	20	0.18 (0.16,0.21)		0.20 (0.17,0.22)		0.20 (0.17,0.23)	
	40	Ref		Ref		Ref	
	≥80	-0.28 (-0.44,-0.13)		-0.30 (-0.45,-0.15)		-0.11 (-0.28,0.05)	
Time from diabetes diagnosis to first statin treatment (years)	≥0 and <1	Ref	<0.001	Ref	<0.001	Ref	0.007
	≥1 and <2	0.02 (-0.02,0.06)		0.02 (-0.02,0.06)		-0.04 (-0.08,-0.00)	
	≥2 and <3	0.03 (-0.02,0.08)		0.03 (-0.01,0.08)		-0.04 (-0.09,0.01)	
	≥3 and <4	0.04 (-0.01,0.09)		0.04 (-0.01,0.09)		-0.08 (-0.12,-0.03)	
	≥4 and <6	0.09 (0.04,0.13)		0.09 (0.04,0.14)		-0.04 (-0.08,0.01)	
	≥6 and <8	0.12 (0.07,0.17)		0.12 (0.07,0.17)		-0.04 (-0.10,0.01)	
	≥8 and <10	0.14 (0.09,0.20)		0.15 (0.09,0.21)		-0.04 (-0.10,0.02)	
	≥10 and <15	0.16 (0.11,0.21)		0.18 (0.13,0.23)		-0.03 (-0.08,0.03)	
	≥15 and <20	0.16 (0.09,0.22)		0.17 (0.11,0.24)		0.04 (-0.03,0.11)	
	≥20	0.24 (0.18,0.30)		0.23 (0.17,0.29)		0.05 (-0.02,0.11)	
Calender year of first statin prescription	2004	-0.04 (-0.09,0.02)	<0.001	-0.02 (-0.07,0.04)	<0.001	-0.06 (-0.12,-0.01)	<0.001
	2005	-0.03 (-0.08,0.03)		-0.01 (-0.07,0.04)		-0.10 (-0.16,-0.05)	
	2006	0.05 (-0.01,0.10)		0.06 (0.00,0.11)		-0.06 (-0.12,-0.01)	
	2007	0.09 (0.04,0.15)		0.10 (0.04,0.16)		0.02 (-0.03,0.08)	
	2008	0.08 (0.02,0.14)		0.09 (0.03,0.15)		0.03 (-0.03,0.09)	
	2009	0.00 (-0.07,0.06)		0.00 (-0.06,0.06)		-0.02 (-0.08,0.04)	
	2010	Ref		Ref		Ref	
	2011	0.00 (-0.07,0.07)		0.00 (-0.07,0.07)		0.07 (0.00,0.13)	
	2012	-0.05 (-0.13,0.03)		-0.05 (-0.13,0.03)		0.02 (-0.06,0.10)	

*Regression analysis for age, sex and ethnicity adjusted for each other
CI: confidence interval; BMI: body mass index; HDL: high-density lipoprotein; LDL: low-density lipoprotein; HbA_{1c}: glycosylated haemoglobin; GFR: glomerular filtration rate; Ref: Reference category

6.2.4 Discussion

In this chapter, I used routinely collected, primary care data to investigate which health indicators measured in the 6 months before initiating statin treatment were associated with greater total cholesterol reduction before and after initiating statin treatment in patients with type II diabetes. The motivation for this study was to discover which patients with type II diabetes clinicians could identify as unlikely to respond to statin treatment, such as those with atherogenic dyslipidaemia, and suggest a possible alternative procedure for treating these patients.

The main strength of this study is analysing a large, nationally representative primary care database. However, one limitation is the substantial missing health indicator values, which could potentially result in biased estimates and SE because a plausible MCAR assumption is unlikely. Therefore, I also considered different imputation approaches, which give unbiased estimates and standard errors under the more plausible MAR assumption, and including auxiliary variables in the imputation model increased the MAR assumption plausibility. The simulation results in chapter 5 found the estimates for time-dependent variables from analysing data imputed using the two-fold FCS algorithm were more precise compared to a 'baseline' imputation, so it is a suitable method to consider in this context.

To aid interpretation of the results in sections 6.2.3.2, I produced a summary table of the p-values (Table 6.16). The complete records analysis found older patients with higher LDL cholesterol, lower GFR and larger statin dose were associated with greater total cholesterol reduction (Table 6.16). These findings may suggest older patients had a higher total cholesterol at baseline and, therefore, more room for reduction compared to younger patients prescribed a lower statin dose at baseline. Another possible explanation is older patients with higher total cholesterol at baseline may be more likely to adhere to the statin treatment because they had higher CVD risk compared to younger patients with low total cholesterol measurement at baseline. These observations contradict one study which showed a higher treated total cholesterol ($>$5mmol l^{-1}) and younger age ($<$51 years) were less likely to adhere to statin treatment[109].

Low GFR (due to nephropathy) is common in patients with type 2 diabetes indicating higher risk of CKD[110] and CVD[111]. From the complete records analysis, patients with lower GFR had greater total cholesterol reduction after initiating statin treatment, again potentially because patients with higher CVD risk were more likely to adhere to statin treatment.

Table 6.16: Summary of p-values from using different methods of handling missing data for fully adjusted analysis to investigate the association between the baseline variables and the difference between total cholesterol measurements records before initiating statin treatment and in the first six months after first statin treatment for patients with type II diabetes

	Complete records analysis	Baseline MI	Two-fold FCS algorithm (before baseline)	Two-fold FCS algorithm (before and after baseline)
	p-value	p-value	p-value	p-value
Age (years) at first statin prescription	<0.001	<0.001	<0.001	<0.001
Sex	0.119	0.026	0.019	0.044
Townsend deprivation score quintile	0.456	0.219	0.090	0.064
Ethnic group	0.188	<0.001	<0.001	0.001
BMI (per 100kg m^{-2})	0.952	0.111	0.132	0.060
Systolic blood pressure (per 100mmHg)	0.569	0.130	0.104	0.064
HDL cholesterol (mmol l^{-1})	0.345	0.718	0.745	0.810
LDL cholesterol (mmol l^{-1})	<0.001	<0.001	<0.001	<0.001
HbA_{1c} (per 100%)	0.244	0.002	0.007	0.006
GFR (per 100ml min^{-1})	0.001	<0.001	<0.001	0.009
Smoking status	0.195	<0.001	<0.001	<0.001
Statin dose (mg)	<0.001	<0.001	<0.001	<0.001
Time from diabetes diagnosis to first statin treatment (years)	0.329	0.029	0.014	0.007
Calender year of first statin prescription	0.005	<0.001	<0.001	<0.001

FCS: fully conditional specification BMI: body mass index; HDL: high-density lipoprotein; LDL: low-density lipoprotein; HbA_{1c}: glycosylated haemoglobin; GFR: glomerular filtration rate

The analysis gained power when analysing data imputed using MI compared to complete records analysis. From analysing greater total cholesterol reduction following baseline MI, in addition to variables from the complete records analysis, male patients, white ethnicity, higher HbA_{1c}, not current smokers, shorter time from first diabetes diagnosis to initiating statin treatment and calendar year were also associated with greater total cholesterol reduction (Table 6.16).

Analysing data imputed using the two-fold FCS algorithm imputation method conditional on measurements before baseline had a little more power but found similar results to analysing data imputed using baseline MI

for the analysis of greater total cholesterol reduction. A possible explanation for finding similar results could be because of the substantial missing data for the health indicators at time blocks except baseline and the two-fold FCS algorithm can not gain much additional information from these time blocks. However, Townsend deprivation score quintile, BMI and systolic blood pressure were almost statistically significantly associated with greater total cholesterol reduction from analysing data imputed using the two-fold FCS algorithm conditional on measurements before and after baseline. The two-fold FCS algorithm could compensate for the substantial missing data at each time point by gaining more information from the additional time points after baseline.

Ferro *et al.* found FIML analysis showed the same results as MI[112] and recommended investigating FIML analysis compared to MI which conditions on the longitudinal data. Therefore, I compared FIML to using the two-fold FCS algorithm and I found they gave similar results, which agrees with Ferro *et al.* findings. FIML may not be a suitable method in this context because it was not developed to use the longitudinal data.

The two-fold FCS algorithm conditional on measurements before and after baseline was preferred to complete records analysis because more variables were associated with greater total cholesterol reduction following baseline MI compared to the complete records analysis, and these findings were similar to the descriptive analysis results (i.e. in the expected direction). The coefficient for HbA_{1c} and statin dose $\geq$80 also change direction of association, but all other coefficients were similar to the complete records analysis. Possibly just having more power resulted in these variables becoming significant, but also some of the odds ratios increased, for example ethnicity. Even though time from diabetes diagnosis to first statin treatment and calendar year of first statin prescription were statistically significantly associated with greatest total cholesterol reduction, it was only small differences and these are possibly not clinically significant differences (Table 6.16).

One limitation of this study was the 6 time blocks may have been too small when using the two-fold FCS algorithm. One possible explanation why the two-fold FCS algorithm did not improve the precision of more variables is because recording was low at other time blocks except baseline due to the small time blocks, so the correlations between repeated measurements at different time blocks were underestimated, even though I analysed data after 2004 when QOF was introduced to try and minimise the missing data. However, QOF required measurements every 15 month for these health indicators, so using 6 month time blocks reduced the records within each time block. I chose to use 6 month time blocks because I was interested in total cholesterol with 6 months after initiating statin treatment, so I kept all time blocks as 6 months for consistency. However, the two-fold FCS algorithm is flexible and can impute missing values in time blocks of varying size. Using these wider time points, the associations between Townsend deprivation score quintile, BMI, systolic blood pressure with greater total cholesterol reduction from analysing data imputed using the two-fold FCS algorithm conditional on measurements before and after baseline the two-fold FCS algorithm would possibly have been stronger.

A limitation for all MI methods was that I could only include smoking status measured at baseline in the imputation model because it was too sparsely recorded at other time blocks. But, as I showed in section 5.1, imputing

using repeated smoking status measurements gave more precise estimates compared to imputing at a single time block.

I chose to investigate total cholesterol reduction in the short time period before and after initiating statin treatment, rather than modelling the data longitudinally, because patients respond very quickly to statin treatment and after this initial reduction, total cholesterol levels remain constant if the patient adheres to statin treatment. Patients must adhere to statin therapy to keep a low total cholesterol. Lamberts *et al.* showed patients who initiate statin treatment after the initiation of oral anti-diabetic treatment (which would be the case for patients included in this cohort) were more likely to discontinue statin treatment than patients who initiate statin treatment before the start of oral anti-diabetics[113].

The health indicator measures at baseline for patients with smaller total cholesterol reduction suggest these patients had lower CVD risk at baseline compared to those with greater total cholesterol reduction. However, patients with smaller total cholesterol reduction were more likely to smoke compared to patients with greater total cholesterol reduction, which usually increases CVD risk. One possible explanation for this is because these patient may consider themselves at low risk of CVD so do not consider smoking to be a risk for them.

Recent update to the NICE lipid modification clinical guideline[114] proposes to offer high-intensity statin treatment for the primary prevention of CVD to people with type 2 diabetes and lower 10-year risk of developing CVD, from 20% to 10%, estimating risk using the UK prospective Diabetes Study assessment tool, but patients with low CVD risk may be reluctant to be prescribed long term statin treatment[115].

6.2.4.1 Clinical implications

This study showed that statins reduce total cholesterol in patients with type II diabetes, contrary to Standl *et al*[104]. Patients with type II diabetes prescribed a high dose statin therapy had the largest reduction in total cholesterol, possibly because these patients had high total cholesterol and CVD risk when statin treatment was initiated (such as those with atherogenic dyslipidaemia) and may be more likely to adhere to statin treatment.

I found that patients with type II diabetes had a greater reduction in total cholesterol if GFR was low when statins were initiated. After using MI, the analysis found that male patients, white ethnicity, high HbA_{1c}, not current smokers and longer time from diabetes diagnosis were also important predictors of greater total cholesterol reduction. Clinicians can use these findings to assess the potential CVD risk of patients with type II diabetes and initiate the appropriate statin therapy.

I recommend analysing data imputed using MI, but the selected method depends on the available longitudinal data. If regularly recorded, longitudinal data is available the two-fold FCS algorithm may be more beneficial than a 'baseline' MI.

Chapter 7

Discussion

7.1 Overview

The motivation for this thesis was to develop and validate a practical approach for addressing the issues raised by missing data in longitudinal studies using electronic clinical databases, such as The Health Improvement Network (THIN). Missing data is a common problem for many epidemiological studies using electronic health record databases of information collected in routine clinical practice and naïve application of 'ad–hoc' methods (such as complete records) can result in non-trivial bias and substantial loss of power. Appropriate use of multiple imputation (MI) can give unbiased estimates and standard errors under the assumption the data is missing at random (MAR)[15] and the model of interest and imputation model are congenial (the imputation models conditional distribution is the same as the model of interests predictive distribution)[20]. However, when I started this PhD, no established, validated imputation approach existed for missing data in longitudinal, electronic, clinical databases with many individuals, many time points, many variables and large proportions of missing data.

In previous work, Nevalainen *et al.*[1] proposed the two-fold fully conditional specification (FCS) to potentially impute missing data in longitudinal databases. Nevalainen *et al.* validated the two-fold FCS algorithm in a simple setting with 3 time points and a few variables with no validation of the two-fold FCS algorithm in complex electronic, health record settings such as THIN, and no published, validated software. Therefore, the overarching aim of this thesis was to adapt and evaluate the two-fold FCS algorithm and implement it to impute missing data in large primary care database to take into account the longitudinal and dynamic structures of these data.

The specific objectives were to:

1. investigate the extent and patterns of missing data in a longitudinal clinical database for health indicators associated with risk of cardiovascular disease;
2. develop methods to identify and remove outliers before imputation;
3. adapt and develop the two-fold FCS multiple imputation algorithm to impute missing values in longitudinal clinical data for variables associated with cardiovascular disease;
4. validate the two-fold FCS algorithm through challenging simulation studies; and

5. apply the two-fold FCS algorithm in THIN to (i) model risk for cardiovascular disease and (ii) understand factors associated with greater total cholesterol reduction in patients with type II diabetes.

In this discussion chapter, I firstly summarise the thesis and the findings. Secondly, I discuss the methodological issues I overcame. Next, I discuss the implications of the findings, and finally I present my overall conclusions and review possible areas of future research.

7.2 Summary of thesis and findings

In this thesis, I first described The Health Improvement Network (THIN)[11] primary care database of longitudinal, electronic health records in section 3.1. The data issues raised by missing data directly motivated this thesis. Primary care databases are a valuable resource for research for the following reasons: they contain information on populations not eligible for cohort studies or clinical trials, such as pregnant women and children[5, 6]; large patient numbers allow research on rare conditions[7]; and long follow-up time enable the development of risk prediction models in clinically relevant populations[8, 10]. Primary care databases contain records on patient characteristics, diagnoses, prescriptions and health indicators, such as weight, systolic blood pressure and smoking status, captured through routine consultations with a GP or nurse. For valid inferences, we require repeated measurements of many health indicators. However, issues arise when using data collected to address direct clinical need for research. For example, primary care databases may have substantial missing health indicator measurements if patients did not attend a consultation with a GP or nurse or if they attended a consultation and only information relevant to the consultation was recorded. Analyses of primary care data may be biased if missing data are not handled appropriately.

The THIN data both directly motivated the project and provided an ideal setting to achieve my aims of developing, implementing and evaluating the two-fold FCS algorithm because data was collected for patient care, not for research, so there was a substantial missing data. The large (11 million patients), nationally representative[29], routinely collect data has great potential for research use. However, missing data issues have to be effectively addressed if THINs (and other electronic health records) potential for answering research questions is to be realised.

I addressed many practical considerations before using THIN. The principal ones were selecting the cohort of patients to include in the analysis, or the time period to use. To develop Read code lists[26] for each disease, I received feedback on the lists from the GPs (Irwin Nazareth and Kate Walters) in the MRC funded project 'Missing data imputation in clinical databases: development of a longitudinal model for cardiovascular risk factors' Steering Group. The patient records in THIN are also not immediately available for analysis so, once these decisions were made, I performed considerable data manipulation and management to extract the required data and prepare it for analysis.

Historically, missing data were commonly handled using 'ad–hoc' methods, as discussed in section 3.2. These

methods typically attempt a simple 'fix' for the missing data, yielding a dataset on which we can perform the intended substantive analysis but, unfortunately, not derived using statistical reasoning so highly unlikely to achieve valid inferences for data with missingness mechanisms which were probably not plausible in THIN. For example, a complete records analysis (only patients with observed data for all records were included in the analysis) can achieve unbiased estimates if the missing completely at random (MCAR) assumption is plausible, and observed data are representative of all the data, but will overestimate the SEs because of loss of power[15]. A complete records analysis of THIN would probably give biased results because health indicator data were mainly recorded for unwell patients who attend a consultation and, therefore, were more likely to be included in the complete records analysis compared to well patients, who do not require a consultation. This suggests the observed data would not represent all the data and a complete records analysis would achieve biased estimates and SEs.

Multiple imputation (MI) is an alternative method for handling missing data based on sound statistical reasoning[18]. I described MI in Section 3.3 and explained it was more suitable to impute missing values compared to 'ad–hoc' methods because it gave unbiased estimates as well as standard errors under the assumption the data was missing at random (MAR, missingness was independent of the missing values conditional on the observed data) and the model of interest and imputation model are congenial (the imputation models conditional distribution is the same as the model of interests predictive distribution)[20]. The MAR assumption is a more plausible assumption about the THIN data than MCAR. Suppose we have a model of interest and a dataset we wish to fit the model of interest to, but some values of explanatory variables have missing values. We can use MI to create multiple imputed datasets. First, we specify the imputation model (different from the model of interest). MI imputes the missing data by selecting random draws from the conditional predictive distribution of the missing data given the observed so each imputed dataset is 'complete'. The model of interest is fitted to each 'complete' dataset and the results are correctly combined for final inference using Rubin's MI rules[18]. We can increase the MAR assumption plausibility by including auxiliary variables associated with the chance of observing variables with missing data in the imputation model[20]. Another advantage of MI is we can also include auxiliary variables associated with the value of the variables with missing data in the imputation model to reduce bias and increase precision in the analysis of the imputed data[20]. The attraction of MI is that it is established as broadly applicable and very flexible, so it is an ideal approach to adapt and develop methods to overcome missing data issues in this setting.

Traditionally, the MI proceeds by fitting a joint, adjusted normal conditional distribution. However, a multivariate normal distribution and the model of interest may not be appropriate, for example if the variables are ordered, as in longitudinal primary care data, or categorical, because the appropriate joint distribution is difficult to define[21]. In this situation, a more flexible and convenient method is fully conditional specification (FCS)[21]. This method avoids explicitly specifying the joint distribution, though is consistent with the multivariate normal distribution if appropriate assumptions are satisfied[49], and instead uses a series of conditional regression models. FCS proceeds as follows:

1. To initiate the procedure, it fills in missing values with randomly chosen observed values for each variable

in the imputation model in turn.

2. 'Filled-in' values in the first variable are discarded. A regression model for this variable conditional on other variables in the imputation model is specified and the missing values are replaced by random draws from the conditional distribution.

3. step 2 is repeated for each variable in turn. Imputed draws are 'proper', reflecting uncertainty in the parameter estimates. Once each variable is imputed, one 'cycle' is complete.

4. steps 2 and 3 are repeated for a pre-specified number of cycles, considered sufficient for convergence.

FCS is a useful and flexible alternative when we cannot specify a convenient and realistic joint distribution[45] and allows us to specify the correct distribution for each variable with missing data under the MAR assumption and compatibility of conditionals (the underlying joint multivariate normal distribution exists).

Before this thesis, we did not know which MI approach was most appropriate to impute missing data in longitudinal clinical records like THIN to take account of the longitudinal and dynamic structure of the data. Two possible approaches to impute longitudinal, clinical data are either to impute missing data at all time blocks separately or at all time blocks simultaneously. Imputing each time block separately results in imputed data which underestimates the correlations between time blocks. We can overcome this problem by imputing all time blocks simultaneously, but this may result in non-convergence due to co-linearity because of the high correlations between repeated measures of the same health indicators at different time blocks in large databases with many time blocks.

An alternative approach is to exploit the temporal structure of the data by simplifying the simultaneous method using the two-fold FCS algorithm[1], described in section 3.3.10. The two-fold FCS algorithm avoids co-linearity problems by restricting the time blocks included in the imputation model to a small time window but also takes account of the longitudinal structure of the data because imputation models included measurements at nearby time blocks. The two-fold FCS algorithm assumes conditional independence i.e. measurements outside of the time window are independent if we condition on measurements closer in time. The two-fold FCS algorithm readily copes with the intermittent missingness pattern because the imputation models condition on observed measurements before and after the time block of interest. Various parameters associated with the two-fold FCS algorithm, such as the window width, within-time iterations and among-time iterations, can be modified if appropriate for the missing data context.

7.2.1 Investigate the extent and patterns of missing data in a longitudinal clinical database for health indicators associated with risk of cardiovascular disease

I assessed the plausibility of the MAR assumption by investigating recording of health indicators (e.g. height, weight and systolic blood pressure) associated with cardiovascular disease (CVD). Initially I explored health indicator recording by age and sex and this investigation revealed the MCAR assumption was not plausible

because recording varied by age and sex. From this investigation, I found higher height, weight and systolic blood pressure annual recording for female patients aged less than 75 years compared to male patients. After age 75 years, annual recording decreased similarly for male and female patients. Most health indicator recording per consultation was similar for men and women. However, male patients aged less than 75 years had higher height measurement recording per consultation compared to female patients aged less than 75 years, and female patients had higher weight and systolic blood pressure recording per consultation below age 40 years. The distribution of annual recording of total cholesterol measurements was similar for male and female patients but, after adjusting for number of consultations, male patient had a higher recording of total cholesterol measurements compared to female patients. Male patients consult less frequently than female patients[80] and the health indicators recorded during consultations varied dependent on sex. Female patients were more likely to have weight and systolic blood recorded compared to men, especially at aged less than 40 years, which could be explained by female patients attending contraception or pregnancy related consultations. Male patients were more likely to have total cholesterol recorded compared to female patients, which might be because they have higher CVD risk compared to female patients.

I also investigated health indicator recording from 1995 to 2011 to explore the effects of two incentives introduced by the Department of Health (DoH). The first was 'new patient health checks', financial incentives which encouraged GPs to record health indicators for all newly registered patients, introduced in 1992[12]. The second was the Quality Outcomes Framework (QOF)[14], introduced in 2004, to encourage practices to record health indicators relevant to monitor patients with specific diseases.

I found the MCAR assumption appeared most plausible in the year after registration because practices performed 'new patient health checks' so GPs collected information regardless of age, sex or health status but conditional on registration status. However, after registration approximately two times more patients had health indicators recorded if diagnosed with one of the Quality Outcomes Framework (QOF) specified diseases (such as CVD or diabetes), which is evidence of a MAR assumption. Conditioning on these diseases in the imputation model increases the MAR assumption plausibility because it is less likely the probability of the variable being missing will depend on the missing values themselves, conditional on these observed auxiliary variables[20].

Total cholesterol recording over time was different to other health indicators, with substantially more missing data at registration (5% observed) but recording steadily increased over time. One possible reasons is because total cholesterol is not recorded at registration as part of the 'new patient health check'. Another possible reason is because a blood sample was required to measure total cholesterol, so it was mainly recorded if the GP suspected the patient had a disease, or presented with another indicator for CVD such as high body mass index (BMI) or high blood pressure, and a total cholesterol measurement was required to confirm high CVD risk. However, total cholesterol recording improved after 1999, at approximately 10%, when statins (a lipid-modifying drug) became more commonly prescribed. Statins and other lipid-modifying drugs lower total cholesterol, so GPs monitor patients treated with lipid-modifying drugs to determine if total cholesterol decreases. Total cholesterol recording

also increased with the introduction of QOF in 2004, to approximately 20%, and GPs were required to monitor total cholesterol for patients with QOF specified diseases, such as CVD or diabetes. These results suggest total cholesterol measurements themselves recorded in THIN were probably abnormal (total cholesterol recording more likely for patients with higher total cholesterol) if the patient had one of the QOF specified diseases, but normal if the patient was treated with a lipid modifying drug. Therefore, to ensure a plausible MAR assumption to imputing missing total cholesterol, I included the QOF specified diseases and lipid-modifying drug treatment in the imputation model.

I next considered the best approach to impute smoking status, which required additional reflection because it is a categorical variable (not normally distributed) with three categories non-smoker, ex-smoker and current smoker. In particular, the added complexity of imputing smoking status over time arises when trying to avoid impute impossible values. For example, not imputing previous current smokers as non-smokers. From discussing smoking habits with GPs, I discovered GPs recorded only a few smoking status records for non-smokers during follow-up because adult non-smokers were unlikely to start smoking so thought it unnecessary to update smoking status regularly. However, ever smokers might change smoking status so GPs update smoking status regularly for these patients; ex-smokers would start smoking again and current smokers would stop smoking. Based on these findings, I interrogated THIN to find if the data recorded agreed with the GPs observations. I found that if I assumed that patients who only ever had non-smoker smoking status recorded in THIN were non-smokers at all time, this increased the percentage of non-smokers to a similar percentage of non-smokers reported by the 2006 Health Survey for England (HSE)[77]. Applying this assumption to the data also greatly reduced the missing data and I imputed the remaining missing values to either ex-smoker or current smoker, which substantially simplifies the problem.

7.2.2 Develop methods to identify and remove outliers before imputation

The last data recording issue I considered before using the two-fold FCS algorithm was incorrectly recorded values, i.e. outliers, which are a common occurrence in electronic health records such as primary care databases. If I used MI when outliers were still in the data, they could bias the conditional distribution of the data and influence the imputation of the missing values and bias the associations between the variables. However, because THIN consists of millions of patients, some genuine extreme values may exist in the data and I did not want to lose the diversity of the data. I required a method which would distinguish between genuine extreme values and incorrectly recorded values.

Existing methods to remove outliers did not take account of repeated measures over time. Therefore, I developed a method to remove outliers in two stages. First, I identified population level outliers (values substantially outside the range of the distribution of measurements from the general population) as measurements more extreme than age and gender specific ranges from the 1998 and 2008 HSE data[76, 78]. Once the population level outliers were removed, the next stage was to identify individual level outliers (highly implausible values in the context of measurements collected over time for each patient) using a series of multilevel linear regression models with

patient specific random intercept and slope. This allowed me to identify measurements very different from the mean of all observations for that patient, after allowing for the trend in their measurements over time. Therefore, I identified measurements as outliers if standardised residuals were more extreme than a given cut-off. In total, I identified 1,643 (0.3%) height measurements as outliers after application of both methods to identify population level outliers and individual level outliers using a 'cut off' of $\pm$ 10 standardised residuals[4].

This new method to identify outliers in longitudinal data is available for other users to identify outliers in longitudinal data prior to analysis. Although I only identified a relatively small number of individual level outliers this should still give users confidence in the consistency of remaining data. Pharmacoepidemiology and Drug Safety journal published a paper describing this method[4].

7.2.3 Adapt and develop the two-fold FCS multiple imputation algorithm to impute missing values in longitudinal clinical data for variables associated with CVD

The original study, reported by Nevalainen *et al.*[1], validated the two-fold FCS algorithm using simulated data in a simple setting with measurements at three time points and up to 40% missing values. I implemented a more flexible two-fold FCS algorithm which imputes time-independent as well as time-dependent variables, increase the time window width and imputes missing data for subjects with different entry and exit time points.

I developed a Stata command `twofold`, which implements the two-fold FCS algorithm used in this study in real electronic health care data. The `twofold` command is available for other users to impute missing data in any other large, clinical, longitudinal dataset with a intermittent missingness[2].

A possible advantage of the two-fold FCS algorithm compared to baseline MI is conditioning on repeated measurements at other time blocks potentially increases the MAR assumption plausibility. Measurements at other time blocks may be associated with the probability of the values being missing at the time blocks of interest. For example, if a patient had a high total cholesterol measurement, they may be more likely to have it recorded again at the next time block because the GP will wish to take repeated measurement to monitor the high total cholesterol.

The two-fold FCS algorithm exploits the longitudinal structure, with stronger local dependencies, to simplify the imputation process: values at a given time block were imputed using only measurements at nearby time blocks (plus outcome and time-independent variables). In all settings, we must carefully consider if the simplification is reasonable for the data. For example, if an exploratory analysis found measurements further away in time provide independent information given the adjacent time blocks, we can increase the time window width.

As with standard FCS MI, the two-fold FCS algorithm is iterative, and must perform a sufficient iterations and imputations to ensure the algorithm converges to its stationary distribution. I found the two-fold FCS algorithm converged with relatively few within-time iterations (i.e. 5) and a larger among-time iterations (10-20). The number of within-time and among-time iterations did not affect convergence, i.e. estimates and standard errors,

but data imputed with more among-time iterations had correlations closer to those from the simulated data before values were made missing. However, as with standard FCS MI, to assess if enough within-time and among-time iterations were selected we can evaluate diagnostics (such as plotting means and standard deviations of the imputed data by iteration number) to empirically assess convergence and, if necessary, perform more iterations.

In this thesis I used 5 imputations, which is not uncommon for relative comparisons of methods using simulation, to achieve valid inferences (for which 5 is sufficient[50]) but more would substantially increase computational time. Some studies recommended more imputations, around 30[58, 73, 116], to reduce the impact of the random sampling inherent in MI procedures. Other studies suggest 100-200 imputations were required before the results were sufficiently accurate for critical inferences[46]. However, this depended on what was considered an acceptable loss of power. Statistical power for small effect sizes diminishes as imputations decreases, and Graham *et al.* found the rate of this power fall off was much greater than predicted by changes in relative efficiency[116]. Even if coefficient estimates are accurately imputed, this does not necessarily imply p-values are also accurate[17]. Therefore, if an accurate estimate of the p-value or fraction of missing information are required, use approximately 100 imputations[20].

FCS is only valid if an underlying joint conditional model exists[21]. Even though I did not investigate the validity of this for the two-fold FCS algorithm, I assumed that it was true because, despite a lack of theoretical justification, simulation studies showed FCS works acceptably well[46]. Investigations of direct vs. indirect imputation found similar results are expected from both[47]. However, I validated the two-fold FCS algorithm for imputing data in a longitudinal data setting using simulation studies, which builds on the previous studies, to establish the validity of this approach in a range of settings.

7.2.4 Validate the two-fold FCS algorithm through challenging simulation studies

Once I finished investigating data recording in THIN, the next aim was to evaluate the two-fold FCS algorithm using simulated data. This project is the first study to validate the two-fold FCS algorithm in longitudinal clinical data using comprehensive simulation studies directly informed by, and closely modelled on, the THIN data, a good example of a setting which would benefit from these methods.

An advantage of simulation studies is it allowed me to compare the imputed data to the original data to evaluate the two-fold FCS algorithm because the original coefficients were known from the data generation mechanism. Also, I could re-run the analysis with different among-time iterations and window width to investigate convergence. I directly developed the simulation study with a survival model using THIN data, so the results validate using the two-fold FCS algorithm for commonly used survival models using THIN. Performing a simulation study required careful programming.

I created simulations as realistic as possible to a real life epidemiological analysis. First I simulated data to mimic a THIN cohort extracted to investigate the association between health indicators recorded at a baseline time block

and a future coronary heart disease (CHD) event over 10 time blocks adjusted for age, Townsend deprivation quintile and anti-hypertensive drug treatment. I used an exponential model of interest, a time-to-event model computational tractable for the simulation study (rather than a proportional hazards model, which introduces considerable additional coding complexity).

I fitted the model of interest to the full simulated data to obtain estimates and standard errors for the data with no missing values. Next, I changed measurements in the simulated data to missing for weight, systolic blood pressure and smoking status using a MCAR mechanism because both the complete records analysis and analysis of data imputed using MI should be unbiased under a MCAR mechanism. Another benefit of a MCAR mechanism was I could directly compare the SEs for the full, complete records and analyses of data imputed using MI, if no bias exists after imputation, to see how they compared in terms of recovering information.

First, the FCS imputation model imputed missing values in all time blocks by entering measurements at each time block as separate variables in the imputation model, i.e. the 'full imputation' model. Next, I imputed the missing data using the two-fold FCS algorithm, investigated two different approaches for handling missing smoking status. For the first method, I imputing smoking status as a time-independent variable (measured only at baseline). For the second method, I imputed smoking status as time-dependent variable using the two-fold FCS algorithm, assuming all patients who only ever had non-smoker smoking status recorded were non-smokers at all time blocks and imputed any remaining missing values as either ex-smoker or current smoker. I evaluated the approaches by comparing bias and precision of estimates from fitting the model of interest to the original simulated data, complete records analysis and data imputed using a 'baseline MI' and both methods using the two-fold FCS algorithm.

The 'full imputation' model did not converge on approximately 25% of datasets because repeated measurements of the same health indicator at different time blocks included in the imputation model caused convergence problems. As explained earlier, we expected potential convergence problems due to high correlations between repeated measurements. The simulation study showed that data imputed using the two-fold FCS algorithm improved estimate precision (closer to the 'true' precision in the full data) for time-dependent variables. The improvement was greater for variables with stronger correlations between repeated measurements such as weight. The simulation study also showed improved precision for smoking status when I imputed it as a time-dependent variable compared to time-independent. This result suggest it is preferable, if possible, to use deterministic methods which apply our knowledge of the data to impute missing values before using MI. Similar methods could be applied to impute other categorical variables with missing values with a few categories, such as alcohol consumption (e.g. abstain, moderate drinker, heavy drinker etc.).

Simulated data imputed using the two-fold FCS algorithm achieved more accurate (closer to the 'true' correlations) estimated correlations between repeated measurements with more among-time iterations and for health indicators with high 'true' correlations, possibly because the higher correlations contain more information, which allows more accurate correlation estimation.

I also investigated how increasing the window width to include variables recorded at more time blocks in the imputation model affected the correlation estimates in the imputed data. This additional information achieved correlation estimates closer to the 'true' value for repeated measurements within the specified time window width, systematically underestimating correlations at greater distance. I remedied this to a limited extent by increasing the number of among-time iterations, but this did not eliminate the bias completely and worked less well if the underlying correlations were lower. In practice, therefore, consider a window width the same length as the distance of the correlations to accurately estimate.

I encountered a few issues when performing the simulation study. Firstly, the imputations did not converge if I changed 70% of values for each variable to missing at each time block because of the sparsely observed data. However, I found fewer imputation issues when I switched to a contextually more plausible missing data mechanism. I assigned 30% to have a 'consultation' and I changed all measurements to missing for those who did not consult and, for each health indicator, randomly selected 5% of patients who did consult to change values to missing. This missingness mechanism reflects real life data recording patterns in THIN and the imputation models converged because many patients had more than one health indicator recorded. Therefore, it is unlikely the two-fold FCS algorithm will not converge when imputing missing data in a primary care database like THIN because multiple health indicators measurements were probably recorded for patients who attend consultations.

I also performed a second simulation study to investigate different ways to apply the two-fold FCS algorithm to impute missing total cholesterol values in section 5.2. I used the same model of interest as the previous simulation study, but additionally including the variables total cholesterol and lipid modifying drugs, measured at the baseline time block. I included both because, as explained earlier, total cholesterol recording is different for patients prescribed lipid-modifying drugs compared to those not prescribed lipid-modifying drugs.

I changed more total cholesterol values to missing compared to the previous simulation study. With a high percentage of missing data, just including data recorded at adjacent time blocks in the two-fold FCS algorithm may not be sufficient additional information to achieve unbiased and precise estimates from fitting the model of interest to the imputed data. Therefore, I analysed data imputed using three different imputation approaches, listed below, applying the two-fold FCS algorithm which incorporates measurements at more distant time blocks, and compared the estimates and standard errors to analysing data imputed using the two-fold FCS algorithm with one time block window width. I imputed missing total cholesterol measurements using the two-fold FCS algorithm with a:

- 2 time block window width so measurements at more time blocks were included in the imputation model to inform imputation.
- 1 time block window width. Missing values at time blocks adjacent to the imputed time block were replaced with observed values at the next more extreme time blocks.
- 1 time block window width. Missing values at time blocks adjacent to the imputed time block were replaced

with the closest observed values at more extreme time blocks. Additional variables were included in the imputation model to indicate which time block the measurement was observed to account for the difference between measurements at those time blocks.

None of these methods gave significantly less bias or more precise estimates compared to using the two-fold FCS algorithm with one time block window width. This result suggests that the additional information at time blocks at more extreme time blocks added very little additional information after conditioning on adjacent time blocks, possibly because of the substantial missing data.

7.2.5 Apply the two-fold FCS algorithm in THIN

In this thesis, I investigated recording of health indicators in THIN and concluded a plausible MAR assumption for health indicators associated with CVD and used simulation studies to evaluate the best imputation approach using the two-fold FCS algorithm. Next, I implemented the two-fold FCS algorithm to impute missing values in THIN. I applied the Stata program described above to address two substantive research questions using THIN data, in which missing data could potentially bias estimates and standard errors.

In Section 6.1, I extracted data for patients actively registered to a THIN practice in the year 2000 and followed them up until a CVD event, left the practice or died. Follow-up was censored at 2009. I used an exponential survival model to find the association between health indicators recorded in the year 2000 and future CVD risk adjusted for age, sex and Townsend deprivation score quintile. First I performed a complete records analysis and compared the results to an analysing data imputed using baseline MI (imputing missing values at the baseline time block, year 2000, conditional on observed measurements at the baseline time block) and used the two-fold FCS algorithm to impute missing data for the health indicators height, weight, systolic blood pressure, total cholesterol and HDL cholesterol. I included auxiliary variables in the imputation model to reduce bias of estimates and SEs and to increase the MAR assumption plausibility. I stratified THIN by sex and practice and imputed each strata separately to make the imputations manageable and to account for interactions of sex and practice.

I was encouraged to reach the same conclusions as the simulation study. Specifically, I found that the estimates for the time-dependent health indicators in the exponential survival model were more precise when analysing data imputed using the two-fold FCS algorithm compared to complete records analysis or baseline MI, with similar estimates to other studies.

In section 6.2, I also implemented the two-fold FCS algorithm in an epidemiological study. The motivation for this study was to discover which patients with type II diabetes had a greater than average total cholesterol reduction after initiating statin treatment. For this study, I compared the results from complete records analysis, baseline MI and using the two-fold FCS algorithm in two ways: firstly, the imputation model including covariate information recorded before baseline. Secondly, the imputation model included covariate information recorded before and after baseline.

The results from the complete records analysis showed that patients with lower glomerular filtration rate (GFR) had larger total cholesterol reduction after initiating statin treatment. However, after using baseline MI (imputing missing values at baseline conditional on observed values at baseline only) more variables (male patients, white ethnicity, higher HbA_{1c} and not current smokers) were also important predictors of larger total cholesterol reduction. The extra power gained by using more information in the analysis of data imputed using baseline MI compared to the complete records analysis resulted in a change in the estimates and increased precision. The health indicators for patients with greater total cholesterol reduction indicated greater CVD risk so they were possibly more likely to adhere to statin treatment and prescribed a higher statin dose compared to patients with lower CVD risk and, therefore, smaller total cholesterol reduction. Also, these patients probably had a higher total cholesterol measurement when statins were initiated, so had more room for reduction.

Analysing data imputed using the two-fold FCS algorithm imputation method conditional on measurements before baseline had a little more power but found similar results to analysing data imputed using baseline MI for the analysis of greater total cholesterol reduction, possibly because of the substantial missing data for the health indicators at time blocks, except baseline, and the two-fold FCS algorithm can not gain much additional information from these time blocks. However, Townsend deprivation score quintile, BMI and systolic blood pressure were almost statistically significantly associated with greater total cholesterol reduction from analysing data imputed using the two-fold FCS algorithm conditional on measurements before and after baseline. The two-fold FCS algorithm could compensate for the substantial missing data at each time point by gaining more information from the additional time points after baseline. Therefore, I would still recommend using the two-fold FCS algorithm because we do not know the extent of the information in the longitudinal data in advance of the analysis, but we can explore recording frequency before using MI.

7.3 Methodological implications

This thesis is the first to extensively report the recording of health indicators in a primary care database and understand the reasons for missing data, which could be useful information for other studies using primary care databases.

This project is also the first to examine utility of the two-fold FCS algorithm on a large clinically substantive analysis of THIN data.

7.4 Applied implications

Other studies investigating MI in longitudinal databases considered simpler settings compared to the setting I investigated for this thesis. For example, one study used data with a monotone missing pattern[117], some imputed missing data at a single time point[71, 118], others used a small dataset so included measurements at all time points in the imputation model without any computational issues arising[119, 120] and one other imputed the outcome in a longitudinal cohort study[121]. Similar studies could alternatively use the two-fold FCS algorithm

to allow more complex imputations and analysis.

7.5 Conclusions

In this project, I investigated the recording of health indicators in THIN, implemented the two-fold FCS algorithm in Stata to impute missing values in longitudinal clinical data, evaluated it using simulated data and used the findings to implemented it to impute missing data in THIN. From the simulation studies, I found the two-fold FCS algorithm improved the estimate precision of the regression analysis for time-dependent variables. The precision was more accurately estimated when stronger correlations existed between the repeated measurements. I found the same results from imputing missing values in THIN using the two-fold FCS algorithm.

A few recommendations for researchers planning on using the two-fold FCS algorithm. Firstly, explore the data in detail prior to analysis to evaluate if the MAR assumption is plausible within the relevant dataset. To gain sufficient understanding of the missingness mechainism, I advise consultation with experts in the research area. These experts can possibly suggest reasons why data are missing. Next, build the imputation model so it is congenial with the model of interest. Including several auxiliary variables may increase the MAR assumption plausibility. For example, disease markers such as diabetes and respiratory infections as well as demographic variables. The two-fold FCS algorithm is more efficient compared to other approaches when imputing time-dependent variables, so include repeated measurements for as many variables as possible. If only a few auxiliary variables are available, including many time blocks will increase the MAR assumption plausibility. The two-fold FCS algorithm is also beneficial when there is substantial missing data at each time block, because it can condition on any observed measurements during follow-up. More traditional imputation methods might be preferred if the data has a small number of time points, small amount of missing data or stronger correlations exist within-time rather than among-time.

Next, I suggest a few considerations for future work involving the two-fold FCS algorithm not covered in this thesis.

7.6 Future work

In this thesis, I assumed the two-fold FCS algorithm imputations were congenial with the model of interest. However, further work is required to assess this assumptions plausibility when using the two-fold FCS algorithm to impute longitudinal data. We need further research to explore how to impute longitudinal data with interactions and non-linear effects using the two-fold FCS algorithm. In this thesis, I simplified the problem and handled interactions by dividing the data into stratum and imputed each strata separately. It may be possible to impute all strata together by combining the two-fold FCS algorithm with substantive model compatible MI to ensure compatibility of conditionals[55].

I used five imputations when imputing missing data in this project. However, many imputations may not be

necessary for the two-fold FCS algorithm because of the combination of within-time and among-time iterations used in each imputation, but further work is required to investigate this.

In clinical datasets, recording was more complete in recent years compared to earlier time periods, as shown in chapter 4.3. The two-fold FCS algorithm allows researchers to capture and make use of data recorded in more recent years, to inform and improve the precision of baseline survival models. This has particular relevance to studies using clinical data with a long lag between the exposure of interest and outcome, and researchers may wish to maximize follow-up time and select early cohort entry dates[122]. Also, the two-fold FCS algorithm may be more beneficial than a 'baseline' MI model for studies with smaller sample size, as it can obtain additional information recorded at other times.

For this study, I focused on implementing the two-fold FCS algorithm in longitudinal, electronic health records. However, the two-fold FCS algorithm can potentially be implemented in any observational study with repeated measurements. For example, survey data. Where there are gaps between recording when participants have missed a survey, potentially this information could be imputed using the two-fold FCS algorithm, which may be the best approach because it uses the recorded information when surveys were completed before and after the missing survey data.

Finally, I only investigated models of interest with explanatory variables measured at a single time block. Further work is required to investigate if the two-fold FCS algorithm achieves unbiased results for models using information at more than one time block, for example with time-updated coefficients. Potentially, the two-fold FCS algorithm could also be used to impute missing values in dataset used for developing risk prediction models that include time-updated coefficients.

Appendix A

Quality outcomes framework coronary heart disease Read code list

Read code	Description
G3...00	Ischaemic heart disease
G3...11	Arteriosclerotic heart disease
G3...12	Atherosclerotic heart disease
G3...13	IHD - Ischaemic heart disease
G30..00	Acute myocardial infarction
G30..11	Attack - heart
G30..12	Coronary thrombosis
G30..13	Cardiac rupture following myocardial infarction (MI)
G30..14	Heart attack
G30..15	MI - acute myocardial infarction
G30..16	Thrombosis - coronary
G30..17	Silent myocardial infarction
G300.00	Acute anterolateral infarction
G301.00	Other specified anterior myocardial infarction
G301000	Acute anteroapical infarction
G301100	Acute anteroseptal infarction
G301z00	Anterior myocardial infarction NOS
G302.00	Acute inferolateral infarction
G303.00	Acute inferoposterior infarction
G304.00	Posterior myocardial infarction NOS
G305.00	Lateral myocardial infarction NOS
G306.00	True posterior myocardial infarction
G307.00	Acute subendocardial infarction
G307000	Acute non-Q wave infarction
G307100	Acute non-ST segment elevation myocardial infarction
G308.00	Inferior myocardial infarction NOS
G309.00	Acute Q-wave infarct
G30A.00	Mural thrombosis
G30B.00	Acute posterolateral myocardial infarction
G30X.00	Acute transmural myocardial infarction of unspecif site
G30X000	Acute ST segment elevation myocardial infarction
G30y.00	Other acute myocardial infarction
G30y000	Acute atrial infarction
G30y100	Acute papillary muscle infarction
G30y200	Acute septal infarction
G30yz00	Other acute myocardial infarction NOS
G30z.00	Acute myocardial infarction NOS
G31..00	Other acute and subacute ischaemic heart disease
G310.00	Postmyocardial infarction syndrome
G310.11	Dressler's syndrome
G311.00	Preinfarction syndrome
G311.11	Crescendo angina
G311.12	Impending infarction
G311.13	Unstable angina
G311.14	Angina at rest
G311000	Myocardial infarction aborted

G311011 MI - myocardial infarction aborted
G311100 Unstable angina
G311200 Angina at rest
G311300 Refractory angina
G311400 Worsening angina
G311500 Acute coronary syndrome
G311z00 Preinfarction syndrome NOS
G312.00 Coronary thrombosis not resulting in myocardial infarction
G31y.00 Other acute and subacute ischaemic heart disease
G31y000 Acute coronary insufficiency
G31y100 Microinfarction of heart
G31y200 Subendocardial ischaemia
G31y300 Transient myocardial ischaemia
G31yz00 Other acute and subacute ischaemic heart disease NOS
G32..00 Old myocardial infarction
G32..11 Healed myocardial infarction
G32..12 Personal history of myocardial infarction
G33..00 Angina pectoris
G330.00 Angina decubitus
G330000 Nocturnal angina
G330z00 Angina decubitus NOS
G331.00 Prinzmetal's angina
G331.11 Variant angina pectoris
G332.00 Coronary artery spasm
G33z.00 Angina pectoris NOS
G33z000 Status anginosus
G33z100 Stenocardia
G33z200 Syncope anginosa
G33z300 Angina on effort
G33z400 Ischaemic chest pain
G33z500 Post infarct angina
G33z600 New onset angina
G33z700 Stable angina
G33zz00 Angina pectoris NOS
G34..00 Other chronic ischaemic heart disease
G340.00 Coronary atherosclerosis
G340.11 Triple vessel disease of the heart
G340.12 Coronary artery disease
G340000 Single coronary vessel disease
G340100 Double coronary vessel disease
G342.00 Atherosclerotic cardiovascular disease
G343.00 Ischaemic cardiomyopathy
G344.00 Silent myocardial ischaemia
G34y.00 Other specified chronic ischaemic heart disease
G34y000 Chronic coronary insufficiency
G34y100 Chronic myocardial ischaemia
G34yz00 Other specified chronic ischaemic heart disease NOS
G34z.00 Other chronic ischaemic heart disease NOS
G34z000 Asymptomatic coronary heart disease
G35..00 Subsequent myocardial infarction
G350.00 Subsequent myocardial infarction of anterior wall
G351.00 Subsequent myocardial infarction of inferior wall
G353.00 Subsequent myocardial infarction of other sites
G35X.00 Subsequent myocardial infarction of unspecified site
G36..00 Certain current complication follow acute myocardial infarct
G360.00 Haemopericardium/current comp folow acut myocard infarct
G361.00 Atrial septal defect/curr comp folow acut myocardal infarct
G362.00 Ventric septal defect/curr comp fol acut myocardal infarctn
G363.00 Ruptur cardiac wall w'out haemopericard/cur comp fol ac MI
G364.00 Ruptur chordae tendinae/curr comp fol acute myocard infarct
G365.00 Rupture papillary muscle/curr comp fol acute myocard infarct
G366.00 Thrombosis atrium,auric append and vent/curr comp foll acute MI
G38..00 Postoperative myocardial infarction
G380.00 Postoperative transmural myocardial infarction anterior wall
G381.00 Postoperative transmural myocardial infarction inferior wall
G382.00 Postoperative transmural myocardial infarction other sites
G383.00 Postoperative transmural myocardial infarction unspec site
G384.00 Postoperative subendocardial myocardial infarction
G38z.00 Postoperative myocardial infarction, unspecified
G3y..00 Other specified ischaemic heart disease
G3z..00 Ischaemic heart disease NOS

Appendix B

Quality outcomes framework stroke Read code list

Read code	Description
G65..00	Transient cerebral ischaemia
G65..11	Drop attack
G65..12	Transient ischaemic attack
G65..13	Vertebro-basilar insufficiency
G650.00	Basilar artery syndrome
G650.11	Insufficiency - basilar artery
G651.00	Vertebral artery syndrome
G651000	Vertebro-basilar artery syndrome
G652.00	Subclavian steal syndrome
G653.00	Carotid artery syndrome hemispheric
G654.00	Multiple and bilateral precerebral artery syndromes
G656.00	Vertebrobasilar insufficiency
G65y.00	Other transient cerebral ischaemia
G65z.00	Transient cerebral ischaemia NOS
G65z000	Impending cerebral ischaemia
G65z100	Intermittent cerebral ischaemia
G65zz00	Transient cerebral ischaemia NOS
F423600	Amaurosis fugax
G61..00	Intracerebral haemorrhage
G61..11	CVA - cerebrovascular accid due to intracerebral haemorrhage
G61..12	Stroke due to intracerebral haemorrhage
G610.00	Cortical haemorrhage
G611.00	Internal capsule haemorrhage
G612.00	Basal nucleus haemorrhage
G613.00	Cerebellar haemorrhage
G614.00	Pontine haemorrhage
G615.00	Bulbar haemorrhage
G616.00	External capsule haemorrhage
G618.00	Intracerebral haemorrhage, multiple localized
G61X.00	Intracerebral haemorrhage in hemisphere, unspecified
G61X000	Left sided intracerebral haemorrhage, unspecified
G61X100	Right sided intracerebral haemorrhage, unspecified
G61z.00	Intracerebral haemorrhage NOS
G63y000	Cerebral infarct due to thrombosis of precerebral arteries
G63y100	Cerebral infarction due to embolism of precerebral arteries
G64..00	Cerebral arterial occlusion
G64..11	CVA - cerebral artery occlusion
G64..12	Infarction - cerebral
G64..13	Stroke due to cerebral arterial occlusion
G640.00	Cerebral thrombosis
G640000	Cerebral infarction due to thrombosis of cerebral arteries
G641.00	Cerebral embolism
G641.11	Cerebral embolus
G641000	Cerebral infarction due to embolism of cerebral arteries
G64z.00	Cerebral infarction NOS
G64z.11	Brainstem infarction NOS
G64z.12	Cerebellar infarction
G64z000	Brainstem infarction
G64z100	Wallenberg syndrome
G64z111	Lateral medullary syndrome
G64z200	Left sided cerebral infarction

Appendix C

Quality outcomes framework diabetes Read code list

Read code	Description
C10E.00	Type 1 diabetes mellitus
C10E.11	Type I diabetes mellitus
C10E.12	Insulin dependent diabetes mellitus
C10E000	Type 1 diabetes mellitus with renal complications
C10E011	Type I diabetes mellitus with renal complications
C10E012	Insulin-dependent diabetes mellitus with renal complications
C10E100	Type 1 diabetes mellitus with ophthalmic complications
C10E111	Type I diabetes mellitus with ophthalmic complications
C10E112	Insulin-dependent diabetes mellitus with ophthalmic comps
C10E200	Type 1 diabetes mellitus with neurological complications
C10E211	Type I diabetes mellitus with neurological complications
C10E212	Insulin-dependent diabetes mellitus with neurological comps
C10E300	Type 1 diabetes mellitus with multiple complications
C10E311	Type I diabetes mellitus with multiple complications
C10E312	Insulin dependent diabetes mellitus with multiple complicat
C10E400	Unstable type 1 diabetes mellitus
C10E411	Unstable type I diabetes mellitus
C10E412	Unstable insulin dependent diabetes mellitus
C10E500	Type 1 diabetes mellitus with ulcer
C10E511	Type I diabetes mellitus with ulcer
C10E512	Insulin dependent diabetes mellitus with ulcer
C10E600	Type 1 diabetes mellitus with gangrene
C10E611	Type I diabetes mellitus with gangrene
C10E612	Insulin dependent diabetes mellitus with gangrene
C10E700	Type 1 diabetes mellitus with retinopathy
C10E711	Type I diabetes mellitus with retinopathy
C10E712	Insulin dependent diabetes mellitus with retinopathy
C10E800	Type 1 diabetes mellitus - poor control
C10E811	Type I diabetes mellitus - poor control
C10E812	Insulin dependent diabetes mellitus - poor control
C10E900	Type 1 diabetes mellitus maturity onset
C10E911	Type I diabetes mellitus maturity onset
C10E912	Insulin dependent diabetes maturity onset
C10EA00	Type 1 diabetes mellitus without complication
C10EA11	Type I diabetes mellitus without complication
C10EA12	Insulin-dependent diabetes without complication
C10EB00	Type 1 diabetes mellitus with mononeuropathy
C10EB11	Type I diabetes mellitus with mononeuropathy
C10EB12	Insulin dependent diabetes mellitus with mononeuropathy
C10EC00	Type 1 diabetes mellitus with polyneuropathy
C10EC11	Type I diabetes mellitus with polyneuropathy
C10EC12	Insulin dependent diabetes mellitus with polyneuropathy
C10ED00	Type 1 diabetes mellitus with nephropathy
C10ED11	Type I diabetes mellitus with nephropathy
C10ED12	Insulin dependent diabetes mellitus with nephropathy
C10EE00	Type 1 diabetes mellitus with hypoglycaemic coma
C10EE11	Type I diabetes mellitus with hypoglycaemic coma
C10EE12	Insulin dependent diabetes mellitus with hypoglycaemic coma

Read code	Description
C10EF00	Type 1 diabetes mellitus with diabetic cataract
C10EF11	Type I diabetes mellitus with diabetic cataract
C10EF12	Insulin dependent diabetes mellitus with diabetic cataract
C10EG00	Type 1 diabetes mellitus with peripheral angiopathy
C10EG11	Type I diabetes mellitus with peripheral angiopathy
C10EG12	Insulin dependent diab mell with peripheral angiopathy
C10EH00	Type 1 diabetes mellitus with arthropathy
C10EH11	Type I diabetes mellitus with arthropathy
C10EH12	Insulin dependent diabetes mellitus with arthropathy
C10EJ00	Type 1 diabetes mellitus with neuropathic arthropathy
C10EJ11	Type I diabetes mellitus with neuropathic arthropathy
C10EJ12	Insulin dependent diab mell with neuropathic arthropathy
C10EK00	Type 1 diabetes mellitus with persistent proteinuria
C10EK11	Type I diabetes mellitus with persistent proteinuria
C10EL00	Type 1 diabetes mellitus with persistent microalbuminuria
C10EL11	Type I diabetes mellitus with persistent microalbuminuria
C10EM00	Type 1 diabetes mellitus with ketoacidosis
C10EM11	Type I diabetes mellitus with ketoacidosis
C10EN00	Type 1 diabetes mellitus with ketoacidotic coma
C10EN11	Type I diabetes mellitus with ketoacidotic coma
C10EP00	Type 1 diabetes mellitus with exudative maculopathy
C10EP11	Type I diabetes mellitus with exudative maculopathy
C10EQ00	Type 1 diabetes mellitus with gastroparesis
C10ER00	Latent autoimmune diabetes mellitus in adult
C10F.00	Type 2 diabetes mellitus
C10F.11	Type II diabetes mellitus
C10F000	Type 2 diabetes mellitus with renal complications
C10F011	Type II diabetes mellitus with renal complications
C10F100	Type 2 diabetes mellitus with ophthalmic complications
C10F111	Type II diabetes mellitus with ophthalmic complications
C10F200	Type 2 diabetes mellitus with neurological complications
C10F211	Type II diabetes mellitus with neurological complications
C10F300	Type 2 diabetes mellitus with multiple complications
C10F311	Type II diabetes mellitus with multiple complications
C10F400	Type 2 diabetes mellitus with ulcer
C10F411	Type II diabetes mellitus with ulcer
C10F500	Type 2 diabetes mellitus with gangrene
C10F511	Type II diabetes mellitus with gangrene
C10F600	Type 2 diabetes mellitus with retinopathy
C10F611	Type II diabetes mellitus with retinopathy
C10F700	Type 2 diabetes mellitus - poor control
C10F711	Type II diabetes mellitus - poor control
C10F900	Type 2 diabetes mellitus without complication
C10F911	Type II diabetes mellitus without complication
C10FA00	Type 2 diabetes mellitus with mononeuropathy
C10FA11	Type II diabetes mellitus with mononeuropathy
C10FB00	Type 2 diabetes mellitus with polyneuropathy
C10FB11	Type II diabetes mellitus with polyneuropathy
C10FC00	Type 2 diabetes mellitus with nephropathy
C10FC11	Type II diabetes mellitus with nephropathy
C10FD00	Type 2 diabetes mellitus with hypoglycaemic coma
C10FD11	Type II diabetes mellitus with hypoglycaemic coma
C10FE00	Type 2 diabetes mellitus with diabetic cataract
C10FE11	Type II diabetes mellitus with diabetic cataract
C10FF00	Type 2 diabetes mellitus with peripheral angiopathy
C10FF11	Type II diabetes mellitus with peripheral angiopathy
C10FG00	Type 2 diabetes mellitus with arthropathy
C10FG11	Type II diabetes mellitus with arthropathy
C10FH00	Type 2 diabetes mellitus with neuropathic arthropathy
C10FH11	Type II diabetes mellitus with neuropathic arthropathy
C10FJ00	Insulin treated Type 2 diabetes mellitus
C10FJ11	Insulin treated Type II diabetes mellitus
C10FK00	Hyperosmolar non-ketotic state in type 2 diabetes mellitus
C10FL00	Type 2 diabetes mellitus with persistent proteinuria
C10FL11	Type II diabetes mellitus with persistent proteinuria
C10FM00	Type 2 diabetes mellitus with persistent microalbuminuria
C10FM11	Type II diabetes mellitus with persistent microalbuminuria
C10FN00	Type 2 diabetes mellitus with ketoacidosis
C10FN11	Type II diabetes mellitus with ketoacidosis
C10FP00	Type 2 diabetes mellitus with ketoacidotic coma
C10FP11	Type II diabetes mellitus with ketoacidotic coma
C10FQ00	Type 2 diabetes mellitus with exudative maculopathy
C10FQ11	Type II diabetes mellitus with exudative maculopathy
C10FR00	Type 2 diabetes mellitus with gastroparesis
C10FS00	Maternally inherited diabetes mellitus

Appendix D

Quality outcomes framework coronary obstructive pulmonary disease Read code list

Read code	Description
H3...00	Chronic obstructive pulmonary disease
H3...11	Chronic obstructive airways disease
H31..00	Chronic bronchitis
H310.00	Simple chronic bronchitis
H310000	Chronic catarrhal bronchitis
H310z00	Simple chronic bronchitis NOS
H311.00	Mucopurulent chronic bronchitis
H311000	Purulent chronic bronchitis
H311100	Fetid chronic bronchitis
H311z00	Mucopurulent chronic bronchitis NOS
H312.00	Obstructive chronic bronchitis
H312000	Chronic asthmatic bronchitis
H312011	Chronic wheezy bronchitis
H312100	Emphysematous bronchitis
H312300	Bronchiolitis obliterans
H312z00	Obstructive chronic bronchitis NOS
H313.00	Mixed simple and mucopurulent chronic bronchitis
H31y.00	Other chronic bronchitis
H31y100	Chronic tracheobronchitis
H31yz00	Other chronic bronchitis NOS
H31z.00	Chronic bronchitis NOS
H32..00	Emphysema
H320.00	Chronic bullous emphysema
H320000	Segmental bullous emphysema
H320100	Zonal bullous emphysema
H320200	Giant bullous emphysema
H320300	Bullous emphysema with collapse
H320311	Tension pneumatocoele
H320z00	Chronic bullous emphysema NOS
H321.00	Panlobular emphysema
H322.00	Centrilobular emphysema
H32y.00	Other emphysema
H32y000	Acute vesicular emphysema
H32y100	Atrophic (senile) emphysema
H32y111	Acute interstitial emphysema
H32y200	MacLeod's unilateral emphysema
H32yz00	Other emphysema NOS
H32yz11	Sawyer - Jones syndrome
H32z.00	Emphysema NOS
H36..00	Mild chronic obstructive pulmonary disease
H37..00	Moderate chronic obstructive pulmonary disease
H38..00	Severe chronic obstructive pulmonary disease
H39..00	Very severe chronic obstructive pulmonary disease
H3y..00	Other specified chronic obstructive airways disease
H3y..11	Other specified chronic obstructive pulmonary disease
H3y0.00	Chronic obstruct pulmonary dis with acute lower resp infectn
H3y1.00	Chron obstruct pulmonary dis wth acute exacerbation, unspec
H3z..00	Chronic obstructive airways disease NOS
H3z..11	Chronic obstructive pulmonary disease NOS

Appendix E

Longitudinal recording - additional figures

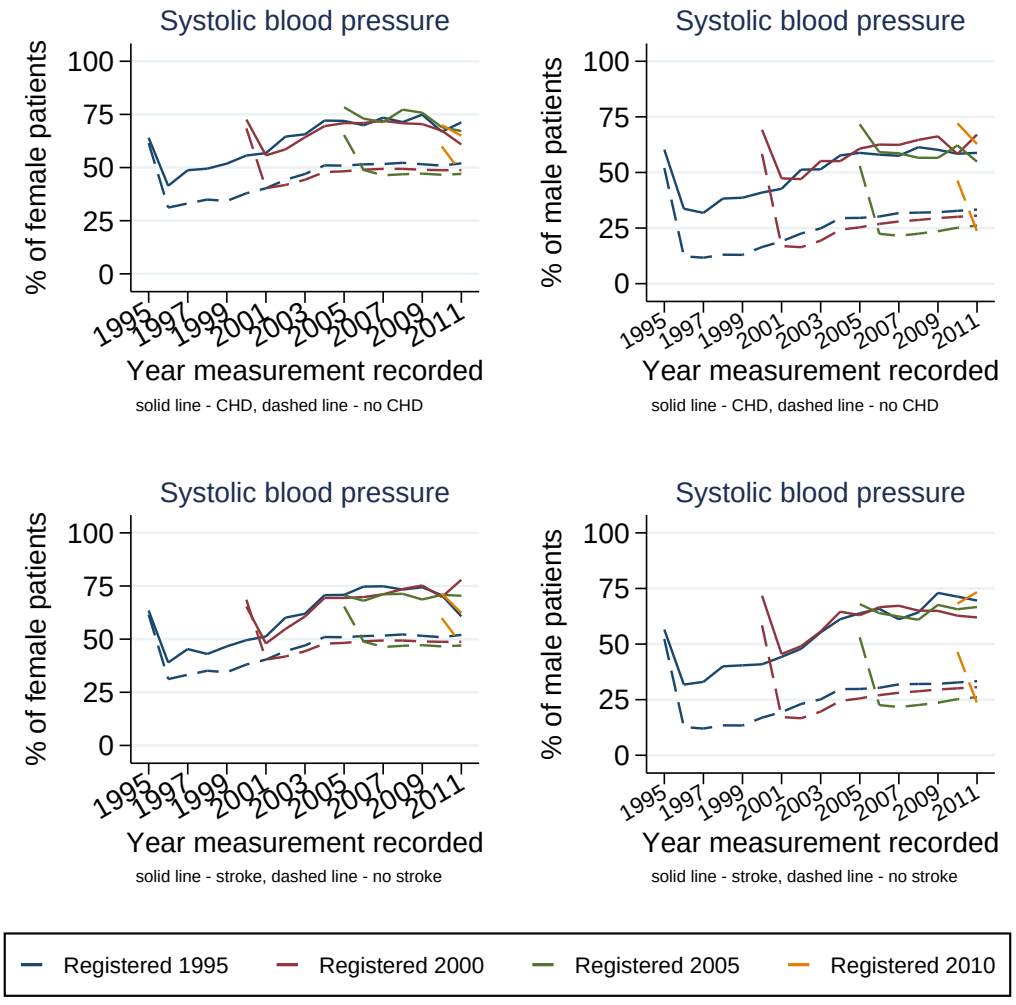

Figure E.1: Percentage of female patients (left) and male patients (right) in each cohort with and without coronary heart disease (CHD, top) or stroke (bottom) and at least one systolic blood pressure measurement recorded each year during follow-up.

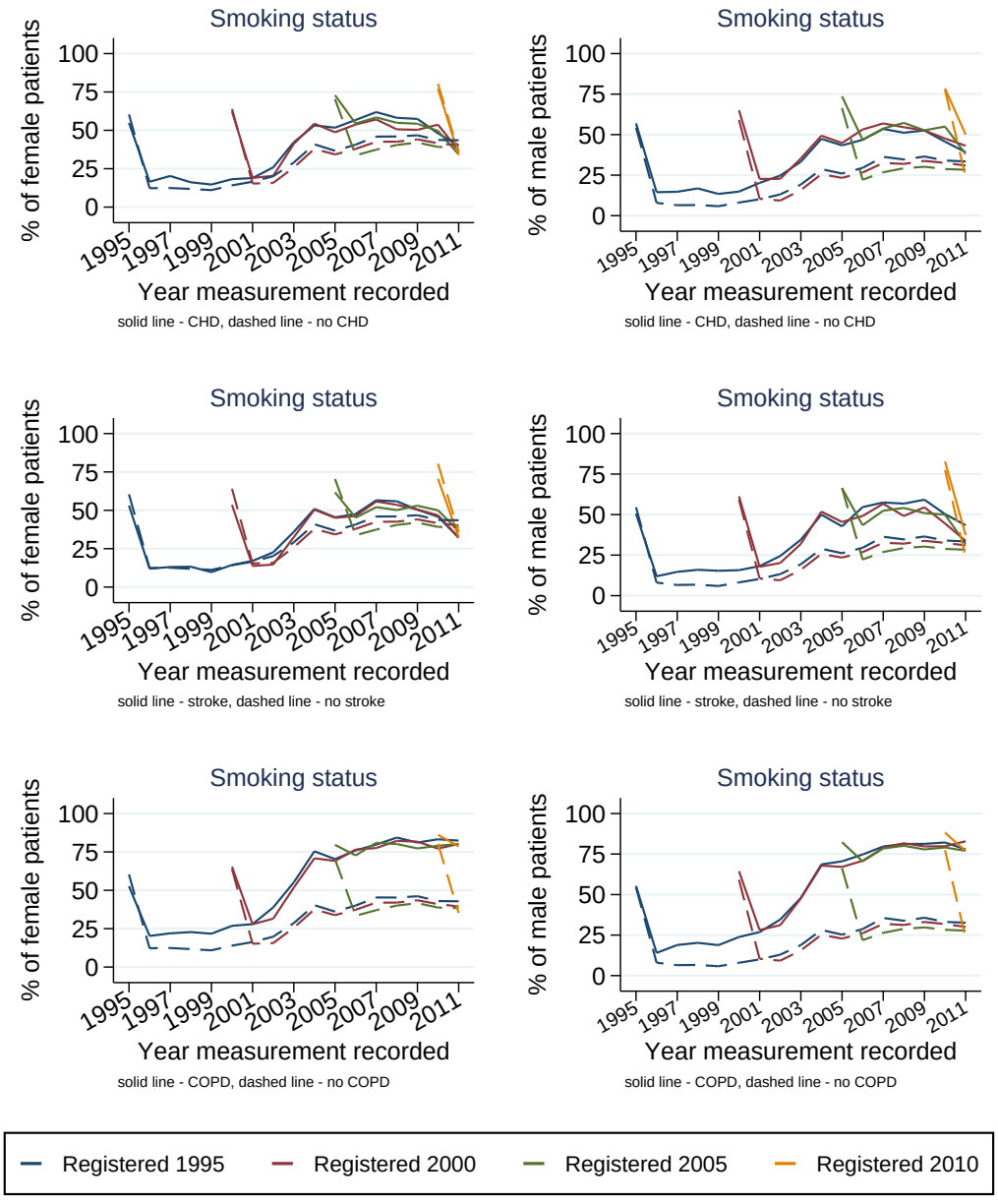

Figure E.2: Percentage of female patients (left) and male patients (right) in each cohort with and without coronary heart disease (CHD, top), stroke (middle) or coronary heart disease (bottom) and at least one smoking status recorded each year during follow-up.

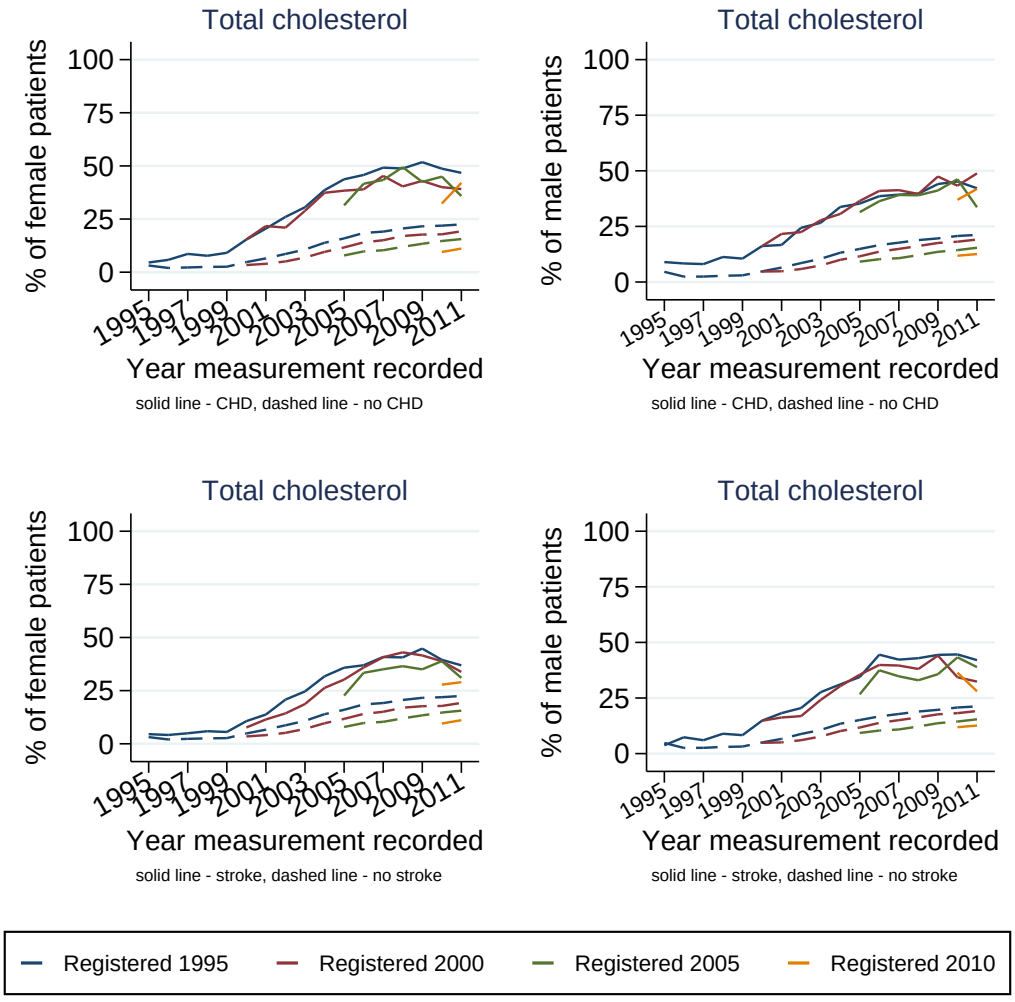

Figure E.3: Percentage of female patients (left) and male patients (right) in each cohort with and without coronary heart disease (CHD, top) or stroke (bottom) and at least one total cholesterol measurement recorded each year during follow-up.

Appendix F

Outliers results

Table F.1: Maximum +10% and minimum -10% values of weight, systolic blood pressure, diastolic blood pressure, total serum cholesterol and HDL cholesterol found in Health Survey for England data from 1998 and 2008 by age and sex

Health indicator	**Age range (years)**	**Male**		**Female**	
		Min-10%	**Max+10%**	**Min-10%**	**Max+10%**
Weight (kg)	18 - 24	36.1	141.5	31.8	142.5
	25 - 34	41.0	143.0	32.7	143.0
	35 - 44	44.9	143.0	28.4	143.0
	45 - 54	40.1	143.0	33.3	143.0
	55 - 64	34.6	142.9	31.7	143.0
	65 - 74	34.8	143.0	27.5	141.7
	75+	36.0	131.7	29.7	137.3
Systolic blood pressure (mmHg)	18 - 24	68.4	207.9	52.2	194.7
	25 - 34	61.2	214.5	72.0	221.1
	35 - 44	55.8	218.9	64.8	235.4
	45 - 54	55.8	242.0	54.9	269.5
	55 - 64	63.0	256.3	72.0	256.3
	65 - 74	72.0	253.0	54.9	255.2
	75+	78.3	249.7	61.2	265.1
Diastolic blood pressure (mmHg)	18 - 24	29.7	128.7	27.9	127.6
	25 - 34	36.0	139.7	30.6	125.4
	35 - 44	29.7	141.9	35.1	144.1
	45 - 54	27.9	159.5	27.9	163.9
	55 - 64	28.8	157.3	35.1	167.2
	65 - 74	35.1	154.0	34.2	165.0
	75+	33.3	158.4	34.2	161.7
Total serum cholesterol ($\mathrm{mmol\,l^{-1}}$)	18 - 24	1.8	8.9	2.5	9.1
	25 - 34	2.0	13.1	2.4	11.8
	35 - 44	2.5	13.0	2.3	10.0
	45 - 54	2.4	12.3	2.8	11.3
	55 - 64	2.9	16.8	2.8	11.6
	65 - 74	1.9	12.3	2.6	13.2
	75+	1.8	9.2	2.3	12.4
HDL cholesterol ($\mathrm{mmol\,l^{-1}}$)	18 - 24	0.4	3.3	0.5	3.2
	25 - 34	0.4	3.2	0.5	3.7
	35 - 44	0.5	3.3	0.5	4.1
	45 - 54	0.3	3.6	0.7	4.7
	55 - 64	0.5	3.4	0.3	3.7
	65 - 74	0.4	3.4	0.5	3.9
	75+	0.3	3.9	0.6	3.9

Table F.2: Details of numbers of outliers identified using the two stage method described

	Weight[1]	**Systolic blood pressure**[2]	**Diastolic blood pressure**[2]	**Total serum cholesterol**[2]	**HDL cholesterol**[3]
Patients included[4]	552,426	591,877	592,369	663,637	717,923
Number of patients with at least one measurement	333,352	374,669	374,643	133,680	100,573
Number of measurements	1,144,978	2,939,313	2,934,031	512,137	312,613
Measurements excluded because outside range of acceptable values defined in previous table	7,162	923	802	1,910	846
Number of outliers identified after fitting random effects model[5] 3 times	105	81	152	63	123

[1] Measurements with standardised residuals more extreme than ± 10 were identified as outliers, or the most extreme of two measurements was identified as an outlier when at least one of the measurement was more extreme than ± 8

[2] Measurements with standardised residuals more extreme than ± 6 were identified as outliers, or the most extreme of two measurements was identified as an outlier when at least one of the measurement was more extreme than ± 5

[3] Measurements with standardised residuals more extreme than ± 7 were identified as outliers, or the most extreme of two measurements was identified as an outlier when at least one of the measurement was more extreme than ± 6

[4] Including all patients with measurements recorded from the date of patient registration with study practices or after the practice had acceptable levels of data recording (good quality records and acceptable mortality rates) or the patient was aged 16 years or older until they were transferred out of the practice, died or before the last date that the practice submitted data. This included data from the time period 1995 to 2009. Measurements were excluded if they were recorded as 0 or missing a measurement date

[5] With random intercept and random slope

Appendix G

Two-stage method to remove population- and individual-level outliers from longitudinal data in a primary care database - article

Appendix H

Evaluation of two-fold fully conditional specification multiple imputation for longitudinal electronic health record data - article

Appendix I

Coronary heart disease Read code list

Read code	Description
14AL.00	H/O: Treatment for ischaemic heart disease
792..00	Coronary artery operations
792..11	Coronary artery bypass graft operations
7920.00	Saphenous vein graft replacement of coronary artery
7920.11	Saphenous vein graft bypass of coronary artery
7920000	Saphenous vein graft replacement of one coronary artery
7920100	Saphenous vein graft replacement of two coronary arteries
7920200	Saphenous vein graft replacement of three coronary arteries
7920300	Saphenous vein graft replacement of four+ coronary arteries
7920y00	Saphenous vein graft replacement of coronary artery OS
7920z00	Saphenous vein graft replacement coronary artery NOS
7921.00	Other autograft replacement of coronary artery
7921.11	Other autograft bypass of coronary artery
7921000	Autograft replacement of one coronary artery NEC
7921100	Autograft replacement of two coronary arteries NEC
7921200	Autograft replacement of three coronary arteries NEC
7921300	Autograft replacement of four of more coronary arteries NEC
7921y00	Other autograft replacement of coronary artery OS
7921z00	Other autograft replacement of coronary artery NOS
7922.00	Allograft replacement of coronary artery
7922.11	Allograft bypass of coronary artery
7922000	Allograft replacement of one coronary artery
7922100	Allograft replacement of two coronary arteries
7922200	Allograft replacement of three coronary arteries
7922300	Allograft replacement of four or more coronary arteries
7922y00	Other specified allograft replacement of coronary artery
7922z00	Allograft replacement of coronary artery NOS
7923.00	Prosthetic replacement of coronary artery
7923.11	Prosthetic bypass of coronary artery
7923000	Prosthetic replacement of one coronary artery
7923100	Prosthetic replacement of two coronary arteries
7923200	Prosthetic replacement of three coronary arteries
7923300	Prosthetic replacement of four or more coronary arteries
7923y00	Other specified prosthetic replacement of coronary artery
7923z00	Prosthetic replacement of coronary artery NOS
7924.00	Revision of bypass for coronary artery
7924000	Revision of bypass for one coronary artery
7924100	Revision of bypass for two coronary arteries
7924200	Revision of bypass for three coronary arteries
7924300	Revision of bypass for four or more coronary arteries
7924400	Revision of connection of thoracic artery to coronary artery
7924500	Revision of implantation of thoracic artery into heart
7924y00	Other specified revision of bypass for coronary artery
7924z00	Revision of bypass for coronary artery NOS
7925.00	Connection of mammary artery to coronary artery
7925.11	Creation of bypass from mammary artery to coronary artery
7925000	Double anastomosis of mammary arteries to coronary arteries
7925011	LIMA sequential anastomosis
7925012	RIMA sequential anastomosis
7925100	Double implant of mammary arteries into coronary arteries
7925200	Single anast mammary art to left ant descend coronary art

Read code	Description
7925300	Single anastomosis of mammary artery to coronary artery NEC
7925311	LIMA single anastomosis
7925312	RIMA single anastomosis
7925400	Single implantation of mammary artery into coronary artery
7925y00	Connection of mammary artery to coronary artery OS
7925z00	Connection of mammary artery to coronary artery NOS
7926.00	Connection of other thoracic artery to coronary artery
7926000	Double anastom thoracic arteries to coronary arteries NEC
7926100	Double implant thoracic arteries into coronary arteries NEC
7926200	Single anastomosis of thoracic artery to coronary artery NEC
7926300	Single implantation thoracic artery into coronary artery NEC
7926y00	Connection of other thoracic artery to coronary artery OS
7926z00	Connection of other thoracic artery to coronary artery NOS
7927300	Transposition of coronary artery NEC
7927400	Exploration of coronary artery
7927500	Open angioplasty of coronary artery
7927y00	Other specified other open operation on coronary artery
7927z00	Other open operation on coronary artery NOS
7928.00	Transluminal balloon angioplasty of coronary artery
7928.11	Percutaneous balloon coronary angioplasty
7928000	Percut transluminal balloon angioplasty one coronary artery
7928100	Percut translum balloon angioplasty mult coronary arteries
7928200	Percut translum balloon angioplasty bypass graft coronary a
7928300	Percut translum cutting balloon angioplasty coronary artery
7928y00	Transluminal balloon angioplasty of coronary artery OS
7928z00	Transluminal balloon angioplasty of coronary artery NOS
7929.00	Other therapeutic transluminal operations on coronary artery
7929000	Percutaneous transluminal laser coronary angioplasty
7929100	Percut transluminal coronary thrombolysis with streptokinase
7929111	Percut translum coronary thrombolytic therapy- streptokinase
7929200	Percut translum inject therap subst to coronary artery NEC
7929300	Rotary blade coronary angioplasty
7929400	Insertion of coronary artery stent
7929500	Insertion of drug-eluting coronary artery stent
7929600	Percutaneous transluminal atherectomy of coronary artery
7929y00	Other therapeutic transluminal op on coronary artery OS
7929z00	Other therapeutic transluminal op on coronary artery NOS
792A.00	Diagnostic transluminal operations on coronary artery
792Ay00	Diagnostic transluminal operation on coronary artery OS
792Az00	Diagnostic transluminal operation on coronary artery NOS
792B.00	Repair of coronary artery NEC
792B000	Endarterectomy of coronary artery NEC
792By00	Other specified repair of coronary artery
792Bz00	Repair of coronary artery NOS
792C.00	Other replacement of coronary artery
792C000	Replacement of coronary arteries using multiple methods
792Cy00	Other specified replacement of coronary artery
792Cz00	Replacement of coronary artery NOS
792D.00	Other bypass of coronary artery
792Dy00	Other specified other bypass of coronary artery
792Dz00	Other bypass of coronary artery NOS
792y.00	Other specified operations on coronary artery
792z.00	Coronary artery operations NOS
793G.00	Perc translumin balloon angioplasty stenting coronary artery
793G000	Perc translum ball angio insert 1-2 drug elut stents cor art
793G100	Perc tran ball angio ins 3 or more drug elut stents cor art
793G200	Perc translum balloon angioplasty insert 1-2 stents cor art
793G300	Percutaneous cor balloon angiop 3 more stents cor art NEC
793Gy00	OS perc translumina balloon angioplast stenting coronary art
793Gz00	Perc translum balloon angioplasty stenting coronary art NOS
793H.00	Transluminal operations on cardiac conduit
793H000	Percutaneous transluminal balloon dilation cardiac conduit
7A54500	Rotary blade angioplasty
7A6G100	Peroperative angioplasty
7A6H300	Prosthetic graft patch angioplasty
7A6H400	Percutaneous transluminal angioplasty of vascular graft
G3...00	Ischaemic heart disease
G3...11	Arteriosclerotic heart disease
G3...12	Atherosclerotic heart disease
G3...13	IHD - Ischaemic heart disease
G30..00	Acute myocardial infarction

Read code	Description
G30..11	Attack - heart
G30..12	Coronary thrombosis
G30..13	Cardiac rupture following myocardial infarction (MI)
G30..14	Heart attack
G30..15	MI - acute myocardial infarction
G30..16	Thrombosis - coronary
G30..17	Silent myocardial infarction
G300.00	Acute anterolateral infarction
G301.00	Other specified anterior myocardial infarction
G301000	Acute anteroapical infarction
G301100	Acute anteroseptal infarction
G301z00	Anterior myocardial infarction NOS
G302.00	Acute inferolateral infarction
G303.00	Acute inferoposterior infarction
G304.00	Posterior myocardial infarction NOS
G305.00	Lateral myocardial infarction NOS
G306.00	True posterior myocardial infarction
G307.00	Acute subendocardial infarction
G307000	Acute non-Q wave infarction
G307100	Acute non-ST segment elevation myocardial infarction
G308.00	Inferior myocardial infarction NOS
G309.00	Acute Q-wave infarct
G30A.00	Mural thrombosis
G30B.00	Acute posterolateral myocardial infarction
G30X.00	Acute transmural myocardial infarction of unspecif site
G30X000	Acute ST segment elevation myocardial infarction
G30y.00	Other acute myocardial infarction
G30y000	Acute atrial infarction
G30y100	Acute papillary muscle infarction
G30y200	Acute septal infarction
G30yz00	Other acute myocardial infarction NOS
G30z.00	Acute myocardial infarction NOS
G31..00	Other acute and subacute ischaemic heart disease
G310.00	Postmyocardial infarction syndrome
G310.11	Dressler's syndrome
G311.00	Preinfarction syndrome
G311.11	Crescendo angina
G311.12	Impending infarction
G311.13	Unstable angina
G311.14	Angina at rest
G311000	Myocardial infarction aborted
G311011	MI - myocardial infarction aborted
G311100	Unstable angina
G311200	Angina at rest
G311300	Refractory angina
G311400	Worsening angina
G311500	Acute coronary syndrome
G311z00	Preinfarction syndrome NOS
G312.00	Coronary thrombosis not resulting in myocardial infarction
G31y.00	Other acute and subacute ischaemic heart disease
G31y000	Acute coronary insufficiency
G31y100	Microinfarction of heart
G31y200	Subendocardial ischaemia
G31y300	Transient myocardial ischaemia
G31yz00	Other acute and subacute ischaemic heart disease NOS
G32..00	Old myocardial infarction
G32..11	Healed myocardial infarction
G32..12	Personal history of myocardial infarction
G33..00	Angina pectoris
G330.00	Angina decubitus
G330000	Nocturnal angina
G330z00	Angina decubitus NOS
G331.00	Prinzmetal's angina
G331.11	Variant angina pectoris
G332.00	Coronary artery spasm
G33z.00	Angina pectoris NOS
G33z000	Status anginosus
G33z100	Stenocardia
G33z200	Syncope anginosa
G33z300	Angina on effort
G33z400	Ischaemic chest pain

Read code	Description
G33z500	Post infarct angina
G33z600	New onset angina
G33z700	Stable angina
G33zz00	Angina pectoris NOS
G34..00	Other chronic ischaemic heart disease
G340.00	Coronary atherosclerosis
G340.11	Triple vessel disease of the heart
G340.12	Coronary artery disease
G340000	Single coronary vessel disease
G340100	Double coronary vessel disease
G342.00	Atherosclerotic cardiovascular disease
G343.00	Ischaemic cardiomyopathy
G344.00	Silent myocardial ischaemia
G34y.00	Other specified chronic ischaemic heart disease
G34y000	Chronic coronary insufficiency
G34y100	Chronic myocardial ischaemia
G34yz00	Other specified chronic ischaemic heart disease NOS
G34z.00	Other chronic ischaemic heart disease NOS
G34z000	Asymptomatic coronary heart disease
G35..00	Subsequent myocardial infarction
G350.00	Subsequent myocardial infarction of anterior wall
G351.00	Subsequent myocardial infarction of inferior wall
G353.00	Subsequent myocardial infarction of other sites
G35X.00	Subsequent myocardial infarction of unspecified site
G36..00	Certain current complication follow acute myocardial infarct
G360.00	Haemopericardium/current comp folow acut myocard infarct
G361.00	Atrial septal defect/curr comp folow acut myocardal infarct
G362.00	Ventric septal defect/curr comp fol acut myocardal infarctn
G363.00	Ruptur cardiac wall w'out haemopericard/cur comp fol ac MI
G364.00	Ruptur chordae tendinae/curr comp fol acute myocard infarct
G365.00	Rupture papillary muscle/curr comp fol acute myocard infarct
G366.00	Thrombosis atrium,auric append and vent/curr comp foll acute MI
G38..00	Postoperative myocardial infarction
G380.00	Postoperative transmural myocardial infarction anterior wall
G381.00	Postoperative transmural myocardial infarction inferior wall
G382.00	Postoperative transmural myocardial infarction other sites
G383.00	Postoperative transmural myocardial infarction unspec site
G384.00	Postoperative subendocardial myocardial infarction
G38z.00	Postoperative myocardial infarction, unspecified
G3y..00	Other specified ischaemic heart disease
G3z..00	Ischaemic heart disease NOS
G501.00	Post infarction pericarditis
Gyu3.00	[X]Ischaemic heart diseases
Gyu3100	[X]Other current complicatns following acute myocard infarct
Gyu3200	[X]Other forms of acute ischaemic heart disease
Gyu3300	[X]Other forms of chronic ischaemic heart disease
Gyu3400	[X]Acute transmural myocardial infarction of unspecif site
Gyu3500	[X]Subsequent myocardial infarction of other sites
Gyu3600	[X]Subsequent myocardial infarction of unspecified site
ZV45700	[V]Presence of aortocoronary bypass graft
ZV45800	[V]Presence of coronary angioplasty implant and graft
ZV45K00	[V]Presence of coronary artery bypass graft
ZV45K11	[V]Presence of coronary artery bypass graft - CABG
ZV45L00	[V]Status following coronary angioplasty NOS

Appendix J

Application of multiple imputation using the two-fold fully conditional specification algorithm - article

Appendix K

Quality outcomes framework schizophrenia, bipolar disorder and other psychoses Read code list

Read code	Description
E10..00	Schizophrenic disorders
E100.00	Simple schizophrenia
E100.11	Schizophrenia simplex
E100000	Unspecified schizophrenia
E100100	Subchronic schizophrenia
E100200	Chronic schizophrenic
E100300	Acute exacerbation of subchronic schizophrenia
E100400	Acute exacerbation of chronic schizophrenia
E100500	Schizophrenia in remission
E100z00	Simple schizophrenia NOS
E101.00	Hebephrenic schizophrenia
E101000	Unspecified hebephrenic schizophrenia
E101100	Subchronic hebephrenic schizophrenia
E101200	Chronic hebephrenic schizophrenia
E101300	Acute exacerbation of subchronic hebephrenic schizophrenia
E101400	Acute exacerbation of chronic hebephrenic schizophrenia
E101500	Hebephrenic schizophrenia in remission
E101z00	Hebephrenic schizophrenia NOS
E102.00	Catatonic schizophrenia
E102000	Unspecified catatonic schizophrenia
E102100	Subchronic catatonic schizophrenia
E102200	Chronic catatonic schizophrenia
E102300	Acute exacerbation of subchronic catatonic schizophrenia
E102400	Acute exacerbation of chronic catatonic schizophrenia
E102500	Catatonic schizophrenia in remission
E102z00	Catatonic schizophrenia NOS
E103.00	Paranoid schizophrenia
E103000	Unspecified paranoid schizophrenia
E103100	Subchronic paranoid schizophrenia
E103200	Chronic paranoid schizophrenia
E103300	Acute exacerbation of subchronic paranoid schizophrenia
E103400	Acute exacerbation of chronic paranoid schizophrenia
E103500	Paranoid schizophrenia in remission
E103z00	Paranoid schizophrenia NOS
E104.00	Acute schizophrenic episode
E104.11	Oneirophrenia
E105.00	Latent schizophrenia
E105000	Unspecified latent schizophrenia
E105100	Subchronic latent schizophrenia
E105200	Chronic latent schizophrenia
E105300	Acute exacerbation of subchronic latent schizophrenia
E105400	Acute exacerbation of chronic latent schizophrenia
E105500	Latent schizophrenia in remission

Read code	Description
E105z00	Latent schizophrenia NOS
E106.00	Residual schizophrenia
E106.11	Restzustand - schizophrenia
E107.00	Schizo-affective schizophrenia
E107.11	Cyclic schizophrenia
E107000	Unspecified schizo-affective schizophrenia
E107100	Subchronic schizo-affective schizophrenia
E107200	Chronic schizo-affective schizophrenia
E107300	Acute exacerbation subchronic schizo-affective schizophrenia
E107400	Acute exacerbation of chronic schizo-affective schizophrenia
E107500	Schizo-affective schizophrenia in remission
E107z00	Schizo-affective schizophrenia NOS
E10y.00	Other schizophrenia
E10y.11	Cenesthopathic schizophrenia
E10y000	Atypical schizophrenia
E10y100	Coenesthopathic schizophrenia
E10yz00	Other schizophrenia NOS
E10z.00	Schizophrenia NOS
Eu2..00	[X]Schizophrenia, schizotypal and delusional disorders
Eu20.00	[X]Schizophrenia
Eu20000	[X]Paranoid schizophrenia
Eu20011	[X]Paraphrenic schizophrenia
Eu20100	[X]Hebephrenic schizophrenia
Eu20111	[X]Disorganised schizophrenia
Eu20200	[X]Catatonic schizophrenia
Eu20211	[X]Catatonic stupor
Eu20212	[X]Schizophrenic catalepsy
Eu20213	[X]Schizophrenic catatonia
Eu20214	[X]Schizophrenic flexibilatis cerea
Eu20300	[X]Undifferentiated schizophrenia
Eu20311	[X]Atypical schizophrenia
Eu20400	[X]Post-schizophrenic depression
Eu20500	[X]Residual schizophrenia
Eu20511	[X]Chronic undifferentiated schizophrenia
Eu20512	[X]Restzustand schizophrenic
Eu20600	[X]Simple schizophrenia
Eu20y00	[X]Other schizophrenia
Eu20y11	[X]Cenesthopathic schizophrenia
Eu20y12	[X]Schizophreniform disord NOS
Eu20y13	[X]Schizophrenifrm psychos NOS
Eu20z00	[X]Schizophrenia, unspecified

Appendix L

Quality outcomes framework asthma Read code list

Read code	Description
H33..00	Asthma
H33..11	Bronchial asthma
H330.00	Extrinsic (atopic) asthma
H330.11	Allergic asthma
H330.12	Childhood asthma
H330.13	Hay fever with asthma
H330.14	Pollen asthma
H330000	Extrinsic asthma without status asthmaticus
H330011	Hay fever with asthma
H330100	Extrinsic asthma with status asthmaticus
H330111	Extrinsic asthma with asthma attack
H330z00	Extrinsic asthma NOS
H331.00	Intrinsic asthma
H331.11	Late onset asthma
H331000	Intrinsic asthma without status asthmaticus
H331100	Intrinsic asthma with status asthmaticus
H331111	Intrinsic asthma with asthma attack
H331z00	Intrinsic asthma NOS
H332.00	Mixed asthma
H333.00	Acute exacerbation of asthma
H334.00	Brittle asthma
H33z.00	Asthma unspecified
H33z.11	Hyperreactive airways disease
H33z000	Status asthmaticus NOS
H33z011	Severe asthma attack
H33z100	Asthma attack
H33z111	Asthma attack NOS
H33z200	Late-onset asthma
H33zz00	Asthma NOS
H33zz11	Exercise induced asthma
H33zz12	Allergic asthma NEC
H33zz13	Allergic bronchitis NEC

Appendix M

Quality outcomes framework chronic kidney disease Read code list

Read code	Description
1Z12.00	Chronic kidney disease stage 3
1Z13.00	Chronic kidney disease stage 4
1Z14.00	Chronic kidney disease stage 5
1Z15.00	Chronic kidney disease stage 3A
1Z16.00	Chronic kidney disease stage 3B
1Z1B.00	Chronic kidney disease stage 3 with proteinuria
1Z1B.11	CKD stage 3 with proteinuria
1Z1C.00	Chronic kidney disease stage 3 without proteinuria
1Z1C.11	CKD stage 3 without proteinuria
1Z1D.00	Chronic kidney disease stage 3A with proteinuria
1Z1D.11	CKD stage 3A with proteinuria
1Z1E.00	Chronic kidney disease stage 3A without proteinuria
1Z1E.11	CKD stage 3A without proteinuria
1Z1F.00	Chronic kidney disease stage 3B with proteinuria
1Z1F.11	CKD stage 3B with proteinuria
1Z1G.00	Chronic kidney disease stage 3B without proteinuria
1Z1G.11	CKD stage 3B without proteinuria
1Z1H.00	Chronic kidney disease stage 4 with proteinuria
1Z1H.11	CKD stage 4 with proteinuria
1Z1J.00	Chronic kidney disease stage 4 without proteinuria
1Z1J.11	CKD stage 4 without proteinuria
1Z1K.00	Chronic kidney disease stage 5 with proteinuria
1Z1K.11	CKD stage 5 with proteinuria
1Z1L.00	Chronic kidney disease stage 5 without proteinuria
1Z1L.11	CKD stage 5 without proteinuria

Appendix N

Quality outcomes framework atrial fibrillation

Read code list

Read code	Description
G573.00	Atrial fibrillation and flutter
G573000	Atrial fibrillation
G573200	Paroxysmal atrial fibrillation
G573300	Non-rheumatic atrial fibrillation
G573400	Permanent atrial fibrillation
G573500	Persistent atrial fibrillation
G573z00	Atrial fibrillation and flutter NOS

Appendix O

Respiratory infection Read code list

Read code	Description
H0...00	Acute respiratory infections
H00..00	Acute nasopharyngitis
H00..11	Common cold
H00..12	Coryza - acute
H00..13	Febrile cold
H00..14	Nasal catarrh - acute
H00..15	Pyrexial cold
H00..16	Rhinitis - acute
H01..00	Acute sinusitis
H01..11	Sinusitis
H010.00	Acute maxillary sinusitis
H010.11	Antritis - acute
H011.00	Acute frontal sinusitis
H012.00	Acute ethmoidal sinusitis
H013.00	Acute sphenoidal sinusitis
H014.00	Acute rhinosinusitis
H01y.00	Other acute sinusitis
H01y000	Acute pansinusitis
H01yz00	Other acute sinusitis NOS
H01z.00	Acute sinusitis NOS
H02..00	Acute pharyngitis
H02..11	Sore throat NOS
H02..12	Viral sore throat NOS
H02..13	Throat infection - pharyngitis
H020.00	Acute gangrenous pharyngitis
H021.00	Acute phlegmonous pharyngitis
H022.00	Acute ulcerative pharyngitis
H023.00	Acute bacterial pharyngitis
H023000	Acute pneumococcal pharyngitis
H023100	Acute staphylococcal pharyngitis
H023z00	Acute bacterial pharyngitis NOS
H024.00	Acute viral pharyngitis
H025.00	Allergic pharyngitis
H02z.00	Acute pharyngitis NOS
H03..00	Acute tonsillitis
H03..11	Throat infection - tonsillitis
H03..12	Tonsillitis
H030.00	Acute erythematous tonsillitis
H031.00	Acute follicular tonsillitis
H032.00	Acute ulcerative tonsillitis
H033.00	Acute catarrhal tonsillitis
H034.00	Acute gangrenous tonsillitis
H035.00	Acute bacterial tonsillitis
H035000	Acute pneumococcal tonsillitis
H035100	Acute staphylococcal tonsillitis
H035z00	Acute bacterial tonsillitis NOS
H036.00	Acute viral tonsillitis
H037.00	Recurrent acute tonsillitis
H03z.00	Acute tonsillitis NOS
H04..00	Acute laryngitis and tracheitis
H040.00	Acute laryngitis
H040000	Acute oedematous laryngitis
H040100	Acute ulcerative laryngitis

Read code	Description
H040200	Acute catarrhal laryngitis
H040300	Acute phlegmonous laryngitis
H040400	Acute haemophilus influenzae laryngitis
H040500	Acute pneumococcal laryngitis
H040600	Acute suppurative laryngitis
H040w00	Acute viral laryngitis unspecified
H040x00	Acute bacterial laryngitis unspecified
H040z00	Acute laryngitis NOS
H041.00	Acute tracheitis
H041000	Acute tracheitis without obstruction
H041100	Acute tracheitis with obstruction
H041z00	Acute tracheitis NOS
H042.00	Acute laryngotracheitis
H042.11	Laryngotracheitis
H042000	Acute laryngotracheitis without obstruction
H042100	Acute laryngotracheitis with obstruction
H042z00	Acute laryngotracheitis NOS
H043.00	Acute epiglottitis (non strep)
H043.11	Viral epiglottitis
H043000	Acute epiglottitis without obstruction
H043100	Acute epiglottitis with obstruction
H043200	Acute obstructive laryngitis
H043211	Croup
H043z00	Acute epiglottitis NOS
H044.00	Croup
H04z.00	Acute laryngitis and tracheitis NOS
H05..00	Other acute upper respiratory infections
H050.00	Acute laryngopharyngitis
H051.00	Acute upper respiratory tract infection
H052.00	Pharyngotracheitis
H053.00	Tracheopharyngitis
H054.00	Recurrent upper respiratory tract infection
H055.00	Pharyngolaryngitis
H05y.00	Other upper respiratory infections of multiple sites
H05z.00	Upper respiratory infection NOS
H05z.11	Upper respiratory tract infection NOS
H05z.12	Viral upper respiratory tract infection NOS
H06..00	Acute bronchitis and bronchiolitis
H060.00	Acute bronchitis
H060.11	Acute wheezy bronchitis
H060000	Acute fibrinous bronchitis
H060100	Acute membranous bronchitis
H060200	Acute pseudomembranous bronchitis
H060300	Acute purulent bronchitis
H060400	Acute croupous bronchitis
H060500	Acute tracheobronchitis
H060600	Acute pneumococcal bronchitis
H060700	Acute streptococcal bronchitis
H060800	Acute haemophilus influenzae bronchitis
H060900	Acute neisseria catarrhalis bronchitis
H060A00	Acute bronchitis due to mycoplasma pneumoniae
H060B00	Acute bronchitis due to coxsackievirus
H060C00	Acute bronchitis due to parainfluenza virus
H060D00	Acute bronchitis due to respiratory syncytial virus
H060E00	Acute bronchitis due to rhinovirus
H060F00	Acute bronchitis due to echovirus
H060v00	Subacute bronchitis unspecified
H060w00	Acute viral bronchitis unspecified
H060x00	Acute bacterial bronchitis unspecified
H060z00	Acute bronchitis NOS
H061.00	Acute bronchiolitis
H061000	Acute capillary bronchiolitis
H061100	Acute obliterating bronchiolitis
H061200	Acute bronchiolitis with bronchospasm
H061300	Acute exudative bronchiolitis
H061400	Obliterating fibrous bronchiolitis
H061500	Acute bronchiolitis due to respiratory syncytial virus
H061600	Acute bronchiolitis due to other specified organisms
H061z00	Acute bronchiolitis NOS
H062.00	Acute lower respiratory tract infection
H06z.00	Acute bronchitis or bronchiolitis NOS
H06z000	Chest infection NOS
H06z011	Chest infection
H06z100	Lower resp tract infection
H06z111	Respiratory tract infection
H06z112	Acute lower respiratory tract infection
H06z200	Recurrent chest infection
H07..00	Chest cold
H0y..00	Other specified acute respiratory infections
H0z..00	Acute respiratory infection NOS

Appendix P

Heavy drinker Read code list

Read code	Description
1364.00	Moderate drinker - 3-6u/day
1365.00	Heavy drinker - 7-9u/day
1366.00	Very heavy drinker - ¿9u/day
136K.00	Alcohol intake above recommended sensible limits
136O.00	Moderate drinker
136P.00	Heavy drinker
136Q.00	Very heavy drinker
136R.00	Binge drinker
136T.00	Harmful alcohol use
13Y8.00	Alcoholics anonymous
14C5.00	H/O: liver disease
7L1f.00	Compensation for liver failure
7L1fy00	Other specified compensation for liver failure
7L1fz00	Compensation for liver failure NOS
9kX..00	Hepatitis status 6 months post treatment - enhanced serv adm
9kX..11	Hepatitis status 6 months post treatment
C251.11	Wernicke's encephalopathy
C253.00	Wernicke's encephalopathy
D307000	Deficiency of coagulation factor due to liver disease
E01..00	Alcoholic psychoses
E011.00	Alcohol amnestic syndrome
E011000	Korsakov's alcoholic psychosis
E011100	Korsakov's alcoholic psychosis with peripheral neuritis
E011200	Wernicke-Korsakov syndrome
E011z00	Alcohol amnestic syndrome NOS
E012.00	Other alcoholic dementia
E012.11	Alcoholic dementia NOS
E012000	Chronic alcoholic brain syndrome
E013.00	Alcohol withdrawal hallucinosis
E01y.00	Other alcoholic psychosis
E01y000	Alcohol withdrawal syndrome
E01yz00	Other alcoholic psychosis NOS
E01z.00	Alcoholic psychosis NOS
E23..00	Alcohol dependence syndrome
E23..11	Alcoholism
E23..12	Alcohol problem drinking
E231300	Chronic alcoholism in remission
E250.00	Nondependent alcohol abuse
E250000	Nondependent alcohol abuse, unspecified
E250200	Nondependent alcohol abuse, episodic
Eu10011	[X]Acute alcoholic drunkenness
Eu10212	[X]Chronic alcoholism
Eu10300	[X]Mental and behav dis due to use alcohol: withdrawal state
Eu10400	[X]Men and behav dis due alcohl: withdrawl state with delirium
Eu10411	[X]Delirium tremens, alcohol induced
Eu10500	[X]Mental and behav dis due to use alcohol: psychotic disorder
Eu10511	[X]Alcoholic hallucinosis
Eu10512	[X]Alcoholic jealousy
Eu10514	[X]Alcoholic psychosis NOS
Eu10611	[X]Korsakov's psychosis, alcohol induced
Eu10700	[X]Men and behav dis due alcoh: resid and late-onset psychot dis
Eu10711	[X]Alcoholic dementia NOS

Read code	Description
Eu10712	[X]Chronic alcoholic brain syndrome
Eu10800	[X]Alcohol withdrawal-induced seizure
F11x000	Cerebral degeneration due to alcoholism
F11x011	Alcoholic encephalopathy
G852200	Oesophageal varices in cirrhosis of the liver
G852300	Oesophageal varices in alcoholic cirrhosis of the liver
J600000	Acute hepatic failure
J600011	Acute liver failure
J601000	Subacute hepatic failure
J61..00	Cirrhosis and chronic liver disease
J611.00	Acute alcoholic hepatitis
J612.00	Alcoholic cirrhosis of liver
J612.11	Florid cirrhosis
J612.12	Laennec's cirrhosis
J612000	Alcoholic fibrosis and sclerosis of liver
J613.00	Alcoholic liver damage unspecified
J613000	Alcoholic hepatic failure
J614.00	Chronic hepatitis
J614000	Chronic persistent hepatitis
J614100	Chronic active hepatitis
J614200	Chronic aggressive hepatitis
J614300	Recurrent hepatitis
J614400	Chronic lobular hepatitis
J614y00	Chronic hepatitis unspecified
J614z00	Chronic hepatitis NOS
J615.11	Portal cirrhosis
J615000	Unilobular portal cirrhosis
J615100	Multilobular portal cirrhosis
J615111	Postnecrotic cirrhosis of liver
J615200	Mixed portal cirrhosis
J615300	Diffuse nodular cirrhosis
J615400	Fatty portal cirrhosis
J615500	Hypertrophic portal cirrhosis
J615600	Capsular portal cirrhosis
J615700	Cardiac portal cirrhosis
J615711	Congestive cirrhosis
J615A00	Pipe-stem portal cirrhosis
J615B00	Toxic portal cirrhosis
J615C00	Xanthomatous portal cirrhosis
J615y00	Portal cirrhosis unspecified
J615z11	Macronodular cirrhosis of liver
J615z13	Cirrhosis of liver NOS
J615z14	Laennec's cirrhosis, non-alcoholic
J615z15	Hepatic fibrosis
J617.00	Alcoholic hepatitis
J617000	Chronic alcoholic hepatitis
J61y300	Portal fibrosis without cirrhosis
J61y400	Hepatic fibrosis
J61y600	Hepatic fibrosis with hepatic sclerosis
J61z.00	Chronic liver disease NOS
J625.00	[X] Hepatic failure
J625.11	[X] Liver failure
J62y.00	Other sequelae of chronic liver disease
J62y.11	Hepatic failure NOS
J62y.12	Liver failure NOS
J62y.13	Hepatic failure
J633.00	Hepatitis unspecified
J633000	Toxic hepatitis
J633z00	Hepatitis unspecified NOS
J635.00	Toxic liver disease
J635000	Toxic liver disease with cholestasis
J635100	Toxic liver disease with hepatic necrosis
J635200	Toxic liver disease with acute hepatitis
J635300	Toxic liver disease with chronic persistent hepatitis
J635400	Toxic liver disease with chronic lobular hepatitis
J635500	Toxic liver disease with chronic active hepatitis
J635600	Toxic liver disease with fibrosis and cirrhosis of liver
J635X00	Toxic liver disease, unspecified
J63B.00	Autoimmune hepatitis
J63X.00	Granulomatous hepatitis, not elsewhere classified
J63y100	Nonspecific reactive hepatitis
J671000	Alcohol-induced chronic pancreatitis
Jyu7000	[X]Toxic liver disease with other disorders of liver
Jyu7100	[X]Other and unspecified cirrhosis of liver
Jyu7200	[X]Other specified inflammatory liver diseases
Jyu7600	[X]Toxic liver disease, unspecified
Jyu7700	[X]Granulomatous hepatitis, not elsewhere classified
ZC2CH11	Dietary advice for liver disease

Appendix Q

Total cholesterol non-response to statin treatment for patients diagnosed with type 2 diabetes - protocol

Q.1 Background

Patients diagnosed with type 2 diabetes areat high risk of developing cardiovascular disease (CVD). National Institute for Clinical Excellence (NICE) recommends prescribing statins to all patients with diabetes to control lipids and reduce CVD risk[103]. However, some patients with diabetes prescribed statins have a residual CVD risk due to atherogenic dyslipidaemia, a blood fat disorder causing artery walls to thicken. It is characterised by low levels of high-density lipoprotein (HDL) cholesterol and high triglycerides and low-density lipoprotein (LDL) cholesterol, resulting in high total cholesterol values,which are associated with increased CVD risk[104]. Diabetic patients with this condition do not always respond to statins despite high-dose statin therapy[105], but it is unclear which other factors are associated with non-response.

Many studies use CVD events to assess the performance of statins. However, CVD events may not occur for some time, if at all, during the study period. Therefore, total cholesterol may act as a surrogate measure for CVD risks if we assume a patients CVD risk reduces if total cholesterol decreases.

The National Institute for Clinical Excellence (NICE) identifies the following health indicators for CVD in patients diagnosed with diabetes: overweight, high blood pressure, high serum albumin, smoking and high-risk lipid profile (low HDL cholesterol and high LDL cholesterol)[103] and in 2004, the Quality Outcomes Framework (QOF)[107] was introduced to encourage regular recording of health indicators in people with chronic diseases like diabetes. Under the scheme, practices receive points for providing good quality of care. QOF award points for recording health indicators to monitor patients with specified diseases. For example, GPs record body mass index (BMI), smoking status, glycosylated haemoglobin (HbA1C), blood pressure and total cholesterol measurements for patients diagnosed with type 1 or type 2 diabetes. It is in the practices interest to gain as many QOF points as possible, not only for financial reward but also to ensure good management of chronic diseases in primary care.

Therefore, health indicators associated with increased risk of the QOF specified diseases are more frequently recorded following the introduction of this legislation[8, 13].

In this study, we will use primary care records of patients with type 2 diabetes and initiated statin treatment to examine the association between sociodemographic variables and health indicators (measured before initiating statin therapy) and the difference between total cholesterol measurements before and within the first 6 months after first statin treatment. We were interested in analysing LDL cholesterol levels, but an exploratory analysis of the data found a large percentage of patients were missing LDL cholesterol measurements. Total cholesterol measurements were recorded more frequently than LDL cholesterol,so we chose to analyse total cholesterol instead of LDL cholesterol. We will analyse the data using a regression model and investigate different methods to deal with missing data using multiple imputation. We want to handle missing data appropriately so we will compare bias and precision of different methods of handling missing data. We expect imputed data are less biased and most precise compared to complete records analysis.

Q.2 Purpose

The aim of this study is to investigate the extent sociodemographic variables and health indicators measured before initiating statin therapy are associated with the difference between total cholesterol measurements before and within the first 6 months after first statin treatment for patients diagnosed with type 2 diabetes.
The specific objectives are, for patients diagnosed with type 2 diabetes:

1. To examine change in total cholesterol after first statin treatment compared to before first statin treatment.
2. To identify factors associated with lack of response to statin treatment for total cholesterol.
3. To evaluate how different methods to handle missing health indicator values impact the analysis results.

Q.3 Data source

We will extract data from The Health Improvement Network (THIN) database, a large, longitudinal, clinical primary care database widely used in epidemiological research. Data are collected from general practices using the Vision practice management software that have elected to join the THIN Quality Data recording scheme, administered by Cegedim Strategic Data (CSD)[11]. Medical events are coded using the hierarchical Read system of coding[26] and prescriptions are coded using multilex encrypted ID codes from the UK Prescription Pricing Authority and classified according to chapters in the British National Formulary (BNF)[27].

The patient data in THIN are approximately representative of the UK population[29]. All data collected are anonymised at source before leaving the GP system and are continually updated. THIN contains data from over 11 million patients registered with more than 500 practices. The recording of both consultations and prescriptions are similar to national consultation and prescription statistics[30, 31]. The data provider CSD has obtained overall ethical approval from the South East MREC (MREC/03/01/073).

Patient information is recorded at irregular times from the point of registration with the practice to the time they leave the practice, providing a longitudinal record of health data. Available data consists of patient characteristics, medical (symptoms and diagnoses) and prescription information. THIN collects additional information on referral to specialists, laboratory results, some lifestyle characteristics and other measurements taken in the GPs practice. Information on area deprivation(Townsend score) is based on the patients electoral ward from the 2001 Population Census (`http://www.statistics.gov.uk/census2001/census2001.asp`).

Q.4 Methods

Q.4.1 Study design

This is a cohort study.

Q.4.2 Study population

We will include practices from the date the computer system is used for clinical consultations (acceptable computer usage (ACU))[75],the practice has acceptable mortality recording (AMR)[74], a list size ¿ 2,000 patients and at least 80% of patients with Townsend score quintile recorded.

We will include patients:

1. diagnosed with type 2 diabetes. This includes patients where:
    - a type 2 diabetes Read code is in the patient record. We will assume patients with non-specific diabetes are type 2 diabetes and include them in the analysis;
    - an AHD code for annual diabetes check, current diabetes status or insulin dosage;
    - the data type is diabetic register, diabetic consultation or diabetes concerns; or
    - they are prescribed type 2 diabetes specific medication.
2. with no previous CVD event before first statin prescription;
3. permanently registered to the practice (patient flag A or C);
4. actively registered with the practice at some time from 1 January 2004 to 31st December 2012;
5. with date of birth, gender and Townsend deprivation quintile recorded.

We will include patients 30 years or over when they enter the dataset, which occurs at the latest date of(i) first type 2 diabetes diagnosis; (ii) ACU date; (iii) AMR date; (iv) when the patient registered with the practice or (v) 1 January 2004.After this first date,we will identify the first statin prescription. To exclude prevalent cases, we will exclude patients prescribed statins before the first date or within 6 months of the date patients enter the study.

We will include patient measurements in the analysis up until the earliest date of (i) death; (ii) when the patient transferred out of the practice; (iii) the last date the practice contributes data to THIN or (iv) 31 December 2012.

We will exclude patients if the time between first statin prescription and last date is less than 6 months. Also, patients must receive at least two statin prescriptions within the first 6 month after first statin treatment.

Q.4.3 Study variables

In this study, we will investigate the association between the difference between total cholesterol measurements before and within the first 6 months after first statin treatment and the following factors measured within 6 months before initiation of statin treatment, called baseline:

- Sociodemographic: age at first statin prescription, gender, deprivation (quintiles of Townsend scores) and ethnicity (white, black, South Asian or other).
- Diseases: chronic kidney disease.
- Health indicators: body mass index (BMI), systolic blood pressure, total cholesterol, HDL cholesterol, LDL cholesterol, glycosylated haemoglobin, glomerular filtration rate and smoking status.
- Statin therapy: statin type and dose.
- Other variables: Calendar year of first statin prescription and time from diagnosis of diabetes to first prescription of statin.

To measure total cholesterol change, we define the outcome variable as the difference between the total cholesterol measurement at baseline and the smallest total cholesterol measurement within the first 6 months after the first statin treatment. Using this outcome, the constant term in the regression model is zero. We will include all patients with a total cholesterol measurement at baseline and within the first 6 months after first statin prescription.When the data are extracted, we will identify and exclude outliers using methods described previously[4].

Q.4.4 Analysis

In the baseline time block, we will include the measurement closest to the date of first statin prescription in the analysis if more than one measurement was recorded.

We will first examine how total cholesterol changes from before to after first statin treatment. We will calculate the difference between the total cholesterol measurement at baseline and within the first 6 months after first statin prescription. We will investigate if this difference is normally distributed and calculate the mean difference for each baseline characteristic. We will also divide the difference into quintiles and produce a table of baseline characteristics for each total cholesterol difference quintile.

We will analyse the observed data using our model of interest: a linear model with a response variable of the difference between the total cholesterol measurement at baseline and within the first 6 months after first statin

prescription. We will condition on a range of sociodeomographic factors, diseases, statin therapy and health indicators recorded at baseline described above.

We will explore the proportions of missing data for variables at baseline. Before imputation, we will evaluate the plausibility of the MAR assumption to assess if MI is appropriate. If the distribution of the continuous variables with missing data is skewed, we will transform the data before imputation[47]. Next, we will impute missing values at baseline using fully conditional specification (FCS)[21] multiple imputation (MI) which accounts for uncertainty due to missing data and provides unbiased estimates under the Missing At Random (MAR) assumption (the reason for missing data is associated with observed data, but not the unobserved)[15, 18]. In addition to the variables measured at baseline in the model of interest (including the outcome), we will include the following auxiliary variables in the imputation model:

1. Patients with the following diseases diagnosed before first statin prescription: chronic obstructive pulmonary disease (COPD); schizophrenia, bipolar disorder and other psychoses (psychoses); asthma; and atrial fibrillation.

2. Patients prescribed following drugs during baseline:

   - Anti-hypertensive drug treatment .
   - Other lipid-modifying drugs dose and type bile acid sequestrants; ezetimibe; fibrate; nicotinic acid and omega 3 polyunsaturated fatty acid (PUFA) ethyl esters.

3. Respiratory infection during baseline.

4. Recognised alcohol problem (i.e. current heavy drinker).

5. Death create variable to indicate which patients die after first statin prescription.

6. CVD event identifies patients with a CVD event at any time after the first statin prescription (separate indicators for CHD and stroke).

Next, we will impute missing values at baseline again using multiple imputation, conditional on heath indicators recorded before baseline. The time before baseline is divided into 6 month time blocks, excluding the earliest time block if it is less than 6 months. We will include a maximum of 4 time blocks before baseline.If health indicators are recorded more than once during each time block, we will select a measurement at random to include in the analysis, except smoking status. If smoking status is recorded more than once,the most predominant category. Time is divided into blocks to simplify imputation of missing data. We will use the two-fold FCS algorithm[1], an extension of the FCS, to take account of the longitudinal and dynamic structure of the data. In addition to the sociodemographic variables, health indicators (measured during each time block), outcome from the model of interest, diseases diagnosed before first statin prescription (described above), recognised alcohol problem, death and CVD event, we will also include the following auxiliary variables in the imputation model:

1. We will identify patients prescribed following drugs in each time block:

- Anti-hypertensive drug treatment.
- Other lipid-modifying drugs: bile acid sequestrants, ezetimibe, fibrate, nicotinic acid, omega 3 polyunsaturated fatty acid (PUFA) ethyl esters.

2. Respiratory infection during each time block.

Finally, we will impute missing values at baseline using multiple imputation, conditional on heath indicators recorded before and after baseline. The time before and after baseline is divided into 6 month time blocks, excluding the earliest or latest time block if it is less than 6 months, as select measurements during each time block using the method described earlier. We will include a maximum of 4 time blocks before baseline and maximum of 4 time blocks after baseline. We will impute using the same imputation model used to impute missing values in time blocks before baseline.

After each imputation, we will fit the model of interest to the observed and imputed explanatory variables at baseline and compare to the results from the complete records analysis and the analysis of the data imputed using baseline imputation.

We will perform all analyses using Stata SE version 12.1.

Q.5 Limitations

Our patient selection method will include some patients with type 1 diabetes in the analysis because some Read codes do not specify the diabetes type. This could affect the results if patients with type 1 diabetes respond differently to statin treatment compared to patients with type 2 diabetes.

We will assume all patients administer statin treatment as directed. We only know if patients were prescribed statins, some may not administer them as directed. This can affect the results because patients who are prescribed statins but do not administer them correctly may show a non-response because they did not take the statin.

There may be unmeasured confounding. It is not possible to adjust for all confounding because this is a cohort study.

Bibliography

[1] J. Nevalainen, M.G. Kenward, and S.M. Virtanen. Missing values in longitudinal dietary data: a multiple imputation approach based on a fully conditional specification. *Stat.Med.*, 28(29):3657–3669, 2009.

[2] C. Welch, J. Bartlett, and Petersen I. Application of multiple imputation using the two-fold fully conditional specification algorithm in longitudinal clinical data. *Stata Journal*, 14(2):418–431, 2014.

[3] C. Welch, Petersen I., J. Bartlett, I. White, L. Marston, R. Morris, I. Nazareth, K. Walters, and J. Carpenter. Evaluation of two-fold fully conditonal specification multiple imputation for longitudinal electronic health record data. *Stat.Med.*, 33(21):3725–3737, 2014.

[4] C. Welch, I. Petersen, K. Walters, R.W. Morris, I. Nazareth, E. Kalaitzaki, I.R. White, L. Marston, and J. Carpenter. Two-stage method to remove population- and individual-level outliers from longitudinal data in a primary care database. *Pharmacoepidemiol.Drug Saf.*, 21(7):725–732, 2011.

[5] L. P. M. M. Wijlaars, I. Nazareth, and I. Petersen. Trends in depression and antidepressant prescribing in children and adolescents: A cohort study in the health improvement network (thin). *PLoS.One*, 7(3):e33181, 2012.

[6] I. Petersen, R. E. Gilbert, S. J. Evans, S. L. Man, and I. Nazareth. Pregnancy as a major determinant for discontinuation of antidepressants: an analysis of data from the health improvement network. *J.Clin.Psychiatry*, 72(7):979–85, 2011.

[7] S. Hardoon, J. F. Hayes, R. Blackburn, I. Petersen, K. Walters, I. Nazareth, and D. P. Osborn. Recording of severe mentall illness in united kingdom primary care, 2000-2010. *PLoS.One*, 8(12):e82365, 2013.

[8] J. Hippisley-Cox, C. Coupland, Y. Vinogradova, J. Robson, M. May, and P. Brindle. Derivation and validation of qrisk, a new cardiovascular disease risk score for the united kingdom: prospective open cohort study. *BMJ*, 335(7611):136, 2007.

[9] J.A. Delaney, S.S. Daskalopoulou, J.M. Brophy, R.J. Steele, L. Opatrny, and S. Suissa. Lifestyle variables and the risk of myocardial infarction in the general practice research database. *BMC Cardiovasc.Disord.*, 7:38, 2007.

[10] J. Hippisley-Cox, C. Coupland, Y. Vinogradova, J. Robson, R. Minhas, A. Sheikh, and P. Brindle. Predicting cardiovascular risk in england and wales: prospective derivation and validation of qrisk2. *BMJ*, 336(7659):1475–1482, 2008.

[11] Cegedim Strategic Data. http://csdmruk.cegedim.com/our-data/our-data.shtml. 2014.

[12] National Health Service. *General Medical Services (GMS) Regulations (Statutory Instrument 1992 No. 635) as amended; Schedule 2, paragraph 14*, 1992.

[13] L. Marston, J. R. Carpenter, K. R. Walters, R. W. Morris, I. Nazareth, and I. Petersen. Issues in multiple imputation of missing data for large general practice clinical databases. *Pharmacoepidemiol.Drug Saf.*, 19(6):618–626, 2010.

[14] The Information Centre. Quality and outcomes framework. http://www.qof.ic.nhs.uk/.

[15] R.J. Little and D.B. Rubin. *Statistical analysis with missing data*. Wiley, New York, 2002.

[16] 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.*, 134(8):895–907, 1991.

[17] J. Carpenter and M.G. Kenward. Missing data in randomised controlled trials: a practical guide. Publication RM03/JH17/MK, 2008.

[18] D.B. Rubin. *Multiple imputation for nonresponse in surveys*. Wiley, New York, 1987.

[19] J.A. 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*, 338:b2393, 2009.

[20] J. Carpenter and M.G. Kenward. *Multiple Imputation and its Application: Statistics in Practice*. Wiley, UK, 2013.

[21] S. van Buuren, J.P.L. Brand, K. Groothuis-Oudshoorn, and D.B. Rubin. Fully conditional specification in multivariate imputation. *J.Stat.Comput.Simul.*, 76(12):1049–1064, 2006.

[22] S. van Buuren, H.C. Boshuizen, and D.L. Knook. Multiple imputation of missing blood pressure covariates in survival analysis. *Stat.Med.*, 18(6):681–694, 1999.

[23] I.R. White, P. Royston, and A.M. Wood. Multiple imputation using chained equations: Issues and guidance for practice. *Stat.Med.*, 30(4):377–399, 2011.

[24] Clinical practice research datalink. *www.cprd.com*, 2014.

[25] Qresearch. *www.qresearch.org*, 2014.

[26] J. Chisholm. The read clinical classification. *BMJ*, 300(6732):1092, 90.

[27] British National Formulary. Joint formulary committee. *59th ed. London: British Medical Association and Royal Pharmaceutical Society of Great Britain*, 2009.

[28] P. Townsend, P. Phillimore, and A. Beattie. The construction of a measure of deprivation. In *Health and Deprivation: Inequality and the North*, pages 34–38. Routledge, London, 1988.

[29] B. T. Blak, M. Thompson, H. Dattani, and A. Bourke. Generalisability of the health improvement network (thin) database: demographics, chronic disease prevalence and mortality rates. *Inform.Prim.Care*, 19(4):251–255, 2011.

[30] A. Bourke, H. Dattani, and M. Robinson. Feasibility study and methodology to create a quality-evaluated database of primary care data. *Inform.Prim.Care*, 12(3):171–177, 2004.

[31] F. D. McClure, J. K. Lee, and D. B. Wilson. Validity of the percent reduction in standard deviation outlier test for screening laboratory means from a collaborative study. *J.AOAC Int.*, 86(5):1045–1055, 2003.

[32] General Medical Service. National health service (general medical and pharmaceutical services) regulations 1974 (statutory instrument 1974 no.160) as amended. 1990.

[33] J. Carpenter, H. Goldstein, and M.G. Kenward. Statistical modelling of partially observed data using multiple imputation: principles and practice. In *Modern Methods for Epidemiology*. Springer, 2012.

[34] J. Carpenter and I. Plewis. Coming to terms with non-response in longitudinal studies. In *The SAGE handbook of Methodological Innovation*. SAGE, London, 2009.

[35] J.W. Graham. Missing data analysis: making it work in the real world. *Annu.Rev.Psychol.*, 60:549–576, 2009.

[36] P.D. Allison. Multiple imputation for missing data: A cautionary tale. *Sociol. Methods Res.*, 28(3):301–309, 2000.

[37] I.R. White and J.B. Carlin. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Stat.Med.*, 29(28):2920–2931, 2010.

[38] R.J. Cook, L. Zeng, and G.Y. Yi. Marginal analysis of incomplete longitudinal binary data: a cautionary note on locf imputation. *Biometrics*, 60(3):820–828, 2004.

[39] M. G. Kenward and G. Molenberghs. Last observation carried forward: a crystal ball? *J.Biopharm.Stat.*, 19(5):872–888, 2009.

[40] G.J. van der Heijden, A.R. Donders, T. Stijnene, and K.G. Moons. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: A clinical example. *J.Clin.Epidemiol.*, 59(10):1102–1109, 2006.

[41] K.I. Penny and I. Atkinson. Approaches for dealing with missing data in health care studies. *J.Clin.Nurs.*, 21(19-20):2722–9, 2011.

[42] P. D. Allison. Handing missing data by maximum likelihood. *SAS Global Forum*, 312, 2012.

[43] D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–92, 1976.

[44] J. W. Graham and D.L. Coffman. Structural equation modelling with missing data. In *Handbook of Structural Equation Modeling*. The Guildford Press, New York, 2011.

[45] S. van Buuren. Multiple imputation of discrete and continuous data by fully conditional specification. *Stat.Methods Med.Res.*, 16(3):219–242, 2007.

[46] M.G. Kenward and J. Carpenter. Multiple imputation: current perspectives. *Stat.Methods Med.Res.*, 16(3):199–218, 2007.

[47] K.J. Lee and J.B. Carlin. Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. *Am.J.Epidemiol.*, 171(5):624–632, 2010.

[48] Liu J and A Gelman. On the stationary distribution of iterative imputations. *Biometrika*, 101(1):155–173, 2014.

[49] R. A. Hughes, I. R. White, S. R. Seaman, J. R. Carpenter, K. Tilling, and J. A. Sterne. Joint modelling rationale for chained equations. *BMC Med.Res.Methodol.*, 14:28, 2014.

[50] J.L. Schafer and M.K. Olsen. Multiple imputation for multivariate missing-data problems: A data analyst's perspective. In *Multivariate Behavioral Research*. Lawrence Erlbaurn Associates, Inc, 1998.

[51] K.G. Moons, R.A. Donders, T. Stijnen, and F.E. Harrell. Using the outcome for imputation of missing predictor values was preferred. *J.Clin.Epidemiol.*, 59(10):1092–1101, 2006.

[52] L.M. Collins, J.L. Schafer, and C.M. Kam. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol.Methods*, 6(4):330–351, 2001.

[53] J.L. Schafer. *Analysis of incomplete multivariate data*. Chapman and Hall, 1997.

[54] S. R. Seaman, J. W. Bartlett, and I. R. White. Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. *BMC.Med.Res.Methodol.*, 12(1):46, 2012.

[55] J. W. Bartlett, S. R. Seaman, I. R. White, and J. R. Carpenter. Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. *Stat.Methods.Med.Res. - Epub ahead of print*, 2014.

[56] T. P. Morris, I. R. White, P. Royston, S. R. Seaman, and A. M. Wood. Multiple imputation for an incomplete covariate that is a ratio. *Stat.Med.*, 33(1):88–104, 2014.

[57] I.R. White and P. Royston. Imputing missing covariate values for the cox model. *Stat.Med.*, 28(15):1982–98.

[58] M. Spratt, J. Carpenter, J.A. Sterne, J.B. Carlin, J. Heron, J. Henderson, and K. Tilling. Strategies for multiple imputation in longitudinal studies. *Am.J.Epidemiol.*, 172(4):478–487, 2010.

[59] M.A. Klebanoff and S.R. Cole. Use of multiple imputation in the epidemiologic literature. *Am.J.Epidemiol.*, 168(4):355–357, 2008.

[60] A.M. Wood, I.R. White, and S.G. Thompson. Are missing outcome data adequately handled? a review of published randomized controlled trials in major medical journals. *Clin.Trials*, 1(4):368–376, 2004.

[61] A. Mackinnon. The use and reporting of multiple imputation in medical research - a review. *J.Intern.Med.*, 268(6):586–593, 2010.

[62] A. Burton and D.G. Altman. Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. *Br.J.Cancer*, 91(1):4–8, 2004.

[63] C. J. Currie, M. Peyrot, C. L. Morgan, C. D. Poole, S. Jenkins-Jones, R. R. Rubin, C. M. Burton, and M. Evans. The impact of treatment non-compliance on mortality in people with type 1 diabetes. *J.Diabetes Complications*, 27(3):219–223, 2013.

[64] D. Ose, J. Rochon, S. M. Campbell, M. Wensing, Lieshout J. van, L. Uhlmann, T. Freund, J. Szecsenyi, and S. Ludt. Secondary prevention in patients with coronary heart diseases: what factors are associated with health status in usual primary care? *PLoS.One*, 7(12):e51726, 2012.

[65] C. J. Hirst, C. Cabrera, and M. Kirby. Epidemiology of castration resistant prostate cancer: a longitudinal analysis using a uk primary care database. *Cancer Epidemiol.*, 36(6):e349–e353, 2012.

[66] C. J. Currie, M. Peyrot, C. L. Morgan, C. D. Poole, S. Jenkins-Jones, R. R. Rubin, C. M. Burton, and M. Evans. The impact of treatment noncompliance on mortality in people with type 2 diabetes. *Diabetes Care*, 35(6):1279–1284, 2012.

[67] G. Collins and D. Altman. Predicting the risk of chronic kidney disease in the uk: an evaluation of qkidney(r) scores using a primary care database. *Br.J.Gen.Pract.*, 62(597):e243–e250, 2012.

[68] N. F. Khan, D. Mant, L. Carpenter, D. Forman, and P. W. Rose. Long-term health outcomes in a british cohort of breast, colorectal and prostate cancer survivors: a database study. *Br.J.Cancer*, 105 Suppl 1:S29–S37, 2011.

[69] Y. Vinogradova, C. Coupland, and J. Hippisley-Cox. Risk of pneumonia in patients taking statins: population-based nested case-control study. *Br.J.Gen.Pract.*, 61(592):742–748, 2011.

[70] D. P. Miller, S. E. Watkins, T. Sampson, and K. J. Davis. Long-term use of fluticasone propionate/salmeterol fixed-dose combination and incidence of cataracts and glaucoma among chronic obstructive pulmonary disease patients in the uk general practice research database. *Int.J.Chron.Obstruct.Pulmon.Dis.*, 6:467–476, 2011.

[71] Y. Vinogradova, C. Coupland, and J. Hippisley-Cox. Exposure to statins and risk of common cancers: a series of nested case-control studies. *BMC.Cancer*, 11:409, 2011.

[72] S. Suissa, L. Azoulay, S. Dell'Aniello, M. Evans, J. Vora, and M. Pollak. Long-term effects of insulin glargine on the risk of breast cancer. *Diabetologia*, 54(9):2254–2262, 2011.

[73] X.L. Meng. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 9(4):538–558, 1994.

[74] A. Maguire, B.T. Blak, and M. Thompson. The importance of defining periods of complete mortality reporting for research using automated data from primary care. *Pharmacoepidemiol.Drug Saf.*, 18(1):76–83, 2009.

[75] L. Horsfall, K. Walters, and I. Petersen. Identifying periods of acceptable computer usage in primary care research databases. *Pharmacoepidemiol.Drug Saf.*, 22(1):64–69, 2013.

[76] National Centre for Social Research, University College London. Department of Epidemiology, and Public Health. Health survey for england, 1998, 5th edition. 1998.

[77] National Centre for Social Research, University College London. Department of Epidemiology, and Public Health. Health survey for england, 2006. 2006.

[78] National Centre for Social Research, University College London. Department of Epidemiology, and Public Health. Health survey for england, 2008, 2nd edition. 2008.

[79] 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*, 3(9):e003389, 2013.

[80] Y. Wang, K. Hunt, I. Nazareth, N. Freemantle, and I. Petersen. Do men consult less than women? an analysis of routinely collected uk general practice data. *BMJ Open*, 3(8):e003320, 13.

[81] S.M. Campbell and M.O. Roland. Why do people consult the doctor? *Fam.Pract.*, 13(1):75–83, 1996.

[82] N. Kapur, I. Hunt, M. Lunt, J. McBeth, F. Creed, and G. Macfarlane. Psychosocial and illness related predictors of consultation rates in primary care–a cohort study. *Psychol.Med.*, 34(4):719–728, 2004.

[83] N. Kapur, I. Hunt, M. Lunt, J. McBeth, F. Creed, and G. Macfarlane. Primary care consultation predictors in men and women: a cohort study. *Br.J.Gen.Pract.*, 55(511):108–113, 2005.

[84] D. Wormser, S. Kaptoge, Angelantonio E. Di, A. M. Wood, L. Pennells, A. Thompson, N. Sarwar, J. R. Kizer, D. A. Lawlor, B. G. Nordestgaard, P. Ridker, V. Salomaa, J. Stevens, M. Woodward, N. Sattar, R. Collins, S. G. Thompson, G. Whitlock, and J. Danesh. Separate and combined associations of body-mass index and abdominal adiposity with cardiovascular disease: collaborative analysis of 58 prospective studies. *Lancet*, 377(9771):1085–1095, 2011.

[85] F.J. Anscombe. Rejection of outliers. *Technometrics*, (2):123–147, 1960.

[86] W.J. Dixon. Analysis of extreme values. *Ann.Math.Statist.*, (21):488–506, 1950.

[87] J.D. Sorkin, D.C. Muller, and R. Andres. Longitudinal change in height of men and women: implications for interpretation of the body mass index: the baltimore longitudinal study of aging. *Am.J.Epidemiol.*, 150(9):969–977, 1999.

[88] M.J. Healy. Outliers in clinical chemistry quality-control schemes. *Clin.Chem.*, 25(5):675–677, 1979.

[89] K. Hayes, A. Kinsella, and N. Coffey. A note on the use of outlier criteria in ontario laboratory quality control schemes. *Clin.Biochem.*, 40(3-4):147–152, 2007.

[90] L.C. Chang, D.K. Jones, and C. Pierpaoli. Restore: robust estimation of tensors by outlier rejection. *Magn.Reson.Med.*, 53(5):1088–1095, 2005.

[91] D. Chen, X. Shao, B. Hu, and Q. Su. Simultaneous wavelength selection and outlier detection in multivariate regression of near-infrared spectra. *Anal.Sci.*, 21(2):161–166, 2005.

[92] R. Clarke, M. Shipley, S. Lewington, L. Youngman, R. Collins, M. Marmot, and R. Peto. Underestimation of risk associations due to regression dilution in long-term follow-up of prospective studies. *Am.J.Epidemiol.*, 150(4):341–353, 1999.

[93] G.S. Collins and D.G. Altman. An independent and external validation of qrisk2 cardiovascular disease risk score: a prospective open cohort study. *BMJ*, 340:c2442, 2010.

[94] L. Szatkowski, S. Lewis, A. McNeill, and T. Coleman. Is smoking status routinely recorded when patients register with a new gp? *Fam.Pract.*, 27(6):673–675, 2010.

[95] T. Coleman, S. Lewis, R. Hubbard, and C. Smith. Impact of contractual financial incentives on the ascertainment and management of smoking in primary care. *Addiction*, 102(5):803–808, 2007.

[96] S. Dave and I. Petersen. Creating medical and drug code lists to identify cases in primary care databases. *Pharmacoepidemiol.Drug Saf.*, 18(8):704–707, 2009.

[97] R. Bender, T. Augustin, and M. Blettner. Generating survival times to simulate cox proportional hazards models. *Stat.Med.*, 24(11):1713–1723, 2005.

[98] A. Burton, D.G. Altman, P. Royston, and R.L. Holder. The design of simulation studies in medical statistics. *Stat.Med.*, 25(24):4279–4292, 2006.

[99] J.D. Lewis, W.B. Bilker, R.B. Weinstein, and B.L. Strom. The relationship between time since registration and measured incidence rates in the general practice research database. *Pharmacoepidemiol.Drug Saf.*, 14(7):443–451, 2005.

[100] S. Lewington, R. Clarke, N. Qizilbash, R. Peto, and R. Collins. Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies. *Lancet*, 360(9349):1903–1913, 2002.

[101] S. Lewington, G. Whitlock, R. Clarke, P. Sherliker, J. Emberson, J. Halsey, N. Qizilbash, R. Peto, and R. Collins. Blood cholesterol and vascular mortality by age, sex, and blood pressure: a meta-analysis of individual data from 61 prospective studies with 55,000 vascular deaths. *Lancet*, 370(9602):1829–1839, 2007.

[102] Angelantonio E. Di, N. Sarwar, P. Perry, S. Kaptoge, K. K. Ray, A. Thompson, A. M. Wood, S. Lewington, N. Sattar, C. J. Packard, R. Collins, S. G. Thompson, and J. Danesh. Major lipids, apolipoproteins, and risk of vascular disease. *JAMA*, 302(18):1993–2000, 2009.

[103] NICE. National institute for health and clinical excellence. lipid modification: Cardiovascular risk assessment and the modification of blood lipids for the primary and secondary prevention of cardiovascular disease. 2008.

[104] E. Standl. Statins and beyond: Concurrent strategies for prevention of cardiovascular disease in patients with type 2 diabetes. *Diab.Vasc.Dis.Res.*, 10(2):99–114, 2012.

[105] J. C. Fruchart, F. Sacks, M. P. Hermans, G. Assmann, W. V. Brown, R. Ceska, M. J. Chapman, P. M. Dodson, P. Fioretto, H. N. Ginsberg, T. Kadowaki, J. M. Lablanche, N. Marx, J. Plutzky, Z. Reiner, R. S. Rosenson, B. Staels, J. K. Stock, R. Sy, C. Wanner, A. Zambon, and P. Zimmet. The residual risk reduction initiative: a call to action to reduce residual vascular risk in patients with dyslipidemia. *Am.J.Cardiol.*, 102(10 Suppl):1K–34K, 2008.

[106] W. T. Friedewald, R. I. Levy, and D. S. Fredrickson. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clin.Chem.*, 18(6):499–502, 1972.

[107] The Information Centre. Quality and outcomes framework http://qof.hscic.gov.uk/index.asp. 2012.

[108] M. Helfand, S. Carson, and C. Kelley. Drug class review on hmg-coa reductase inhibitors (statins). *Oregon Health and Sciences University, Drug Effectiveness Review Project*, 2004.

[109] J. M. Galema-Boers, M. J. Lenzen, R. T. van Domburg, Lennep J. Roeters van, van Bruchem-van de Scheur GG, E. J. Sijbrands, and J. G. Langendonk. Predicting non-adherence in patients with familial hypercholesterolemia. *Eur.J.Clin.Pharmacol.*, 70(4):391–397, 2014.

[110] T. Wada, M. Haneda, K. Furuichi, T. Babazono, H. Yokoyama, K. Iseki, S. I. Araki, T. Ninomiya, S. Hara, Y. Suzuki, M. Iwano, E. Kusano, T. Moriya, H. Satoh, H. Nakamura, M. Shimizu, T. Toyama, A. Hara, and H. Makino. Clinical impact of albuminuria and glomerular filtration rate on renal and cardiovascular events, and all-cause mortality in japanese patients with type 2 diabetes. *Clin.Exp.Nephrol. in press*, 2013.

[111] A. Solini, G. Penno, E. Bonora, C. Fondelli, E. Orsi, M. Arosio, R. Trevisan, M. Vedovato, M. Cignarelli, F. Andreozzi, A. Nicolucci, and G. Pugliese. Diverging association of reduced glomerular filtration rate and albuminuria with coronary and noncoronary events in patients with type 2 diabetes: the renal insufficiency and cardiovascular events (riace) italian multicenter study. *Diabetes Care*, 35(1):143–149, 2012.

[112] M. A. Ferro. Missing data in longitudinal studies: cross-sectional multiple imputation provides similar estimates to full-information maximum likelihood. *Ann.Epidemiol*, http://dx.doi.org/10.1016/j.annepidem.2013.10.007, 2013.

[113] E. J. Lamberts, G. Nijpels, L. M. Welschen, J. G. Hugtenburg, J. M. Dekker, P. C. Souverein, and M. L. Bouvy. Discontinuation of statins among patients with type 2 diabetes. *Diabetes Metab.Res.Rev.*, 28(3):241–245, 2012.

[114] National Clinical Guidline Centre. National institute for health and clinical excellence. lipid modification: Cardiovascular risk assessment and the modification of blood lipids for the primary and secondary prevention of cardiovascular disease - draft for consultation. 2014.

[115] A. Majeed. Increasing the use of statins in people at low cardiovascular risk is difficult. *BMJ*, 347:f6901, 2013.

[116] J.W. Graham, A.E. Olchowski, and T.D. Gilreath. How many imputations are really needed? some practical clarifications of multiple imputation theory. *Prev.Sci.*, 8(3):206–213, 2007.

[117] G.F. Liu and X.J. Zhan. Comparisons of methods for analysis of repeated binary responses with missing data. *J.Biopharm.Stat.*, 21(3):371–392, 2011.

[118] L. Saraceno, J. Heron, M. Munafo, N. Craddock, and M.B. van den Bree. The relationship between childhood depressive symptoms and problematic alcohol use in early adolescence: findings from a large longitudinal population-based study. alcohol problems in depressed boys and girls. *Addiction*, 107(3):567–77, 2011.

[119] U. Grittner, G. Gmel, S. Ripatti, K. Bloomfield, and M. Wicki. Missing value imputation in longitudinal measures of alcohol consumption. *Int.J.Methods Psychiatr.Res.*, 20(1):50–61, 2011.

[120] N. Lewis, L.S. Martinez, D.R. Freres, J.S. Schwartz, K. Armstrong, S.W. Gray, T. Fraze, R.H. Nagler, A. Bourgoin, and R.C. Hornik. Seeking cancer-related information from media and family/friends increases fruit and vegetable consumption among cancer patients. *Health Commun.*, 27(4):380–388, 2011.

[121] G. Howard, L.A. McClure, C.S. Moy, M.M. Safford, M. Cushman, S.E. Judd, B.M. Kissela, D.O. Kleindorfer, V.J. Howard, D.J. Rhodes, P. Muntner, and H.K. Tiwari. Imputation of incident events in longitudinal cohort studies. *Am.J.Epidemiol.*, 174(6):718–726, 2011.

[122] D.P. Osborn, G. Levy, I. Nazareth, I. Petersen, A. Islam, and M.B. King. Relative risk of cardiovascular and cancer mortality in people with severe mental illness from the united kingdom's general practice rsearch database. *Arch.Gen.Psychiatry*, 64(2):242–249, 2007.