THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

APPLYING MISSING DATA METHODS TO ROUTINE DATA USING THE EXAMPLE OF A POPULATION-BASED REGISTER OF PATIENTS WITH DIABETES

Stephanie Read

2015

UNIVERSITY OF EDINBURGH

DISSERTATION SUBMITTED FOR THE DEGREE OF

DOCTOR *of* PHILOSOPHY

DECLARATION

I, Stephanie Read, declare that the following thesis has been composed by me and has not been submitted for any other degree or professional qualification.

More specifically, Dr Steff Lewis and Professor Sarah Wild conceived the idea of identifying methods for handling unobserved data in the Scottish Care Information – Diabetes Care register. I proposed and performed the data manipulation strategy. Each statistical analysis presented in this thesis was planned, conducted and interpreted by myself. I conducted and transcribed all interviews with healthcare practitioners. The writing of this thesis has been done by me alone, under the supervision of Dr Steff Lewis and Professor Sarah Wild.

The study was funded by the MRC Network of Hubs for Trials Methodology Research.

Signature: ...……. Date: ……

ABSTRACT

Background:

Routinely-collected data offer great potential for epidemiological research and could be used to make randomised controlled trials (RCTs) more efficient. The use of routine data for research has been limited by concerns surrounding data quality, particularly data completeness. To fully exploit these information-rich data sources it is necessary to identify approaches capable of overcoming high proportions of missing data.

Using a 2008 extract of the Scottish Care Information – Diabetes Collaboration (SCI-DC) database, a population-based register of people with a diagnosis of diabetes in Scotland, I compared the findings of several methods for handling missing data in a retrospective cohort study investigating the association between body mass index (BMI) and all-cause mortality in patients with type 2 diabetes.

Methods:

Discussions with clinicians and logistic regression analyses were used to determine the likely mechanisms of missingness and the relative appropriateness of a selection of missing data methods, such as multiple imputation. Sequentially more complicated imputation approaches were used to handle missing data. Cox proportional hazard model coefficients for the association between BMI and all-cause mortality were compared for each missing data method. Age-standardised mortality rates by categories of BMI at around the time of diagnosis were also presented.

Results:

There were 66,472 patients diagnosed with type 2 DM between 2004 and 2008. Of these patients, 21% of patients did not have a recording of BMI at time of diagnosis. Amongst patients with complete BMI data, there were 5,491 deaths during 296,584 person years of follow-up. Amongst patients with incomplete data, there were 2,090 deaths during 79,067 person-years of follow-up. Analyses indicated that the primary mechanism of missingness was missing at random, conditional on patient year of diagnosis and vital status. In particular, patients with missing data had considerably worse survival than patients without missing data. Regardless of the method for handling the missing data, a U-shaped relationship between BMI and mortality was observed. Compared to complete case analysis, the association between BMI and all-

cause mortality was weaker using multiple imputation approaches with estimates moving towards the null. Closest observation imputation had the smallest effect on estimates compared to complete case analysis.

Risk of mortality was consistently highest in the less than 25kg/m² BMI group. For example, estimates obtained using multiple imputation using chained equations indicated that patients with a BMI below 25kg/m² had a 38% higher risk of mortality than patients in the 25 to less than 30kg/m² BMI category.

Conclusions:

Alternative methods to complete case analysis can be computationally intensive with many important practical considerations. However, it remains valuable to explore the robustness of estimates to departures from the assumptions made by complete case analysis. The use of these methods can preserve the sample size and therefore may be useful in developing risk prediction scores.

Mortality was lowest amongst overweight or obese patients relative to normal weight. Further work is required to identify optimal approaches to weight management amongst patients with diabetes.

ACKNOWLEDGEMENTS

I am grateful to a number of people who have provided exceptional advice, support and guidance throughout the duration of this PhD. Their contributions have been fundamental to the completion of this thesis.

Steff Lewis, my primary supervisor, has provided many encouraging words along the way. She has been patient, supportive and has provided invaluable and insightful guidance. Her unflappable nature has proven hugely beneficial in the face of my continual fretting.

Sarah Wild, my secondary supervisor, has been extremely generous with her time and advice. Her positivity and encouragement have never once faltered. I feel very fortunate to have had Sarah as a supervisor.

Nynke Halbesma has also been central to the completion of this thesis. Not only has she provided excellent advice regarding the thesis, but she has also provided plentiful light-relief during our morning tea breaks.

Many thanks also to all my colleagues with whom I have enjoyed sharing an office with. They have been extremely kind and supportive despite my incessant "borrowing" of their books and stationary. Morag Leitch has also shown an astounding willingness to help me negotiate any bumps in the road. Her help has been greatly appreciated.

I thank the Scottish Diabetes Research Network Epidemiology Group for providing access to the dataset. Jeremy Walker in particular provided excellent insight into using the dataset. I would also like to extend my gratitude to the MRC Hubs for Trials Methodology Research for funding me and for giving me the opportunity to conduct this piece of research.

On a personal note, I would like to give special thanks to my friends and family, who have very little idea of what it is that I have been busying myself with for the past few years, but have enthusiastically listened to my updates nonetheless.

CONTENTS

1. INTRODUCTION & BACKGROUND 19

1.1 Missing Data in routine data sources 20

1.1.1 Routinely collected data for clinical trials 20

1.1.2 Missing data in routinely collected data 21

1.1.3 Data analysis with missing data 22

1.1.4 Mechanisms of missing data 22

1.1.5 Patterns of missing data 26

1.1.6 Proportion of missing data 27

1.1.7 Methods for handling missing data 28

1.2 Data Source – Scottish Care Information – Diabetes Collaboration (SCI-DC) 29

1.3 Diabetes (DM) 29

1.3.1 Introduction to DM 29

1.3.2 DM and BMI 31

1.4 Study Aims & Implications 31

1.4.1 Objectives 32

1.5 Thesis Outline 33

2. LITERATURE REVIEW OF MISSING DATA METHODS 35

2.1 Introduction 36

2.1.1 Review aims 36

2.1.2 Search Strategy 36

2.2 Simple missing data methods 36

2.2.1 Deletion approaches 36

2.2.2 Single imputation approaches 38

2.3 Methods for MAR data 45

2.3.1 Multiple Imputation 45

2.3.2 Maximum Likelihood Estimation (MLE) 55

2.3.3 Inverse Probability Weighting (IPW) 56

2.3.4 Comparison of MAR Methods 56

2.4 Methods for MNAR data 58

2.4.1 Selection-based Sensitivity Analyses 58

2.4.2 Pattern Mixture-Based Approach 59

2.4.3 Summary 59

2.5 Conclusions 60

3. LITERATURE REVIEW OF THE RELATIONSHIP BETWEEN BMI & MORTALITY 63

3.1.1 Review Aims 64

3.1.2 Search Strategy 64

3.2 BMI and Mortality in Non-DM patients 64

3.3 BMI and Mortality in Patients with type 2 DM 70

3.3.1 Confounders & Effect Modifiers 72

3.4 Trends in BMI in patients with Type 2 DM 74

4. DATA CLEANING & MISSINGNESS IN THE SCI-DC REGISTER 77

4.1 Introduction 78

4.2 Description of dataset 78

4.2.1 Datasets 78

4.2.2 Description of variables 79

4.3 Data cleaning 84

4.3.1 Fixed Variables 85

4.3.2 Repeat Variables 89

4.3.3 BMI 93

4.4 Cohort Characteristics 99

5. PLAUSIBILITY OF MISSING DATA MECHANISMS 103

5.1 Introduction 104

5.2 Mechanisms of missingness & the SCI-DC data 104

5.2.1 MCAR data 105

5.2.2 MNAR data 105

5.2.3 MAR data 106

5.3 Clinicians' views on explanations for missing data 107

5.3.1 Introduction & Methods 107

5.3.2 Recording Procedures 107

5.3.3 Reasons for missing data 108

5.4 Plausibility of MCAR & MAR assumptions - Findings from data 109

5.4.1 Introduction 109

5.4.2 Methods ... 109
5.4.3 Results I: Predictors of missingness ... 112
5.4.4 Results II: Survival and missingness ... 118
5.4.5 Discussion of results from full dataset ... 123
5.5 2004 - 2008 Cohort ... 126
5.5.1 Introduction ... 126
5.5.2 Methods ... 126
5.5.3 Results ... 127
5.5.4 Discussion of results from 2004-2008 subsample ... 131
5.6 Predictors of BMI at date of diagnosis ... 132
5.6.1 Introduction ... 132
5.6.2 Methods ... 132
5.6.3 Results ... 132
5.6.4 Summary ... 140
5.7 Conclusions ... 143
6. DEALING WITH MISSING DATA IN SINGLE BMI MEASUREMENTS ... 145
6.1 Introduction ... 146
6.2 Methods ... 146
6.2.1 Exclusion Criteria ... 147
6.2.2 Methods for handling missing data ... 147
6.2.3 Analysis Model ... 150
6.3 Results ... 151
6.3.1 Exclusion Criteria ... 151
6.3.2 MCAR: Complete case analysis (CCA) ... 155
6.3.3 MCAR: Population Mean Imputation ... 159
6.3.4 MAR: Stochastic Imputation ... 162
6.3.5 MAR: multiple imputation using Chained Equations (MICE) ... 165
6.3.6 MAR: Multiple Imputation using Predictive Mean Matching (PMM) ... 169
6.3.7 MNAR: Sensitivity Analysis using a Pattern-Mixture based Approach ... 172
6.4 Discussion ... 174
6.4.1 Relationship between BMI and all-cause mortality ... 174
6.4.2 Performance of missing data methods ... 177
6.4.3 Problems & limitations ... 182
6.4.4 Conclusions ... 183

7. USING REPEAT MEASUREMENTS OF BMI TO IMPUTE MISSING DATA 185

7.1 Introduction 186

7.2 Description of Data Problem 186

7.2.1 Introduction 186

7.2.2 Number of observations per person 189

7.2.3 Characteristics of patients by number of observations 190

7.2.4 Timing of measurements 194

7.3 Incorporating longitudinal data to improve imputations 195

7.3.1 Introduction 195

7.3.2 Methods 195

7.3.3 Results 199

7.4 Discussion 213

7.4.1 Introduction 213

7.4.2 Relationship between BMI and all-cause mortality 213

7.4.3 Performance of missing data methods 214

7.4.4 Limitations 221

7.4.5 Conclusions 222

8. MISSING DATA IN MULTIPLE VARIABLES 225

8.1 Introduction 226

8.2 Methods 227

8.2.1 Data 227

8.2.2 Data and extent of missingness 228

8.2.3 Predictors of missingness 228

8.2.4 Analysis Model 229

8.2.5 Missing data methods 229

8.3 Results 233

8.3.1 Extent of missingness 233

8.3.2 Predictors of missingness 234

8.3.3 Complete case analysis (CCA) 236

8.3.4 Multiple Imputation using Chained Equations 244

8.3.5 Predictive Mean Matching (PMM) 251

8.3.6 Multivariate Normal (MVN) Multiple imputation 258

8.3.7 All Methods 264

8.4 Discussion 265

8.4.1 Introduction 265
8.4.2 Missing Data in SCI-DC 265
8.4.3 Analysis of Association Between BMI and All-Cause Mortality 268

9. DISCUSSION 273

9.1 Introduction 274

9.2 Findings from missing data 275
9.2.1 Strengths & limitations 279
9.2.2 Implications & future research 280
9.2.3 Summary 282

9.3 Findings from the association between BMI & All-Cause Mortality 282
9.3.1 The obesity paradox 283
9.3.2 Strengths & limitations 289
9.3.3 Implications & future research 292
9.3.4 Summary 293

10. BIBLIOGRAPHY 295

APPENDICES 315

LIST *of* TABLES

Table 4.1: Details of 13 datasets extracted from SCI-DC from 2008 onwards. 80

Table 4.2: Description of variables extracted from SCI-DC to be used in analyses. 81

Table 4.3: Plausibility cut-offs for available continuous variables (obtained from SCI-DC data quality audit) 85

Table 4.4: Data completeness of fixed variables in patients with type 2 DM 86

Table 4.5: Charlson Comorbidity Index (CCI) Scoring System by condition 87

Table 4.6: Frequency distribution of Charlson comorbidity index scores at diagnosis 88

Table 4.7: Completeness of repeat variables based on measurements taken closest to date of diagnosis and within a year of diagnosis date 92

Table 4.8: Median time between diagnosis and closest measurement date for repeatedly measured variables. 93

Table 4.9: Number of available BMI measurements per patient.....94

Table 4.10. Comparison of measurements recorded at diagnosis and measurements recorded before or after diagnosis.....96

Table 4.11: Characteristics of people within full sample and 2004 to 2008 cohort.....100

Table 4.12: Distribution of patients with DM in Scotland by HBA 101

Table 5.1. Factors influencing missingness in BMI at diagnosis in patients with type 2 DM in Scotland.....113

Table 5.2: Relative and absolute measures of survival by groups of missingness.....120

Table 5.3: Absolute and relative measures of survival within defined time-intervals by missingness in BMI at diagnosis 122

Table 5.4: Factors influencing missingness in BMI at diagnosis with type 2 DM between 2004 and 2008 in Scotland.....128

Table 5.5: Relative and absolute measures of survival by groups of missingness in cohort of patients diagnosed with DM between 2004 and 2008.....130

Table 5.6: Clinical and demographic characteristics of study population by categories of BMI at diagnosis with DM 137

Table 5.7: Mean BMI at diagnosis with DM by deprivation status (from SIMD).....139

Table 5.8: Mean BMI at diagnosis with DM by HBA 139

Table 5.9: Summary of relationship of baseline characteristics with BMI at diagnosis and missingness in BMI at date of diagnosis 144

Table 6.1. Proportions of missing data for variables to be included in missing data models 152

Table 6.2. Top 10 patterns of data completeness in variables to be included in imputation models and analysis models 153

Table 6.3. Baseline characteristics of patients included in final CCA by BMI category..... 156

Table 6.4: Estimates of relative hazard of all-cause mortality and age-standardised mortality by categories of BMI using CCA. 157

Table 6.5: Estimates of relative hazard of all-cause mortality and age-standardised mortality by categories of BMI using population mean imputation. 160

Table 6.6: Estimates of relative hazard of all-cause mortality and age-standardised mortality by categories of BMI using stochastic imputation. 163

Table 6.7: Estimates of relative hazard of all-cause mortality and age-standardised mortality by categories of BMI using MICE. 165

Table 6.8: Estimates of relative hazard of all-cause mortality and age-standardised mortality by categories of BMI using PMM. 170

Table 6.9: Estimates of relative hazard of all-cause mortality and age-standardised mortality by categories of BMI using a pattern-mixture based sensitivity analysis. 172

Table 7.1: Number of BMI observations per patient 189

Table 7.2: Characteristics of patients by number of BMI measurement per person 192

Table 7.3: Patterns of observations in relation to date of diagnosis timeframe. 194

Table 7.4: Estimates of relative hazard of all-cause mortality and age-standardised mortality by categories of BMI using closest observation imputation. 200

Table 7.5: Agreement between closest measurement recorded before diagnosis and measurement recorded at diagnosis with DM. 203

Table 7.6: Agreement between closest measurement recorded after diagnosis and measurement recorded at diagnosis with DM. 205

Table 7.7: Estimates of relative hazard of all-cause mortality and age-standardised mortality by categories of BMI using extrapolated BMI imputation 206

Table 7.8: Agreement between categories of extrapolated BMI and BMI at diagnosis.....208

Table 7.9: Estimates of relative hazard of all-cause mortality and age-standardised mortality by categories of BMI using multiple imputation (+extrapolated BMI). 209

Table 7.10: Estimates of relative hazard of all-cause mortality and age-standardised mortality by categories of BMI using two fold fully conditional specification 211

Table 8.1: Completeness of variables included in models of substantive interest.....234

Table 8.2: Estimates for relationship between categories of BMI at diagnosis and all-cause mortality using CCA (All participants).....240

Table 8.3: Estimates for relationship between categories of BMI at diagnosis and all-cause mortality using CCA, stratified by smoking status.....241

Table 8.4: Estimates for relationship between categories of BMI at diagnosis and all-cause mortality using CCA, stratified by presence of comorbid conditions 243

Table 8.5: Estimates for relationship between categories of BMI at diagnosis and all-cause mortality using multiple imputation using chained equations (All participants) 248

Table 8.6: Estimates for relationship between categories of BMI at diagnosis and all-cause mortality using multiple imputation using chained equations. Estimates stratified by smoking status.....249

Table 8.7: Estimates for relationship between categories of BMI at diagnosis and all-cause mortality using multiple imputation using chained equations. Estimates stratified by presence of comorbid conditions.....250

Table 8.8: Estimates for relationship between categories of BMI at diagnosis and all-cause mortality using PMM (All participants) 252

Table 8.9: Estimates for relationship between categories of BMI at diagnosis and all-cause mortality using PMM, estimates stratified by smoking status 253

Table 8.10: Estimates for relationship between categories of BMI at diagnosis and all-cause mortality using PMM, estimates stratified by presence of comorbid conditions.....254

Table 8.11: Estimates for relationship between categories of BMI at diagnosis and all-cause mortality using MVN model (All participants).....259

Table 8.12: Estimates for relationship between categories of BMI at diagnosis and all-cause mortality using MVN model, estimates are stratified by smoking status. 260

Table 8.13: Estimates for relationship between categories of BMI at diagnosis and all-cause mortality using MVN model, estimates are stratified by presence of comorbidities. 261

LIST *of* FIGURES

Figure 1.1: Four common patterns of missing data. (From Enders, 2010) 27

Figure 2.1: Mean imputation scatterplot of x and y. (From Schafer & Graham, 2002).....40

Figure 2.2: Regression imputation scatterplot of x and y. (From Schafer & Graham, 2002) 41

Figure 2.3. Stochastic regression imputation scatterplot of x and y. (From Schafer & Graham, 2002) 42

Figure 2.4: Hot-deck imputation scatterplot of x and y. (From Schafer & Graham, 2002).....44

Figure 3.1 All-cause mortality versus BMI for each sex in the range 15–50kg/m² (excluding the first 5 years of follow-up). (From Prospective Studies Collaboration, (2009)) 65

Figure 4.1: Data cleaning process for repeatedly measured biochemical characteristics in SCI-DC datasets 90

Figure 4.2: Production of predicted BMI in a patient with type 2 DM.....97

Figure 4.3: Illustration of potential problems associated with using individual level linear regression to predict BMI at date of diagnosis.....98

Figure 5.1: Unadjusted odds ratios (& 95% confidence intervals) of having missing BMI at diagnosis by HBA.....116

Figure 5.2: Adjusted ORs for having missing BMI values at date of diagnosis by categories of BMI group (from median ever-recorded BMI). 117

Figure 5.3: Kaplan Meier plot of the probability of survival by groups of missingness.....119

Figure 5.4: Log-log plot for testing the proportional hazards assumption.....119

Figure 5.5: Frequency of deaths amongst patients with DM by calendar year.....123

Figure 5.6. Scatterplots of BMI at diagnosis with DM against (a) Age at diagnosis, (b) Follow-up time, (c) Charlson Comorbidity Index, (d) Individual-median BMI 134

Figure 5.7: Scatterplots of BMI at diagnosis with DM with (a) Systolic BP (b) Diastolic BP (c) Glycated Haemoglobin (d) Creatinine 135

Figure 5.8: Scatterplot of BMI at diagnosis with DM against (a) Total cholesterol (b) Triglycerides (c) HDL-cholesterol.....136

Figure 5.9: Box and whiskers plot of BMI at diagnosis by year of diagnosis 140

Figure 6.1: Full case-flow of excluded patients 154

Figure 6.2. a) Hazard ratio estimates for risk of all-cause mortality and b) deaths per 1,000 person-years by categories of BMI at diagnosis of type 2 DM in CCA..... 158

Figure 6.3: Comparison of a) adjusted hazard ratio estimates for risk of all-cause mortality and b) deaths per 1,000 person-years by categories of BMI using CCA and population mean imputation..... 161

Figure 6.4: Comparison of a) adjusted hazard ratio estimates for risk of all-cause mortality and b) deaths per 1,000 person-years by categories of BMI using CCA and stochastic imputation.164

Figure 6.5: Comparison of a) hazard ratio estimates for risk of all-cause mortality and b) deaths per 1,000 person-years by categories of BMI using CCA and MICE..... 166

Figure 6.6: Absolute mortality by categories of BMI in each imputed dataset..... 168

Figure 6.7: Comparison of a) hazard ratio estimates for risk of all-cause mortality and b) deaths per 1,000 person-years by categories of BMI using CCA and multiple imputation using PMM. 171

Figure 6.8: Comparison of a) hazard ratio estimates for risk of all-cause mortality and b) deaths per 1,000 person-years by categories of BMI using CCA and a pattern-mixture based sensitivity analysis..... 173

Figure 7.1: Pattern of BMI measurements in relation to date of diagnosis in a random selection of 10 patients..... 188

Figure 7.2: Box and whiskers plots of number of measurements by a. Deprivation status (1=Most deprived) and b. Year of diagnosis..... 193

Figure 7.3: Exclusion criteria for analysis involving use of longitudinal data to impute missing data..... 196

Figure 7.4: Comparison of a) Estimates of relative hazard of all-cause mortality and b) deaths per 1,000 person-years by categories of BMI at diagnosis of diabetes after using closest observation imputation and CCA..... 201

Figure 7.5: Comparison of a) hazard ratio estimates for risk of all-cause mortality and b) deaths per 1,000 person-years by categories of BMI after using CCA and extrapolated BMI imputation. 207

Figure 7.6: Comparison of a) hazard ratio estimates for risk of all-cause mortality and b) deaths per 1,000 person-years by categories of BMI after using CCA and multiple imputation..... 210

Figure 7.7: Comparison of a) hazard ratio estimates for risk of all-cause mortality and b) deaths per 1,000 person-years by categories of BMI after using CCA and two-fold fully-conditional specification.....212

Figure 8.1: Probability of having missing data by Charlson Comorbidity index. Stratified by vital status.....235

Figure 8.2: Model I estimates for the relationship between categories of BMI and all-cause mortality with stratification.....238

Figure 8.3: Model II estimates for the relationship between categories of BMI and all-cause mortality with stratification.....239

Figure 8.4: A comparison of absolute and relative estimates for the association between categories of BMI and all-cause mortality using CCA and multiple imputation using chained equations.....246

Figure 8.5: A comparison of relative estimates for the association between categories of BMI and all-cause mortality using CCA and multiple imputation using chained equations, stratified by smoking status.....247

Figure 8.6: A comparison of relative and absolute estimates for the association between categories of BMI and all-cause mortality using CCA and multiple imputation using PMM.....255

Figure 8.7: A comparison of relative estimates for the association between categories of BMI and all-cause mortality using CCA and multiple imputation using PMM, stratified by smoking status at diagnosis.....256

Figure 8.8: A comparison of relative estimates for the association between categories of BMI and all-cause mortality using CCA and multiple imputation using PMM, stratified by comorbidities.....257

Figure 8.9: comparison of relative and absolute estimates for the association between categories of BMI and all-cause mortality using CCA and MVN model.....262

Figure 8.10: A comparison of relative estimates for the association between categories of BMI and all-cause mortality using CCA and multiple imputation using MVN model, stratified by smoking status.....263

Figure 8.11: A comparison of relative estimates for the association between categories of BMI and all-cause mortality using four different methods for handling incomplete data.....264

Figure 8.12: A comparison of absolute estimates for the association between categories of BMI and all-cause mortality using four different methods for handling incomplete data.....265

ABBREVIATIONS

BMI	Body mass index
BP	Blood pressure
CCA	Complete case analysis
CCI	Charlson Comorbidity Index
CI	95% Confidence interval
DM	Diabetes mellitus
GPs	General Practitioners
HBA	Health Board Area
HbA1c	Glycated haemoglobin
HDL	High density lipoprotein
HR	Hazard ratio
ICD	International Classification of Disease
IQR	Interquartile Range
ISD	Information Services Division
LDL	Low density lipoprotein
MAR	Missing at Random
MCAR	Missing Completely At Random
MI	Multiple imputation
MNAR	Missing Not At Random
mv	Multivariate
MVN	Multivariate normal
NHS	National Health Service
NRS	National Records Scotland
OR	Odds Ratio
PMM	Predictive mean matching
QOF	Quality of Outcomes Framework
RCT	Randomised controlled trials
SCI DC	Scottish Care Information - Diabetes Collaboration
SD	Standard deviation
SIMD	Scottish Index of Multiple Deprivation
SMR01	Scottish Morbidity Records
UK	United Kingdom
uni	Univariate

1. INTRODUCTION & BACKGROUND

1.1 MISSING DATA IN ROUTINE DATA SOURCES

1.1.1 ROUTINELY COLLECTED DATA FOR CLINICAL TRIALS

Routinely collected healthcare records are important, information-rich data sources that offer great potential for epidemiological research. These data sources are stored electronically in general practices and hospitals. Recent developments have also allowed the linkage of these datasets to other healthcare-related records, including death registrations and disease registries, therefore enabling the long-term, unobtrusive follow-up of large numbers of patients at a comparatively low cost.

These data sources not only offer huge promise in terms of observational research, including descriptive analyses and development of risk prediction scores, but may also be useful in enhancing and supplementing randomised controlled trials (RCTs). RCTs are widely accepted as the best way to reliably assess treatment effects. However, they are expensive, time-consuming and are often criticised for their lack of external validity and are often limited by ethical considerations (Lewsey et al., 2000; Rothwell, 2005). By utilising routinely assembled health records, it may be possible to improve the quality, efficiency and cost-effectiveness of RCTs in a number of ways. Firstly, routinely collected data may be used to reduce sampling bias (Staa et al., 2012). As a consequence of strict eligibility criteria, trials are regularly conducted on unrepresentative samples and are therefore unable to address questions related to the interventions efficacy in standard clinical care. Routinely collected data may improve recruitment processes by identifying and enrolling large numbers of patients quickly, rather than depending on patients attending a specific clinic. This process may be especially useful in trials conducted in patients with rare conditions. In addition, patients who were not enrolled could then easily be compared to enrolled patients to assess their representativeness (Lewsey et al., 2000).

Secondly, routine data may enhance the feasibility of studies by providing improved estimates of treatment effects, thus leading to the production of more reliable and accurate power calculations.

Finally, the use of linked electronic health records could allow for the long-term follow-up of patients enrolled in RCTs trials, whilst allowing for the assessment of patient outcomes in routine clinical care, thus reducing the need for manual data collection. Future RCTs would be made considerably cheaper and less burdensome for both the patient and the clinician. Indeed, there have already been several examples of linked healthcare registers being successfully used as tools for long-term follow-up of patients and measuring patient outcomes or adverse events (Ford et al., 2007; Hemminki et al., 2008; Dugas et al., 2010)

1.1.2 MISSING DATA IN ROUTINELY COLLECTED DATA

Despite the huge potential for routinely assembled data both for observational studies and to enhance the performance of RCTs, the practice of making use of this data has been limited by two major concerns. Firstly, concerns surrounding ethics approval have proven to be significant barriers to the use of routinely collected data for clinical trials. Secondly, data quality issues, such as missing data have also proven to be considerable challenges in the utility of these data sources. It is this feature which provides the motivation for this thesis.

Though unobserved data are a well-recognised problem throughout medical research, missing data are a particularly substantial problem in studies which use data from clinical databases (Horton & Kleinman, 2007; Marston et al., 2010). This susceptibility to missing data arises from the fact that data collection and recording is typically clinically driven rather than for specific research purposes. The availability of health indicators at specific time-points in clinical databases is therefore largely dependent on their relevance to the clinical care of the patient. Motivations for recording data are likely to vary over time and procedures are likely to differ across large areas. Accordingly, available data are likely to be less reliable than when data are recorded specifically for the purpose of research, are likely to be available only intermittently and there are likely to be substantial differences between patients with and without missing data (Marston et al., 2010). The introduction of the Quality and Outcomes Framework (QOF) in 2004 may have improved this situation. This programme is a National Health Service (NHS) care delivery plan which offers financial incentives to general practices (GPs) to improve and standardise the care of patients with chronic conditions such as diabetes or asthma.

Since its introduction the extent of missing data in key healthcare indicators within one year of registration with GP has gradually improved in primary care databases (Szatkowski et al., 2011; Bhaskaran et al., 2013; Marston et al., 2014). Despite these improvements, missing data in key health indicators are an inevitable problem. To fully exploit the usefulness of electronic health records, methods need to be sought which are capable of overcoming the problems associated with unobserved data.

1.1.3 DATA ANALYSIS WITH MISSING DATA

Where missing data are not appropriately accounted for, the validity of a study's findings are likely to be compromised. In general, missing data can undermine the validity of study conclusions by reducing sample sizes and introducing bias (Rubin, 1987). The default method for conducting research with incomplete data is to restrict analyses to participants with complete data, known as complete case analysis (CCA) (Graham, 2009). Consequently, along with decreasing study sample size, this method for data analysis will reduce study efficiency through the discarding of information. This problem is most pronounced in multivariate analyses and ultimately leads to reductions in statistical power for hypothesis tests.

By excluding patients with unobserved data, CCA can also introduce bias as a result of the presence of systematic differences between patients with observed and unobserved data. This in turn leads to the production of invalid parameter estimates and therefore incorrect inferences (Little & Rubin, 1987).

The extent of these problems and the respective appropriateness of any missing data method are dependent on the plausibility of the assumptions each technique makes, the proportion of unobserved data, the pattern of missing data and the research question (Carpenter & Kenward, 2007).

1.1.4 MECHANISMS OF MISSING DATA

Many researchers mistakenly believe that the proportion of missing data is the most important criterion on which to base the decision on the most appropriate method for handling unobserved data. In fact, probably the most important determining factor of

the suitability of a missing data approach is the mechanism by which missing data came about, known as the mechanism of missingness.

This is a fundamental concept in missing data methodology and describes the relationship between observed variables and the probability of incomplete data. It is formally defined as the probability that a set of values are missing given the values taken by the observed and missing observations (Little & Rubin, 1987). Mathematically this is:

$$\Pr (R \mid \mathbf{y}_{obs}, \mathbf{y}_{mis})$$

Whereby y represents the data that was intended to be collected. Y_{obs} is observed and Y_{mis} is missing. R denotes a missing value indicator which equals 1 if Y is observed and 0 if Y is missing.

In 1976, Rubin and colleagues developed a classification system to broadly define the three mechanisms by which missing data may occur. Within Rubin's classification system there are three categories; Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR) (Little & Rubin, 1987).

1.1.4.1 MCAR

Data are defined as MCAR when the probability of an observation being missing does not depend on observed or unobserved measurements (Donders et al., 2006). Mathematically this can be described as:

$$\Pr(\mathbf{r} \mid y_{obs}\, y_{mis}) = \Pr(\mathbf{r})$$

Put more simply, the likelihood of a single observation being missing is influenced by chance alone. For example, some patients may have a missing blood cholesterol value as a consequence of the laboratory technician dropping test tubes containing blood samples. This is an example of data that are MCAR since this event could have happened to any patient's blood cholesterol measurement, regardless of the patients observed or unobserved data. When using CCA, this mechanism is likely to introduce the lowest level of bias since no systematic differences between patients with and without observed data should exist. Nonetheless, the results are less precise than if the data were complete because of reduced power.

This is perhaps the strictest mechanism of missingness assumption as missingness cannot be related to any variables under study. Consequently, in practice this assumption is likely to be violated (Baraldi & Enders, 2010). One situation in which this assumption may be satisfied is when data are missing by design (Allison, 2000). For example, researchers may not measure a biomarker for a random 25% of patients due to the costs associated with measurement. Data that are not recorded are MCAR.

From the observed data it is not possible to show that data are MCAR; it is only possible to rule this assumption out. While no differences between patients with and without complete data is consistent with MCAR data, it is not possible to verify that differences do not exist by some unmeasured characteristic. However, systematic differences between patients with and without complete data indicate a clear violation of the MCAR assumption (Little & Rubin, 1987).

1.1.4.2 MAR

Data can be described as MAR when given the observed data, the missingness mechanism does not depend on the values of the unobserved data (Little & Rubin, 1987).

$$\Pr(r \mid y_o, y_m) = \Pr(r \mid y_o)$$

The cause of the missing data is therefore unrelated to the missing values but may be related to the observed values of other variables. For instance, differences may be observed in the proportion of missing blood pressure values across age groups since younger patients are less likely to have blood pressure values recorded. Many modern statistical methods such as multiple imputation and maximum likelihood approaches make the assumption that data are MAR. These methods work by conditioning analyses on the observed data so that missingness becomes completely random. Accordingly, the MCAR assumption is a special case of MAR data. These mechanisms are also termed ignorable missingness.

If the missing data model does not control for the causes of missingness then missingness will become the third mechanism of missingness: MNAR (Graham, 2009). Continuing with the above example where missingness in systolic blood pressure was related to patient age, failure to include age in the missing data model would lead to data

becoming MNAR. It is possible to make the MAR assumption more plausible by collecting more potential explanatory variables for the missingness. By doing so, researchers increase the likelihood of identifying variables which predict missingness in incomplete variables. However, the inclusion of more variables ultimately increases the likelihood of missing data and therefore a strong theoretical rationale is required to justify this approach (Enders, 2010).

1.1.4.3 MNAR

Data are MNAR when even after accounting for all the available observed information, the reason for observations being missing remains unexplained. Missingness may therefore be related to some unmeasured data, such as those of the missing values themselves or of some unmeasured characteristic.

$$\Pr(r \mid y_o, y_m) = \Pr(r \mid y_m)$$

As an example of the former, study participants who have been attempting to lose weight but have failed to do so may be more likely to miss appointments or not disclose their weight. Subsequently, these patients may have more missing values (Gadbury et al., 2003). This form of missing data is most common in clinical trials where patients drop out due to poor health. Alternatively, patients may have missing data on activity levels since their GP may not routinely record this information. If data are not available on which GP the patient attended then this data are MNAR. This form of missingness may be termed non-ignorable since the missing data mechanism must be modelled to obtain valid parameter estimates (Bennett, 2001).

Dealing with MNAR data is difficult and the options to do so are limited. Since most mainstream approaches such as CCA are not appropriate where data are MNAR, some methodologists have attempted to develop methods specific to this assumption. These methods include selection models and pattern mixture models (Abraham & Russell, 2004). However, these techniques are difficult to implement and tend to make dubious and untestable assumptions (Baraldi & Enders, 2010). As an alternative, some researchers have recommended addressing MNAR data by using sensitivity analyses to explore the robustness of parameter estimates to departures from the MAR or MCAR assumptions (Sterne et al., 2009).

In reality, it is likely that all three mechanisms are at work within any given dataset and their presence is dependent on the analysis being carried out and which variables are included. As previously discussed, it is only possible to disprove the presence of MCAR data. Information from people involved in data collection is likely to provide the greatest insight into data collection processes and the probable mechanism of missingness, a process which may not be possible when routine data are used.

1.1.5 PATTERNS OF MISSING DATA

The pattern of missingness is an important assumption made by a number of missing data methods. It refers to the configuration of observed and missing values in the dataset. More simply, the missing data pattern describes the location of the missing values but does not attempt to explain the processes leading to the missingness (Enders, 2010). Identification of the pattern of missingness and more specifically, the clarity of the pattern enables researchers to determine how systematic the process of missing data production is. Some missing data strategies are more effective with specific missing data configurations.

There are several different patterns of missingness including general, univariate, monotone and unit nonresponse (Figure 1.1) (McKnight et al., 2007). The most common missing data pattern is the general pattern, also known as arbitrary or non-monotone missingness. This configuration exhibits no clear pattern of unobserved data and missing data appear to be dispersed randomly across the dataset (Figure 1.1, D). In longitudinal data, this may mean patients have missing data for some but not all occasions. Nonetheless, this apparent randomness does not necessarily mean that no relationship between missing values and some other variable exists. The univariate configuration occurs when missing values are restricted to a single variable (Figure 1.1, A). The monotone data pattern is typical of longitudinal studies and occurs when patients drop out of the study. This pattern of missingness is present when the data can be organised in a way so that each successive variable has fewer observed values (Figure 1.1, C) (Rubin, 2009). Finally, the unit non-response patterns occur when information for every participant is available for certain variables, but for other variables completeness is not achieved (Figure 1.1, B).

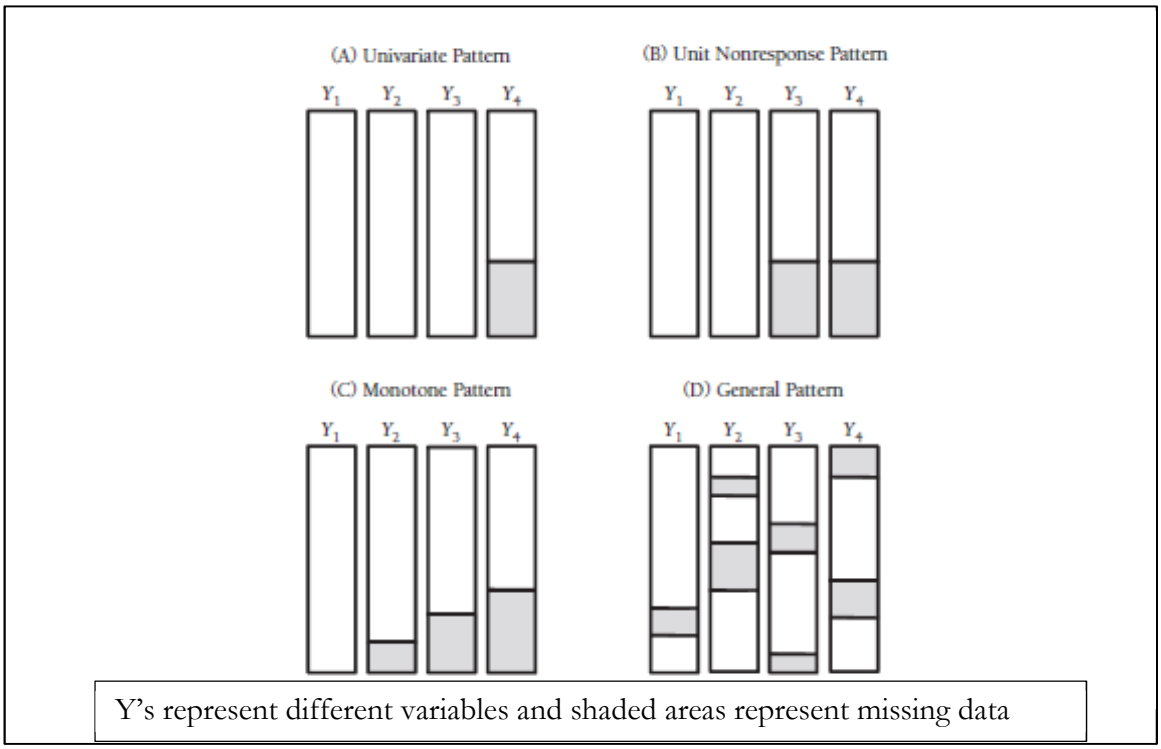

Figure 1.1: Four common patterns of missing data. (From Enders, 2010)

Though determining the missing data configuration is a useful data exploration exercise, its importance is lessening due to the development of flexible missing data methods capable of handling all missing data patterns (Enders, 2010).

1.1.6 PROPORTION OF MISSING DATA

The effectiveness of any given missing data method is dependent on the proportion of unobserved data within the dataset. Previous simulations have generally shown that most techniques tend to perform worse with increasing incompleteness (Janssen et al., 2010; Langkamp et al., 2010). However, there is not currently an established cut-off point for the extent of missing data that is tolerable, though Schafer (1999) stated that 5% missingness was likely to have a negligible impact on parameter estimates, regardless of the approach used.

The impact the proportion of missing data can exert on estimates will depend critically on the underlying missing data mechanism. For example, the error introduced into estimates has been shown to be considerably worse when data are MNAR, even when the proportion of missing data is small. The bias introduced will also depend on the

context of the study. For instance, if missing data occur predominantly in patients who have experienced a rare event, the impact on analyses may be more striking.

1.1.7 METHODS FOR HANDLING MISSING DATA

To overcome the problems associated with missing data a wide variety of methods are commonly employed, some with more attractive features than others. An appropriate statistical analysis with missing data should aim to achieve two important objectives (Graham, 2009). Firstly the missing data technique should be capable of producing unbiased parameter estimates which closely reflect the true population value. These estimates should remain consistent if the analysis were repeated. Secondly, the technique should produce valid standard errors and confidence intervals to allow for the assessment of the uncertainty of the estimate given the presence of missing data.

Unfortunately, the most frequently applied techniques make strong assumptions regarding the mechanism of missingness and are likely to introduce significant bias into the study whilst understating the variability of the data (Sterne et al., 2009). Included amongst these popular techniques are a range of simple approaches such as CCA and single imputation where missing values are replaced with values obtained from the observed data, such as the sample mean. Despite some widely reported flaws, the use of these techniques remains widespread, a possible reflection of their ease of use and accessibility in popular statistical software (Klebanoff & Cole, 2008).

More recently, there has been the evolution of more advanced techniques such as maximum likelihood methods and multiple imputation, both of which are capable of producing unbiased estimates under weaker assumptions. Though these approaches have been available for over 30 years, the use of these techniques is only now becoming more widespread, a movement which may be attributed to a number of software developments.

A more thorough description of these methods is provided in the literature review presented in Chapter 2.

1.2 DATA SOURCE – SCOTTISH CARE INFORMATION – DIABETES COLLABORATION (SCI-DC)

The data for this project comes from the existing SCI-DC register, a population based register of people with a diagnosis of diabetes mellitus (DM) in Scotland. This register was established in 2000 when the Scottish Diabetes Framework recognised the need for an effective information technology system to facilitate the delivery of well-managed, integrated DM care. To achieve this, SCI-DC was formed with the aim of developing a Scottish diabetes register which collated information for almost all patients with diagnosed DM in Scotland. SCI-DC subsequently expanded existing data systems in the Tayside and Lanarkshire health board areas (HBAs) to the remaining 12 HBAs and now, 14 years later, the DM register achieves coverage of over 99.5% of patients with a diagnosis of DM in Scotland. This translated to approximately 230,000 still-living patients and a further 50,000 deceased patients in the extracted database in 2008. The register assembles data on numerous demographic and clinical characteristics from a wide range of primary and secondary care sources. In recent times, register data have also been linked to several other healthcare records such as mortality data, prescription data and renal and cancer registrations.

With such a wealth of information, the SCI-DC register represent a hugely valuable resource that can be utilised to improve service and resource planning as well as offering great potential for epidemiological research. Unfortunately, the registers utility for epidemiological research is often compromised by non-trivial proportions of incomplete and poor quality data. Accordingly, if research is to exploit this large information-rich register, approaches need to be developed to draw reliable conclusions from the data despite the presence of missing data.

1.3 DIABETES (DM)

1.3.1 INTRODUCTION TO DM

DM represents a profound public health concern across Scotland. According to the 2013 Scottish Diabetes Survey, there were 268,154 patients on the DM register at the

end of 2013, representing a crude prevalence of 5.1%, up from 4.9% in the previous year (Scottish Diabetes Survey Monitoring Group, 2013). This increasing prevalence of DM may reflect the ageing demography of Scotland, improved survival times, the obesity epidemic and a lowered average age of onset of type 2 DM (Scottish Intercollegiate Guidelines Network, 2010). Consequently, the challenge of maintaining and improving the standard of DM care in a country where DM prevalence is increasing represents a substantial problem to healthcare providers.

DM is a chronic metabolic disease characterised by an inherited or acquired deficiency in insulin production, or by the insensitivity of cells to any insulin produced. Insulin is an essential hormone responsible for triggering the absorption of glucose into cells and therefore regulating blood glucose levels. In patients with type 1 DM, insulin cannot be produced by the pancreas and therefore the process of glucose absorption is restricted (Diabetes UK Scotland, 2010). In patients with type 2 DM, either the pancreas produces only a limited amount of insulin or the body develops a resistance to the action of insulin. In the event of resistance, the body is said to be insulin insensitive, a phrase which describes the body's inability to respond appropriately to the presence of insulin (Stevens, 2009). As a consequence, DM is typically characterised by high blood sugar levels (hyperglycaemia).

DM and its associated hyperglycaemia is a source of considerable excess morbidity and premature mortality in Scotland. Indeed, the Scottish Diabetes Research Network (2010) found that between 2001 and 2007, mortality was between 10 and 25 per cent higher in patients with DM than those without. The majority of these deaths can be accounted for by significant macro- and micro- vascular problems caused by hyperglycaemia. One study estimates that macro-vascular deaths such as cardiovascular disease (CVD) and stroke are responsible for 70% of all deaths in patients with DM (Fisher, 2006). Patients with DM are also susceptible to numerous micro-vascular complications such as retinopathy, nephropathy and neuropathy.

To overcome this problem, patients are advised to attempt to manage their blood glucose levels through exercise and diet. Where this is not possible, patients are prescribed anti-hyperglycaemic agents such as insulin, metformin, sulfonylureas or thiazolidinediones (TZDs). Consequently, DM care has been estimated to account for

approximately 10% of all NHS expenditure and therefore contributes a significant burden to healthcare providers (Diabetes UK Scotland, 2010).

1.3.2 DM AND BMI

Excess adiposity is rapidly becoming a major global public health problem with the prevalence of obesity and overweight reaching epidemic proportions. In Scotland alone it is estimated that 68.5% of men and 61.8% of women are overweight or obese (defined as 25 to $<30 kg/m^2$ and $\geq 30 kg/m^2$ respectively, according to the World Health Organisation's international classification system), at an annual cost to the NHS of £171 million (Scottish Intercollegiate Guidelines Network, 2010).

Based on cross-sectional data, it has been widely recognised that raised patient BMI is a significant proven risk factor for many adverse health outcomes, regardless of DM status (Scottish Intercollegiate Guidelines Network, 2010). In patients without a DM diagnosis, raised BMI has been associated with an increased risk of CVD, liver disease, osteoarthritis, depression and numerous cancers. Importantly, the risk of developing type 2 DM has also been widely demonstrated to increase with increasing BMI (Chan et al., 1994) (Vazquez et al., 2007). For this reason, the majority of patients with type 2 DM are overweight or obese.

BMI is therefore an important covariate in many analyses involving patients with DM. Unfortunately in the SCI-DC register, completeness levels are low (See chapter 4) and it was deemed beneficial to identify methods for handling unobserved BMI data. Furthermore, current studies investigating the association between BMI and adverse health outcomes fail to adequately assess the impact of missing data on the study's respective findings.

1.4 STUDY AIMS & IMPLICATIONS

This project was multi-disciplinary in nature, occupying overlapping aspects of epidemiology and statistics. Its primary focus was methodological and it attempted to document approaches and challenges to handling unobserved data in a routine data source. These problems and potential solutions are described plainly so that the work

remained accessible to epidemiologists who may typically be deterred from using more appropriate but complicated methods for handling incomplete data. The criteria upon which methods for handling incomplete data were chosen was largely based on their accessibility in general purpose statistical software such as STATA, SAS or R. Flexible missing data suites have been developed in each of these software packages (Horton & Kleinman, 2007). Despite further suites being available in alternative software programmes including WinBUGS and MLWIN, data analyses were restricted to the most commonly used packages as it was deemed unlikely that the majority of epidemiologists would choose to use alternative software for missing data handling purposes. It aims to provide practical guidance and describe issues encountered when these methods are applied in routine data sources. This work will also help to assess the feasibility of using routinely-collected databases for research using observational and interventional study designs.

The project's secondary focus was epidemiological and endeavoured to ascertain the association between BMI at diagnosis with type 2 DM and all-cause mortality in the presence of missing data. Cox proportional hazards models were applied to explore this relationship.

In addition, this thesis aimed to explore missing data problems specific to the SCI-DC register. Missingness in this database was examined and reported, highlighting variables where improvements in data completeness should be sought.

1.4.1 OBJECTIVES

To achieve the study aims the following objectives were formulated:

1. To conduct a thorough literature review to summarise the current available methods for handling incomplete data in a wide range of scenarios. A description of each methods strengths and weaknesses are provided.

2. To summarise the current literature on the relationship between BMI and all-cause mortality in patients with DM. Current analysis strategies to investigate this relationship are documented and important potential confounders and effect modifiers are identified.

3. To document the extent of the missing data problem in the SCI-DC register. Proportions of missing data relative to date of diagnosis are observed, with a particular focus on missingness in the BMI at diagnosis variable

4. To identify the plausibility of the separate mechanisms of missingness by conducting informal interviews with clinicians and data managers.

5. To use the data to explore the plausibility of the separate mechanisms of missingness by examining differences between patients with and without complete data, conducting logistic regression analyses to identify important predictors of missingness in BMI at diagnosis variable. Conducting survival analyses to determine if missingness in BMI at diagnosis was associated with the outcome.

6. To identify important predictors of BMI at diagnosis variable which may be included in imputation models to improve predictions of missing values

7. To use information from points 4,5 and 6 to choose appropriate methods for handling missing data and apply these methods to an analysis of the association between BMI and all-cause mortality in the presence of univariate missingness. Estimates of this relationship obtained from each missing data method are compared.

8. To identify patterns of BMI recording and ascertain the plausibility of using longitudinal BMI measurements to improve imputations. Imputation methods using longitudinal data are applied to an analysis between BMI and all-cause mortality.

9. To explore the relationship between BMI and all-cause mortality when missing data occur in multiple variables. Different methods capable of handling non-monotone missing data are applied.

1.5 THESIS OUTLINE

Chapter 2 presents a literature review of methods for handling missing data and details when these approaches produce valid estimates and their potential limitations.

Chapter 3 presents a literature review of the association between BMI and all-cause mortality in both the general population and in patients with type 2 DM.

Chapter 4 provides a thorough description of the routine datasets that were used throughout this thesis and outlines the thorough data cleaning process that was undertaken.

Chapter 5 explores the extent of the unobserved data in this dataset and attempts to determine the plausibility of the missing data mechanisms for the incomplete BMI variable. To achieve this, findings from informal discussions with a number of clinicians are described and data exploration exercises are presented. More specifically, differences between patients with and without observed data were compared and logistic regression analyses were conducted to ascertain predictors of missing variables.

The relative complexity of the settings to which the missing data methods were applied became progressively more complicated. Chapter 6 applies several missing data methods to an analysis of the association between BMI and all-cause mortality. These approaches were applied in the context of univariate missingness in the BMI at diagnosis variable and did not make use of any longitudinal BMI data.

Chapter 7 builds on the work of Chapter 6 but attempts to improve imputation approaches by incorporating longitudinal patient data to improve the predictive power of imputation models.

Chapter 8 examines the more realistic setting whereby missing data are present in multiple variables. Here, results from three separate multiple imputation models are compared.

In Chapter 9 the thesis is concluded and an overview of the study's main findings is presented along with the important strengths and limitations of the work and finally, potential areas for future research.

2. LITERATURE REVIEW OF MISSING DATA METHODS

2.1 INTRODUCTION

2.1.1 REVIEW AIMS

In an effort to understand current approaches to handling missing data, a literature review was undertaken. This review aimed to provide an insight into important considerations when devising strategies for handling missing data along with outlining current available methods. For each method identified, a description of the methods respective strengths and limitations was reported, as well as the circumstances in which using this technique could result in invalid results. This review is divided into three main sections: simple missing data methods, methods for MAR data and methods for MNAR data.

2.1.2 SEARCH STRATEGY

Conducting a thorough literature review for missing data was a complex process. A search on keywords describing methods for missing data would find many thousands of papers, as it would find clinical papers that merely use the methods to analyse their data – indeed most comprehensive searches would find any paper that mentioned missing data. To overcome this, missing data methodological review articles were sought using Medline, EMBASE and the Web of Knowledge in 2011. Reference lists from reviews and articles citing the reviews were examined to identify further relevant articles. Method-specific searches for missing data approaches were conducted using the above electronic databases. This literature review was updated in 2014. The methods found as a result of this search are summarised in the following narrative review.

2.2 SIMPLE MISSING DATA METHODS

2.2.1 DELETION APPROACHES

Deletion techniques involve removing participants without complete data from analyses. These popular approaches are capable of producing valid estimates when data are MCAR (Horton & Kleinman, 2007). Under this condition, the set of subjects with no unobserved data should form a random sample of the target population. Accordingly,

the removal of participants without complete data should produce a representative sub-sample which can be analysed in the usual way.

List-wise deletion or complete case analysis (CCA) is the most popular missing data method and involves the exclusion of cases with missing information for any studied variable from all analyses. This method is quick and simple to implement and is regularly used when the amount of missing data is small. Under this circumstance, the presence of missing data is unlikely to have a large impact on the overall conclusions. For these reasons researchers typically overlook labour intensive methods and opt for CCA, the default setting in most mainstream statistical packages.

There are scenarios whereby CCA may produce unbiased estimates when data are not MCAR. For example, CCA is robust to violations of MCAR providing missingness on predictors is not dependent on the outcome (Allison, 2000). This point is widely overlooked in missing data reviews, despite it having important implications (Steyerberg & van Veen, 2007).

Unfortunately, there are also a number of scenarios whereby CCA analysis will produce biased parameter estimates when data are not MCAR (Baraldi & Enders, 2010). Commonly, the MCAR assumption does not hold and the resulting sub-samples are no longer representative of the target population. Subsequent analyses produce estimates which do not reflect the true population values (He, 2010). This problem has been confirmed in numerous simulation studies (Wood et al., 2005; van der Heijden et al., 2006; Janssen et al., 2010). In some instances, CCA performed the worst for most population parameter estimates compared to other methods when data were not MCAR. However, the studies also demonstrated that where data were not MCAR and the proportion of missing data was minimal, any bias introduced was negligible. In one simulation study, Janssen et al., (2010) used CCA to overcome MAR missingness in D–dimer values, a predictor of the presence of deep venous thrombosis. Here, CCA only caused bias when the proportion of missing values was 60% or more.

A further drawback of CCA is the potential for significant sample size reductions and resultant losses in study power. This is especially likely in multivariate analyses where missing data in individual variables can combine to produce high rates of case-wise missingness (Klebanoff & Cole, 2008; Janssen et al., 2010). Therefore, not only does

this method act to discard valuable information, it also increases the risk of type II errors.

In the context of clinical trials this method creates additional problems by violating the intention-to-treat principle. This principle asserts that all participants in trials should be analysed in the groups to which they were allocated. Therefore, according to this principle, patients lost to follow-up cannot be excluded despite having unobserved outcomes (Molenberghs et al., 2004).

CCA may also exert a large influence in studies involving rare outcomes, such as childhood cancer (Schafer & Graham, 2002; Abraham & Russell, 2004). In this example, CCA could conceivably discard a large proportion of patients with the rare condition.

A related method for handling missing data is pairwise deletion or available case analysis. This technique involves the use of all available data on an analysis-by-analysis basis resulting in different sample sizes for each individual analysis (Baraldi & Enders, 2010). This makes it difficult to compare related models to each other as analyses are based upon different populations. Though this approach may be considered more efficient than CCA through the reduction of sample size losses, many of the limitations of CCA persist in pairwise deletion. Most important of which is the technique's tendency to produce biased parameter estimates when data are not MCAR.

Furthermore, with this technique analyses are undertaken on different but over-lapping samples of the dataset resulting in analyses being undertaken on significantly different sub-samples of the dataset. This feature can negatively affect the generalisability of any inferences made from the data (Enders, 2010).

2.2.2 SINGLE IMPUTATION APPROACHES

Single imputation is a collection of techniques which involve the replacement of missing values with plausible values. Using the newly complete dataset, the analyses are completed as if there were no unobserved data. A wide range of single imputation techniques are available such as arithmetic mean imputation, last observation carried forward, hot-deck imputation and regression imputation.

Though these approaches have a number of strengths including their simple execution and their ability to account for the observed data, the use of single imputation techniques can very rarely be recommended. Under MCAR, MAR and MNAR conditions, single imputation techniques are likely to produce biased estimates and over-estimate the precision of estimates by ignoring the uncertainty of the imputed values. They can also distort relationships within the data and make implausible assumptions (He, 2010).

2.2.2.1 MEAN IMPUTATION

Mean imputation involves the replacement of missing values with the arithmetic mean value of the variable of interest. In analyses with repeated measurements, this involves imputing the mean value for all non-missing subjects for that specific time-point (Gadbury et al., 2003).

Despite its quick and easy execution, the use of mean imputation is difficult to justify due to a number of shortcomings (Baraldi & Enders, 2010). Firstly, under all missing data mechanisms mean imputation produces biased estimates. Secondly, mean imputation under-represents extreme values leading to the underestimation of variance. This in turn incorrectly attenuates standard errors and increases the risk of observing a type 1 error (Schafer & Graham, 2002). Thirdly, this approach inserts values that are uncorrelated with any other variable, and so correlations between imputed variables and other variables are diluted (Baraldi & Enders, 2010). This problem is illustrated in Figure 2.1 (from Schafer & Graham, 2002). Here the imputed values fall on a straight horizontal line suggesting no relationship between x and y, while the real values show variation around the straight line.

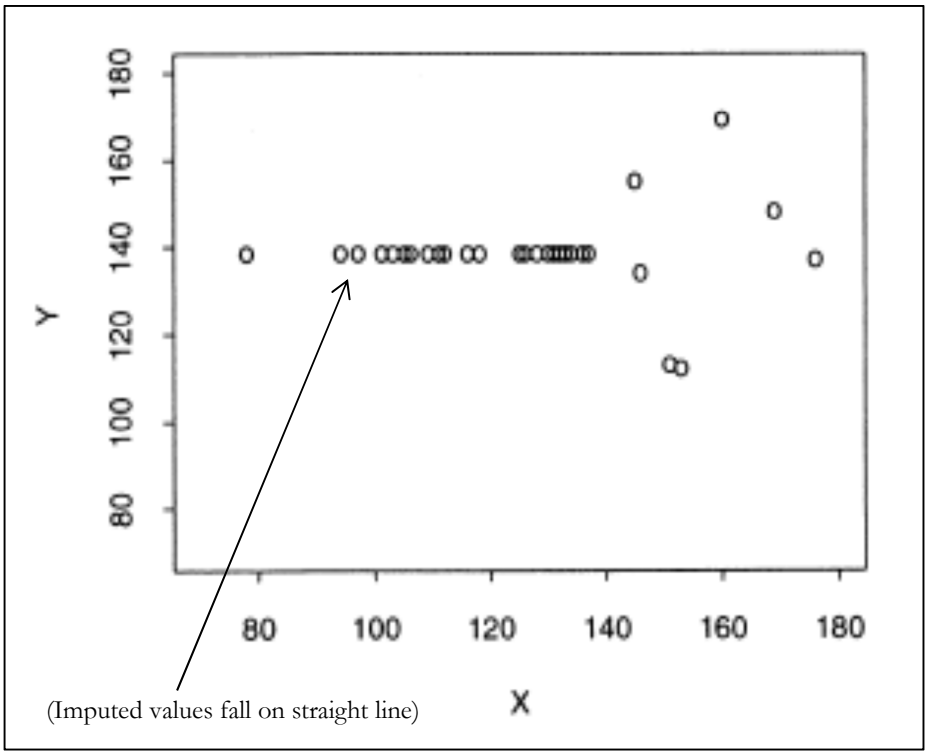

Figure 2.1: Mean imputation scatterplot of x and y. (From Schafer & Graham, 2002)

Finally, where data are non-Normally distributed, the mean may not be the best measure of central tendency of the data and therefore parameter estimates are under or over estimated.

These limitations have been demonstrated in numerous simulation studies (Engels et al, 2003; Schafer & Graham, 2002). In one example, the performance of mean imputation in producing parameter estimates which concur with true population values was compared to that of three other single imputation methods (Schafer & Graham, 2002). Here, mean imputation performed consistently worse than all other methods under all mechanisms of missingness. Indeed, Graham (2012) commented that this is perhaps the worst missing data approach available.

2.2.2.2 Regression Imputation (Conditional mean imputation)

Regression imputation is a technique which replaces missing observations with values obtained from regression equations. This approach uses the observed data to create regression models to predict values of the missing variable(s) (Little & Rubin, 1987). For example, when the *y* variable has missing values and the *x* variable does not, conditional mean imputation fits a model for predicting *y* given *x*. Inserting corresponding values of

x into the newly-formed regression equation will therefore identify predicted values for y. The resulting predictive regression equation is as follows

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 (x_i)$$

where y_i is the predicted value of y for case i. Like other single imputation techniques, the main virtue of this approach is that all information is preserved and study sample size remains intact. However, regression imputation has several limitations which have restricted its utility in medical research. As well as underestimating standard errors, this approach overstates relationships in univariate analyses leading to imputed value falling on a straight regression line creating overly correlated scores (Figure 2.2) (Baraldi & Enders, 2010). In multivariate analyses, this problem is present but to a lesser extent (Enders, 2010).

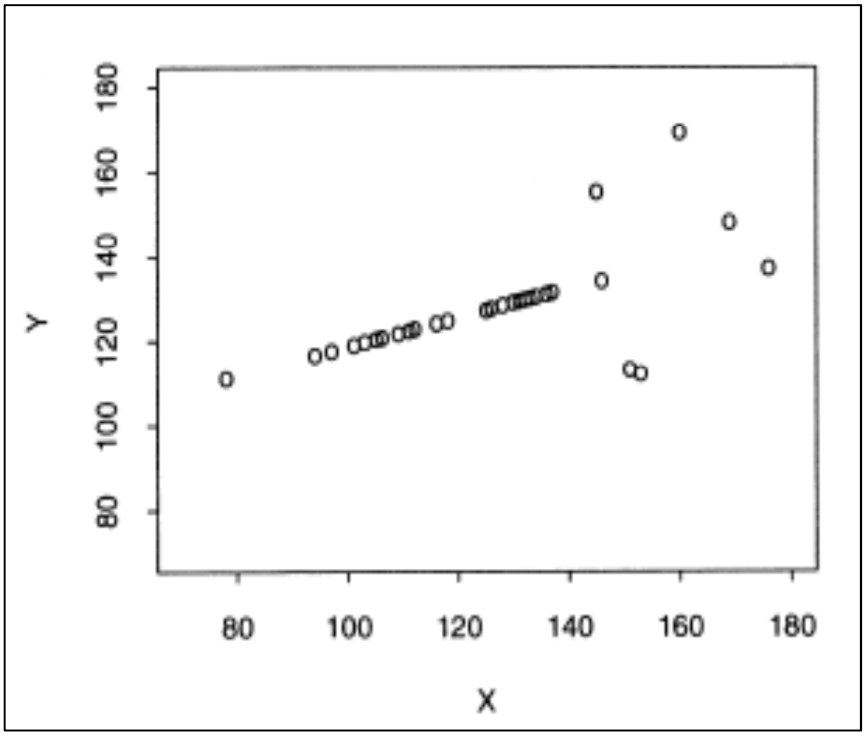

Figure 2.2: Regression imputation scatterplot of x and y. (From Schafer & Graham, 2002)

A variant of regression imputation exists for missing repeated measures data. This technique predicts missing values using a regression equation which models the relationship between variable scores from two separate time-points in subjects with observed data. Values are taken from these regression equations to impute missing data

at one of these time-points. Unfortunately, this technique suffers from the same problems as standard regression imputation for cross-sectional data.

2.2.2.3 STOCHASTIC IMPUTATION

Stochastic imputation attempts to overcome the problem of underestimated variability in regression imputation by adding random error to predicted values (Baraldi & Enders, 2010). The method is identical to regression imputation, but for the addition of a Normally-distributed residual term with a mean of zero and a variance equal to the residual variance. The regression equation would therefore become:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 (x_i) + z_i$$

where z_i is the residual term. By including this residual term this method is capable of restoring lost variability.

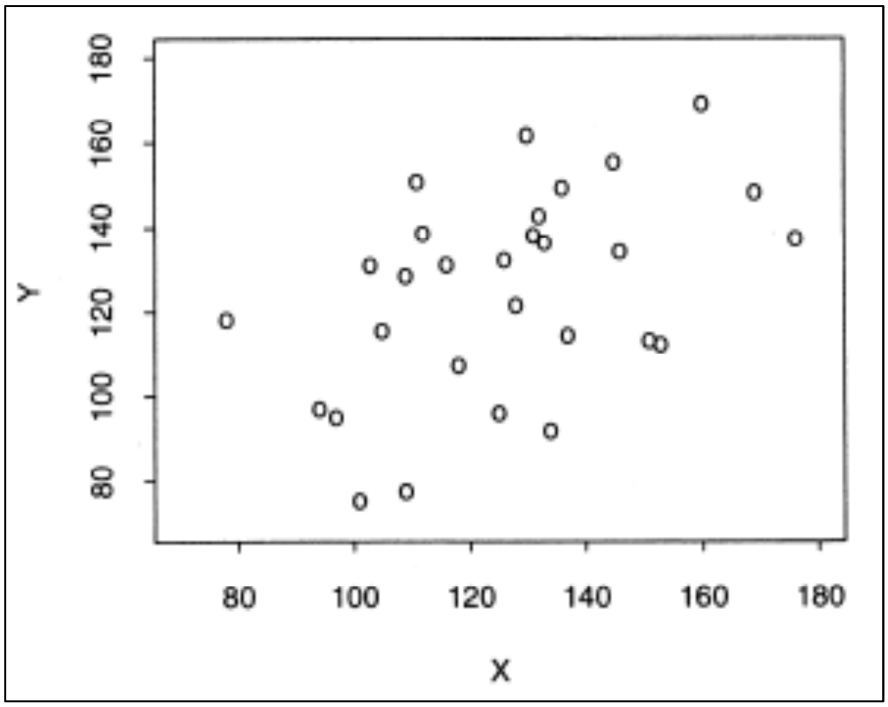

Figure 2.3. Stochastic regression imputation scatterplot of x and y. (From Schafer & Graham, 2002)

Figure 2.3 shows a stochastic regression imputation scatterplot of x and y where the missing data are MAR. Clearly, the imputed values no longer follow a straight line therefore stochastic regression imputation successfully restores lost variability whilst preserving the relationship between the x and y variables. The scatterplot closely

resembles the original scatterplot with complete data. Stochastic imputation is frequently the best performing of each of the single imputation techniques due to its ability to yield unbiased parameter estimates with MAR data (Little & Rubin, 2002; Newman, 2003). Though simulation studies have reported similar estimates to those from more advanced techniques, this technique still incorrectly attenuates standard errors.

2.2.2.4 HOT DECK IMPUTATION

Hot deck imputation is the name of a series of imputation techniques which use responses from similar respondents to fill in missing values (Enders, 2010). In its most basic form, hot-deck imputation will randomly replace missing values with the scores of other respondents. In this form, the likelihood of selecting a single value will depend on the rate of its occurrence amongst respondents. More typically however, hot-deck imputation involves the replacement of missing values with a random value drawn from a subsample of similar respondents. This technique is often referred to as closest match and is easily applied to repeated measurements by using values obtained from one or more participants who have similar scores on the same measure assessed at other time-points (Elliott & Hawthorne, 2005). Alternatively, missing data may be replaced with the 'average closest match' by identifying participants with the lowest sum of absolute differences and using the mean value from these participants for the variable of interest.

With hot deck imputation, the variability of the imputed data can be attenuated but not to the extent shown by other single imputation techniques (Figure 2.4). The approaches may also distort the relationship between variables and Figure 2.4 clearly demonstrates this feature (Schafer & Graham, 2002). Here, the scatterplot should resemble a positive correlation but does not.

Despite this, one simulation study comparing the efficacy of closest match, average closest match, list-wise deletion and last observation carried forward for imputing repeated measures data reported better performances from the two variants of hot-deck imputation than from the other methods (Elliott & Hawthorne, 2005).

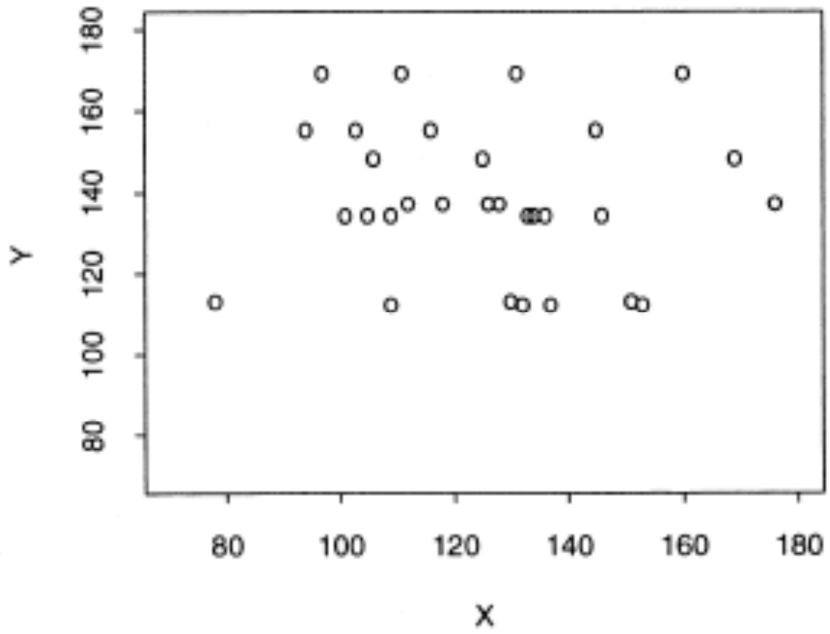

Figure 2.4: Hot-deck imputation scatterplot of x and y. (From Schafer & Graham, 2002)

2.2.2.5 LAST OBSERVATION CARRIED FORWARD

A popular technique for handling missing data in longitudinal studies is last observation carried forward (LOCF). To achieve a complete dataset, LOCF replaces missing values with previous observations. This technique therefore makes the unlikely assumption that the most recent observation for each individual is the best guess for any subsequent unobserved values (Carpenter & Kenward, 2007). This assumption seldom holds and resultant estimates are generally biased (Mallinckrodt et al., 2003). For example, in analyses of treatment effects on progressive diseases, LOCF will produce overly optimistic results for each treatment group. Where missing data occur in an outcome of interest that is likely to deteriorate through time, such as in Alzheimer's, replacing values with previous observations will cause patients to seem healthier than they actually are. If more withdrawals occur in the active treatment arm there will be a larger bias in favour of this group and so an incorrect treatment effect will be identified (European Medicines Agency, 2010).

Variants of the LOCF approach designed to handle unobserved values in the middle of a set of longitudinal data also exist. These techniques are able to take advantage of recorded values taken before and after the missing value. In one example, Engels and Diehr (2003) compared the use of 14 separate single imputation techniques. These 14 separate approaches could be grouped according to the type of data they used to

complete the imputations; 'Population', 'Baseline', 'Before' and 'Before and After'. The population methods used data from the whole population such as population median or mean values to fill in missing values in a specific year. The baseline techniques used other covariates to group individuals into classes before using group mean/median to impute missing values. This group of techniques included hot-deck imputation and regression imputation. The before methods included LOCF and mean or median of the subjects previous known values. Before and after methods used data available from before and after the missing value for imputation. Here, next observation carried backward, mean of the last known and next known values, mean/median of the participant's values before and after missing value were used.

Each imputation technique was found to produce biased results and underestimated variances, regardless of the imputation technique used. However, some techniques such as the population-based methods performed worse than others. The study concluded that those methods which used person-specific longitudinal information were superior to those that did not. More specifically, the authors recommended the use of the method that use the mean or median from the last and next available values for each individual. Despite promising results taken from this simulation study which did not compare these techniques to more advanced approaches, most methodologists actively discourage the use of all forms of LOCF due to its previously discussed unrealistic assumptions (Mallinckrodt et al., 2003; Carpenter et al., 2004; Kenward & Molenberghs, 2009).

2.3 METHODS FOR MAR DATA

There are a number of methods that assume data are MAR; whereby the probability of missing data is conditional on observed data only. The two main methods which are effective when data MAR are multiple imputation and maximum likelihood approaches. These methods are described below.

2.3.1 MULTIPLE IMPUTATION

Multiple imputation is an important modern statistical technique that is rapidly becoming the most widely used "filling-in" method for dealing with unobserved values

(Schafer & Graham, 2002). Unlike single imputation, multiple imputation is able to take into account the uncertainty of the imputed values, whilst producing unbiased parameter estimates. It predominantly assumes MAR data, though there is potential for its use in datasets with MCAR and MNAR data when the MNAR mechanism is known (White & Carlin, 2010).

The technique is multi-step and involves the creation of several plausible datasets using the distribution of observed values, the analysis of each individual imputed dataset and finally the combining of results obtained from each dataset (Sterne et al., 2009). This approach therefore maintains the natural variability of the data whilst preserving relationships between variables (Wayman, 2003). There are a number of advantages of this technique including its ability to produce complete datasets which can be subsequently analysed using familiar data analysis techniques. It is also a fairly general approach whose basic steps may be applied in a number of contexts (Kenward & Carpenter, 2007). Indeed, imputed datasets may even be used for multiple different analyses.

Despite multiple imputation being considered 'state of the art' amongst most methodologists, its use remains low in comparison to CCA. Upon searching 4 high impact journals from 2002 and 2007, Sterne et al (2009), found only 59 articles in which multiple imputation had been used in original research. Similarly, MacKinnon (2010) searched the same 4 high impact journals from their earliest searchable full text articles until the end of 2008 and found only limited use of multiple imputation. It is suggested that this pattern of use could reflect the perceived complexity and multi-step nature of multiple imputation which can prove too time-consuming for some researchers. Nonetheless, the recent advent of statistical software capable of performing multiple imputation has allowed the method to become more accessible to researchers.

2.3.1.1 Stages of Multiple Imputation

To produce estimates from incomplete data, multiple imputation involves three distinct stages, often referred to as the imputation phase, analysis phase and the pooling phase (Baraldi & Enders, 2010).

Imputation Stage

The imputation phase involves the use of statistical models to obtain plausible substitutes for the missing values using the MAR assumption. There are two chief imputation models that have been implemented in standard statistical software, these are the multivariate normal model and the imputation using chained equation approaches.

The multivariate normal (MVN) modelling approach, also named data augmentation and joint modelling was first popularised by Schafer (1997). This approach assumes all variables in the model exhibit a multivariate normal distribution. When imputation models include different types of variables such as continuous and categorical variables the construction of a joint model is difficult.

The MVN imputation model is an iterative procedure which has two distinct steps, an imputation step (I-step) and a posterior step (P-step). The I-step begins by generating starting values for data parameters including means, variances and covariances using the expectation maximisation (EM) algorithm. Briefly and simply, this algorithm attempts to find the maximum likelihood estimates of parameters and uses observed values to generate an initial guess of the parameters. Using the observed data and these initial guesses, the EM algorithm then re-estimates the missing values. This process continues until the estimates converge upon a stable set of estimates. The I-step then regresses variables containing missing values on the complete variables and from these equations missing values are replaced with predicted scored. A random draw from the normal distribution of residuals is augmented to these predicted scores to restore variability. In the P-step, new values of the parameters are simulated by drawing from their posterior distribution and random residual terms are added to each of the resulting estimates. These new parameters are then used to recalculate the regression model in the I-step. This process is repeated until convergence is achieved.

The second most widely used imputation model is the multiple imputation using chained equations (MICE) approach. This method, also termed fully conditional specification, sequential regression and multiple imputation with regression switching was first developed by van Buuren (1999). MICE is a variable by variable approach with separate imputation models for each incomplete variable. To begin with, incomplete

values are replaced with observed values using a simple random sampling with replacement approach (White et al., 2011). The variable with the least missing data is then regressed on all other variables and missing values for this variable are replaced with simulated draws from the corresponding posterior predictive distribution. This is repeated for each incomplete variable to form one cycle. To stabilise results, this procedure is repeated for several cycles to produce a single imputed dataset. The whole procedure is then repeated to produce several imputed datasets. When missing data occur in only one variable, iterations are not required.

Unlike in the MVN approach, MICE can use the most appropriate regression model for imputing different variable types. For example, linear regression may be used in imputation models for normally distributed continuous variables, logistic regression for binary variables and multinomial logistic regression for nominal or ordinal categorical variables. MICE is therefore a more flexible approach that does not make the assumption that data exhibit a multivariate normal distribution. The downside of this approach is that MICE lacks the theoretical basis that MVN imputation has. Accordingly, the conditional distributions may be incompatible making convergence difficult to achieve (White et al., 2011). Both approaches have been found to produce unbiased estimates in a wide range of settings and therefore no general consensus regarding the appropriateness of each imputation model has been reached (Kropko et al., 2014).

A slight alternative to this imputation approach is predictive mean matching (PMM). This approach mimics MICE but instead of replacing missing values with random draws from the corresponding posterior predictive distribution, PMM matches unobserved cases with a pre-specified number of nearest neighbours, known as donors, whose predictive means are closest to that of the unobserved case. Missing values are replaced with observed values from one of the randomly selected donors (Schenker & Taylor, 1996). In this way PMM is similar to hot-deck imputation since it imputes values from nearby complete cases. As a result, PMM imputes realistic values with greater validity but imputed values are limited to the range of observed data. PMM has been proposed as a method which provides a greater degree of robustness against model misspecification. For example, this method may be appropriate when data are non-normal or the model contains non-linear associations (Morris et al., 2014). However,

this technique is not recommended when the sample size is small since this will result in almost identical imputations in each imputed dataset.

Each of these imputation approaches are widely implemented in standard statistical software including R, STATA and SAS (Horton & Kleinman, 2007).

Analysis Phase

The second stage of multiple imputation is the separate analysis of each imputed dataset. This step is straightforward and requires only traditional complete-case analysis approaches. Each imputed dataset will produce distinct parameter estimates due to imputed data differing in each newly created dataset (White et al., 2011).

Pooling Phase

The final step involves the aggregation of estimates from each imputed dataset according to Rubin's combining rules (1987). To obtain the point estimate for the parameter of interest, such as a regression coefficient, the arithmetic average of parameter estimates obtained from each imputed dataset is calculated. Calculating pooled standard errors is slightly more complicated. Standard errors from multiple imputation combine two sources of sampling fluctuation; firstly the within-imputation variance and secondly the between-imputation variance. The within-imputation variance reflects the uncertainty of the results from a single imputed datasets whilst between-imputation variance allows for the additional uncertainty due to the missing observations. It is the between-imputation variance which simple imputation approaches are unable to account for since they treat imputed values as real data and do not account for the additional uncertainty of the imputations.

To calculate the within-imputation variance, the arithmetic average of the sampling variances, (i.e. squared sample errors) from each imputed dataset is taken. Mathematically this is:

$$V_W = \frac{1}{m}\sum_{t=1}^{m} s_t^2$$

Where V_W represents the within-imputation variance, m denotes the number of imputed datasets and s_t^2 is the standard error for dataset t.

The between-imputation variance is equal to the sample variance of the parameter estimates across the *m* datasets and quantifies the uncertainty due to the missing data. The formula is as follows:

$$V_B = \frac{1}{m-1}\sum_{t=1}^{m}(\hat{\theta}_t - \bar{\theta})^2$$

Where V_B denotes the between-imputation variance $\hat{\theta}_t$ is the parameter estimate from dataset *t* and $\bar{\theta}$ is the average point estimate (Enders, 2010). To obtain the multiple imputation standard error, $V_B$and V_W are summed, with the addition of a correction factor to V_W due to the average parameter estimate being subject to sampling error too. The square root of the sum of these values is then taken, as shown below:

$$SE = \sqrt{V_w + V_B + \frac{V_B}{m}}$$

where SE denotes the standard error. From the between-imputation variance, it is also possible to calculate the fraction of missing imputation which is able to quantify the missing data's influence on the sampling variance of a parameter estimate.

Most statistics may be combined using Rubin's rules. However, some statistics may need to be transformed to achieve Normality such as odds ratios and hazard ratios. Statistics that may not be combined using Rubin's rules are those whose values change with the sample size such as measures of strength including P-values (White et al., 2011). Most software packages which implement multiple imputation have built-in processes which automate the pooling process.

2.3.1.2 Model Specification

Incorrectly specified imputation models are the source of the majority of problems associated with multiple imputation. Important considerations when specifying the imputation model include the number of and choice of variables, the number of imputed datasets created and the compatibility of the imputation model and the model of substantive interest.

Variable Selection & Form

To obtain unbiased parameter estimates, variable selection is a vital step of imputation model specification. Despite the risk of over-fitting and the resulting potential for reduced precision, the general consensus is that the inclusion of more variables is better (Rubin, 1996; Kenward & Carpenter, 2007). Briefly, it is important that the imputation model includes the following:

- All variables to be included in the model of substantive interest
- All variables that are predictive of missingness in the unobserved variable
- All variables associated with the unobserved variable

To avoid bias, the imputation model should include all variables that are in the model of substantive interest. To preserve relationships of interest, imputation models should include predictors and outcomes that are involved in these later analyses (Schafer & Graham, 2002). Failure to do so will weaken the association between these variables, as was unfortunately illustrated in the development of the QRISK tool for cardiovascular disease prediction (Hippisley-Cox et al., 2007). In this high-profile mistake, researchers excluded their cardiovascular disease outcome from their imputation model for imputing cholesterol values and subsequently reported a null association between cholesterol value and cardiovascular disease risk. The inclusion of all variables to be included in analysis models may be difficult to achieve in settings where the imputed datasets are to be used for multiple analyses (White et al., 2011).

The imputation model must also include these variables in their correct functional form. For example, if the model of substantive interest contains transformed variables, interaction terms or higher-order terms to account for non-linearity, the imputation model also needs to include these terms.

However, there is currently no consensus regarding the best way of incorporating non-linear and interaction terms. A number of approaches have been suggested, such as passive imputation, just another variable (JAV), PMM and polynomial combination (Von Hippel, 2009; White et al., 2011; Seaman et al., 2012; Vink & van Buuren, 2013; Bartlett et al., 2014). In passive imputation, all non-linear and interaction terms are excluded from the imputation model and are then calculated from the imputed value. JAV involves the inclusion of interaction terms in a MVN imputation model.

Unfortunately, both of these approaches and PMM have been found to produce biased estimates when data are MAR, particularly for logistic regression (Seaman et al., 2012). These approaches have yet to be investigated in a survival analysis setting. Alternative means of handling higher terms such as polynomial combination are in their infancy and have not been applied in a wide range of settings (Vink & van Buuren, 2013).

The appropriate method for incorporating ratio variables, such as BMI and total cholesterol in multiple imputation also remains unclear. Only one study was identified which explored this issue (Morris et al., 2014). This study advises against the use of a passive imputation approach, whereby the denominator and numerator are imputed and then the ratio is calculated using imputed data. Further work is required on this topic.

Similarly, there has been uncertainty regarding the most appropriate means of incorporating outcome variables when the model of substantive interest is a survival analyses. In this setting, the outcome comprises of both the time (t) and the censoring indicator variable (D) and failure to incorporate these in imputation models may dilute the association between the covariate and survival. Previously these variables have been included in imputation models as D, t and log t with little justification (van Buuren et al., 1999). However, White and Royston (2009) found that this approach biased associations between covariates and survival towards the null and instead proposed the inclusion of D and the Nelson-Aalen estimator of the cumulative baseline hazard. This approach had the lowest bias and highest precision in the majority of simulations carried out.

The imputation model should also include those variables which are predictive of missingness in the unobserved variable. This makes the MAR assumption more plausible (Sterne et al., 2009). The MAR mechanism assumption made my multiple imputation only holds if the variables which are predictive of missing data are included in the imputation model (Enders, 2010).

Variables which are correlated with the missing variables to be imputed should be included in the imputation model too. These variables will not necessarily be included in the final analysis but can provide valuable information for the imputation process. Selecting these variables will be based around the biological plausibility of any associations. For example, when attempting to impute weight it would be advisable to

include height (since these variables are likely to be correlated) even if height is not to be included in the final analysis.

This fully inclusive analysis strategy has been widely recommended to avoid bias. Reduced precision through over-fitting has been observed, though in practice these reductions are typically small and unlikely to be a problem in large datasets (Kenward & Carpenter, 2007).

Handling non-Normally distributed variables

Appropriately handling non-Normally distributed continuous variables is a further important step in model specification. Both MICE and the MVN imputation model assume incomplete continuous variables are Normally-distributed. There is currently no agreement regarding the impact of including non-Normal variables in imputation models. A number of studies have suggested that their inclusion may bias results and imputed data will not resemble that of the observed values (Sterne et al., 2009; White et al., 2011; Lee & Carlin, 2012). For example, in their empirical study Lee and Carlin (2012) observed considerable bias when estimating the effect of a skewed continuous covariate even when missingness was moderate. On the other hand, other studies have shown that normality violations do not pose a threat to the accuracy of parameter estimates from multiple imputation (Schafer & Graham, 2002; Demirtas et al., 2008)

To mitigate the potential for bias, transformations may be used prior to imputation before back-transforming afterwards. Proposed transformations include the Box-Cox transformation, a procedure which identifies an appropriate exponent to add to observed values until the value which best removes the skew is identified and the shifted-log transformation which involve taking the log of the observed values and subtracting a positive or negative value dependent on whether the distribution is right or left skewed. Alternatively, PMM may be used (White et al., 2011).

Number of imputations

The number of imputed datasets to include is an important consideration when specifying the imputation model. Earlier work suggested that a small number of imputes (e.g. m=3 to 5) would suffice, though more recently there has been a shift in opinion towards larger number of imputations. White et al. (2011) have argued that considerably more imputations are required to provide an appropriate level of reproducibility. Here,

the authors recommend that the number of imputations should be similar to the percentage of incomplete cases. This recommendation comes from simulation studies which have explored Monte Carlo errors and identified that more imputations are better for improving reproducibility along with providing good statistical efficiency and limiting loss of power.

2.3.1.3 Situations when Multiple imputation may be beneficial

Multiple imputation is widely advocated as a better missing data handling approach than CCA. Whilst this is true for a wide range of settings, there are a number of settings in which more traditional approaches may be more appropriate. This is an important consideration which is rarely emphasised in the missing data literature (White & Carlin, 2010).

Both multiple imputation and CCA produce valid estimates when data are MCAR. In this setting, the use of multiple imputation may be advocated in preference of CCA since it is able to use all available information and therefore may improve precision. However, it is important to note that when missing data levels are high, the impact of model misspecification is likely to increase with increasing fractions of imputed data. There are no current guidelines regarding how much missing data may be imputed. For example, Marshall et al. (2010) identified that multiple imputation can produce biased estimates with more than 50% MAR missingness in binary or continuous variables where the outcome is survival. White et al. (2011) meanwhile suggest caution is required when more than 30% of data are to be imputed.

When missing data occur in the outcome variable only and no additional variables can be included in the imputation model that are not in the model of substantive interest, CCA and multiple imputation will yield almost identical results (Enders, 2010). If there are auxiliary variables which are associated with the unobserved outcome that are not included in the model of substantive interest then multiple imputation may be better.

When missing data occur in covariates, the additional benefit gained from using multiple imputation is dependent on the role of the covariate in the analysis. According to recent simulation studies, if the incomplete covariate is included in the analysis model for confounding adjustment only, multiple imputation may be beneficial since it is able to recover information regarding the relationship between the predictor of interest and the

outcome (White & Carlin, 2010; Lee & Carlin, 2012). However, when missingness occurred in the predictor variable of interest only, the benefits of using multiple imputation were minimal, especially where there are no strongly correlated auxiliary variables. It remains uncertain if this effect would be observed across a wider range of settings in real-life data.

The use of multiple imputation therefore requires a careful consideration of the mechanism of missingness, the patterns of missing data and the context of the study. Clearly, its potential for information recovery needs to be balanced against the numerous sources of bias which may be introduced when the imputation model is misspecified.

2.3.2 Maximum Likelihood Estimation (MLE)

Maximum likelihood estimation (MLE) is a broad set of approaches which attempt to identify the population parameter estimates that have the highest probability of producing the sample data, using all available data (Peugh & Enders, 2004). Consequently, this technique can be applied to both complete and incomplete data with minimal alterations. Two main approaches to MLE with missing data are commonly applied; Full-information maximum likelihood, also known as direct maximum likelihood and the EM algorithm. Both approaches assume data are multivariate normal and are MAR.

To achieve their goal, MLE approaches use a log likelihood function to identify the parameter estimates which minimise the standardised distance between the observed data points and the parameters of interest (Enders, 2010). The method uses an iterative process to 'audition' different values for the parameter of interest until the best fit of the data is identified. When MLE approaches are applied, no data are discarded whilst no data are permanently imputed either.

Since its development, numerous real-data and generated-data simulation studies have demonstrated that MLE approaches will produce unbiased estimated with analyses using MAR data, especially where sample sizes are large (Barnard & Meng, 1999; Horton & Kleinman, 2007; Enders, 2010; Jeličić et al., 2010). Even when data are

MCAR, maximum likelihood approaches are more powerful than traditional techniques since they make use of all the observed data (Horton & Kleinman, 2007).

Current options for using MLE approaches for missing data analysis are limited. Though software are available, often the model requires manual specification especially when auxiliary variables are to be included (Enders, 2010). The unavailability of accessible software may have contributed to the limited uptake of MLE methods.

2.3.3 Inverse Probability Weighting (IPW)

An alternative means of dealing with MAR data is IPW. This approach involves weighting complete cases by the inverse of their probability of being a complete case (Horton & Kleinman, 2007). This weighting system gives greater weight in the CCA to cases who had the lowest probability of being a complete case, and less weight to those who had higher probabilities. This therefore aims to rebalance the complete cases so that they become representative of the population of interest. Theoretically this approach is considerably more straightforward than the MAR methods previously discussed. In practical terms, IPW is easier to apply using standard statistical software and unlike multiple imputation, involves specifying a comparatively simple model.

Unfortunately, a serious drawback of this approach is the requirement for data to exhibit monotone missingness when occurring in multiple variables. In large clinical databases, data are typically non-monotone and therefore IPW's utility is limited in these datasets. There are extensions to allow IPW for non-monotone data though these are complicated and not currently available in statistical packages. In addition, IPW is less efficient than multiple imputation and MLE approaches since it only uses complete cases and obtains no other information from incomplete records (Seaman & White, 2013). Due to these important limitations, the use of this approach was not considered appropriate for this large dataset and not pursued further.

2.3.4 Comparison of MAR Methods

Three general methods for handling data that are MAR have been presented, each with a number of important strengths and limitations.

Multiple imputation has a number of strengths over MLE and weighting methods. Firstly, multiple imputation is considerably more flexible and is not model specific. This means the basic steps may be applied very generally and regardless of the analysis of substantive interest, the specification of the imputation model is the same. Each complete dataset may then be used to conduct any analytic model. In contrast, the specification of a MLE model is problem specific since the handling of missing data is in-built into the final analysis model. Specialised software is required to conduct these model-specific analyses. For example whilst MLE methods for linear models may be available in certain software, this same software may not be able to conduct MLE for logistic regression analyses (Allison, 2012).

Secondly, since multiple imputation is based upon two distinct models, it is more straightforward to incorporate auxiliary variables into the missing data model (Johnson & Young, 2011). Unlike in MLE approaches, these auxiliary variables need not be included in the final analysis model. In MLE the inclusion of these auxiliary variables may alter the parameter estimates obtained from the model of interest (Peugh & Enders, 2004). Furthermore, some findings have suggested multiple imputation may give more accurate standard errors than MLE (Abraham & Russell, 2004).

However, whilst multiple imputations will produce slightly different results each time it is applied, thus raising concerns that different investigators could plausibly reach different conclusions, MLE approaches will always produce the same result given the set of data. The complexity of multiple imputation also means this approach is considerably more labour intensive than the other MAR methods. To apply multiple imputation a number of issues require considerable forethought. For example, thorough consideration of the imputation model to be used, the number of datasets to produce and the handling of interactions is required. In addition, the imputation and final analysis model need to be compatible to prevent bias. Failure to include variables to be included in the final model in the imputation model is likely to dilute associations of interest. In comparison, the specification of MLE models is much more straightforward (providing the software is available) since all analyses are confined to a single analysis.

Providing the same set of cases and variables are used and the number of imputed datasets is sufficiently large, these approaches will yield largely similar parameter

estimates. The practical considerations described above are therefore likely to influence their relative utility by researchers

2.4 METHODS FOR MNAR DATA

Missing data may also be MNAR meaning that given the observed data, the missingness mechanism depends on the unobserved data. Selection models and pattern-mixture models are two proposed methods for handling MNAR data and whilst a great deal of literature is devoted to these approaches, much of it remains inaccessible to non-statisticians. These approaches attempt to jointly model both the distribution of the data and the probability of missing data but in very different ways. Selection models combine the substantive analysis with an additional regression equation that predicts response probabilities whilst pattern mixture models produce sub-samples of cases with similar missing data patterns and estimates the substantive analysis model for each pattern (Abraham & Russell, 2004). Unfortunately, in practice these models are difficult to specify and are not currently implemented in standard statistical software.

Basic sensitivity analyses have therefore been proposed which use the general principles of the above approaches but are considerably more accessible in standard statistical software. Rather than producing joint MMAR models, these sensitivity analyses investigate the robustness of estimates obtained from multiple imputation under the MAR assumption.

2.4.1 SELECTION-BASED SENSITIVITY ANALYSES

One selection-based approach is the reweighting method, first proposed by Carpenter et al. (2007). This technique reweights parameter estimates obtained from the imputed datasets to reflect the distribution of imputations under a MNAR mechanism. In ordinary multiple imputation, the point estimates from each imputed dataset are simply averaged, whilst in this approach, weighted averages are used. Here, estimates acquired from imputations that are more likely under MNAR are up-weighted relative to the others. The value of the weight, often referred to as δ may be elicited from experts' prior beliefs (Scharfstein et al., 2003). Alternatively, as proposed by Carpenter et al.

(2007), δ may reflect how much the probability of observing the incomplete variable increases for a one unit increase in that variable, given the other variables. Accordingly, this method is a local sensitivity analysis approach which assumes the true MNAR inference lies within the range of estimates from MAR. Furthermore, if the number of imputations is small, one imputation may receive a disproportionally large weight and therefore dominate the pooled estimate. However, Carpenter and colleagues suggest that this will not necessarily mean the estimate is poor. At the time of writing, this latter approach had only been reported in a single observational study (Heraud-Bousquet et al., 2012). In this approach, Heraud-Bousquet and colleagues applied this approach to an investigation of the risk factors that were predictive of severe liver disease, in which data for a number of risk factors were missing. The authors were able to explore the application of this approach and encouraged its use where there are non-trivial proportions of missing data.

2.4.2 Pattern Mixture-Based Approach

An alternative approach is the pattern-mixture sensitivity analysis which involves changing imputations obtained under the MAR assumption to reflect a MNAR assumption. This approach was first proposed by Rubin who described it as an ad-hoc imputation approach (Rubin, 1987). In this simple model, Rubin proposes that imputations are generated under an MAR mechanism and then a constant term is added to 50% of these imputed values to compensate for the possibility that these may be too high or low. Alternatively, a prior distribution of the differences between the observed and unobserved values for the incomplete variable could be elicited from experts (White et al., 2007). From this prior distribution, values are drawn and added to the imputed values. In practice eliciting this prior distribution is not straightforward and may not be possible in routine data sources where there may be no experts on the likely distribution of the unobserved variable.

2.4.3 Summary

In summary, both of these approaches rely on potentially poorly estimated MNAR models and make assumptions that are not only tenuous but also cannot be verified. Where these assumptions are violated it is likely that they will produce severely biased

estimates (Enders, 2010). Schafer (2002) suggested that it would be preferable to specify a richer imputation model under the MAR assumption with more auxiliary variables that were related to missingness. By doing so, the MAR assumption is made more plausible.

2.5 CONCLUSIONS

There are currently an abundance of methods for handling unobserved data, each with a number of strengths and limitations which will make them more or less appropriate in a given setting. Ultimately, the use of a given method is dependent on a number of important considerations:

1. Extent of missing data

   It has been repeatedly shown that the level of bias is likely to increase with increasing proportions of unobserved data. However it is not sufficient to base the decision on the appropriateness of a missing data method on the extent of missing data alone.

2. Pattern of missing data

   Some methods are more readily suited to specific patterns of missing data, though this is becoming less important with the advent of flexible missing data techniques. The appropriateness of a method will also depend on whether the missing data occurs in a covariate or in the outcome variable. For example, multiple imputation is not worthwhile in settings where only the outcome variable has unobserved data and there are no auxiliary variables predictive of the probability of missingness in this variable.

3. Method assumptions

   The most important assumption made my missing data methods is the missingness mechanism. Violation of this assumption will result in biased estimates. A thorough exploration of the mechanisms of missingness is therefore recommended. Furthermore, some methods make distributional assumptions that may not be plausible in the given dataset.

4. Accessibility

A further consideration is the accessibility of the method. Many of the approaches detailed above remain inaccessible to non-statisticians and may not be implemented in standard statistical software. However, a number of recent developments have reduced this problem. Unfortunately, maximum likelihood approaches and MNAR methods still remain unavailable without considerable user input.

3. Literature Review of the Relationship between BMI & Mortality

3.1.1 REVIEW AIMS

To illustrate the techniques for handling missing data, an analysis of the relationship of both cross-sectional and longitudinal patterns of BMI on all-cause mortality was undertaken. In an effort to determine the current approaches to measuring this relationship and the current understanding of the association between patterns of BMI and all-cause mortality, a small-scale literature review was carried out. The chosen method for handling missing data in each of the relevant studies was also documented.

3.1.2 SEARCH STRATEGY

To identify relevant articles, Medline, EMBASE, and the Web of Knowledge were searched through August 2011. Searches of papers investigating the association between BMI and all-cause mortality were not limited to diabetes populations. In addition to those identified using these searches, reference lists from relevant articles including systematic reviews were examined. This literature review was continually updated with new studies.

3.2 BMI AND MORTALITY IN NON-DM PATIENTS

The association between cross-sectional BMI and mortality has been extensively explored in the general population. A number of large-scale studies have demonstrated a U or J-shaped association between BMI and all-cause mortality (Pischon et al., 2008; Prospective Studies Collaboration, 2009; Berrington de Gonzalez et al., 2010; Hotchkiss & Leyland, 2011; Zheng et al., 2011). Figure 3.1 illustrates a convincing U-shaped relationship in a population of 900,000 adults from 57 prospective studies. The large sample size of this and many other studies investigating this association have enabled researchers to establish a more precise dose-response relationship.

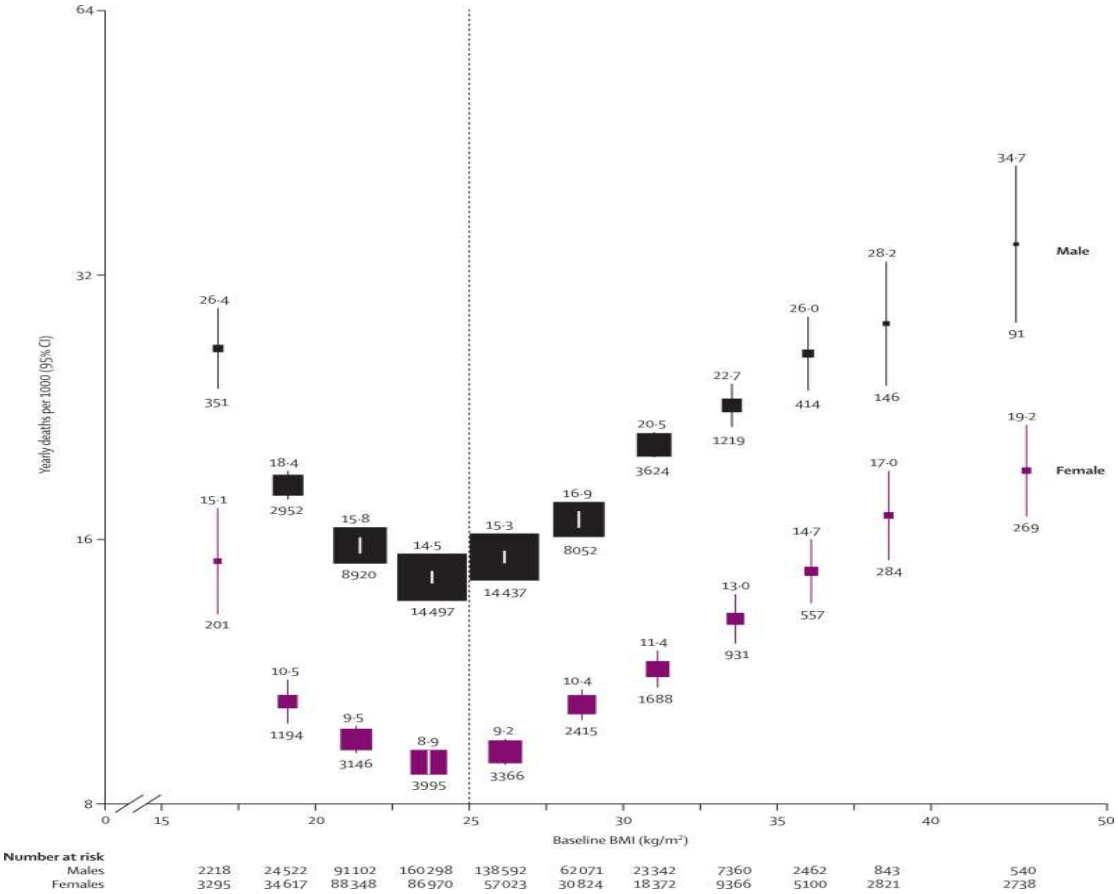

Figure 3.1 All-cause mortality versus BMI for each sex in the range 15–50kg/m² (excluding the first 5 years of follow-up). (From Prospective Studies Collaboration, (2009))

Typically the lowest risk of mortality is observed in the 22.5 to 30kg/m² BMI ranges with progressively increased risk in higher and lower BMI categories. Whilst the risk of mortality is consistent in the upper BMI categories, a great deal of uncertainty remains in the lower BMIs. The uncertainty of the relationship at these BMIs likely reflects the complicated nature of this association. More specifically, the overall association may be confused by the risk of residual confounding by smoking and reverse causation by chronic pre-existing disease (Nilsson, 2008).

Smoking is strongly associated with both weight loss and premature mortality and therefore excess mortality in below-normal BMI groups may be a consequence of smoking rather than a mechanism of low BMI. Adequately controlling for smoking and smoking intensity is difficult, especially where the quality of smoking data is poor. Where smoking data are available, it may be possible to exclude current or former smokers to produce unbiased estimates (Stampfer, 2005), though this notion has been

dismissed elsewhere (The BMI in Diverse Populations Collaborative Group, 1999). One argument against the exclusion of smokers is that results from non-smokers are not generalizable.

In instances where analyses have been limited to non-smokers, the relationship between BMI and mortality has become increasingly J-shaped. For example, in a pooled analysis of 19 prospective studies from the National Cancer Institute Cohort Consortium, the hazard ratio (HR) for all-cause mortality in the 20 to 22.5kg/m² BMI category decreased from 1.60 to 1.01 amongst healthy male participants when smokers were excluded (Berrington de Gonzalez et al., 2010). The decrease was less apparent amongst females. Similarly, in a cohort of U.S. men and women aged between 50 and 71 years, the association between BMIs below 23.5kg/m² and mortality was strongest amongst current smokers than never or former smokers (Adams et al., 2006). This association was weakened in former smokers and further still amongst never smokers. Whilst this study had a number of strengths including a large sample size which allowed for detailed exploration of the relationship in various subgroups, the use of self-reported weight was a clear limitation of this research. Self-reported weights have been demonstrated to be under-estimated amongst overweight and obese patients, though it seems unlikely that the level of underestimation would differ by the patients smoking status. Evidence of reduced mortality in the lowest BMI categories amongst never smokers compared to never smokers has also been reported in a population of elderly patients (ages ≥65 years) participating in the Leisure World Cohort Study in California and in the large scale European Prospective Investigation into Cancer and Nutrition (EPIC) study (Corrada et al., 2006; Pischon et al., 2008). In each of these studies, obesity was more strongly associated with mortality risk in never smokers than ever smokers.

Nonetheless, Hotchkiss and Leyland (2011) observed the opposite when performing a sensitivity analysis using sequential exclusions in their study based in Scotland. This study used cross sectional surveys linked to NHS administrative databases and had a sample size of 20,117 people. Hazard ratios (HR) for mortality in the less than 18.5kg/m² BMI category initially decreased from 2.63 [95% Confidence interval (CI): 1.95-3.55] to 2.54 [95% CI: 1.82-3.55] with the exclusion of patients with cancer or cardiovascular disease, before increasing to 2.71 [95% CI: 0.84-8.70] following the exclusion of ever smokers. However, as the confidence intervals indicate, these

exclusions resulted in the exclusion of over 60% of the sample and therefore the precision of estimates was considerably reduced.

Research using the California Teachers Study also did not identify a difference in the relationship between BMI and mortality between smokers and non-smokers (Bessonova et al., 2011). The method of categorisation may have contributed to this finding, with former smokers being grouped with current smokers. Of the patients in this group, 85% were former smokers and therefore the effect of smoking was likely to be diluted in this analysis. When the relative risks of mortality in the lowest BMI categories of BMI were observed for all participants and never smokers only, the risk of mortality was lowest for never smokers. This study also relied upon self-reported weights and the sample was made up of California women in the teaching profession only. These may be possible sources of bias.

The presence of chronic pre-existing disease in patients may also act to complicate the relationship between lower BMIs and mortality through reverse causation. Patients with chronic diseases may lose weight through a number of biological processes or may intentionally lose weight in response to the disease diagnosis. Lower BMI values could therefore be indicative of disease rather than a direct cause of mortality. To overcome this problem, most researchers opt to exclude deaths that occur in the first few years after follow-up. Though this approach will limit the risk of reverse causation, some conditions such as depression and lung disease have a longer duration and may proceed undiagnosed. This problem was illustrated by Berrington de Gonzalez et al. (2010) whereby they were able to show that the association between underweight (BMI below 18.5kg/m^2) and risk of mortality was weaker amongst patients with a follow-up of 15 years compared to patients with a follow-up of 5 years (HR: 1.21 vs. 1.73). Nonetheless, weight gain amongst underweight participants during this lengthy follow-up could bias these results. Likewise, the Prospective Studies Collaboration (2009) demonstrated that the inverse association between lower BMI levels (<25kg/m^2) and mortality weakened upon exclusion of patients who died within 10 years of follow-up.

An alternative approach to preventing reverse causation through chronic disease has been the exclusion of patients with pre-existing disease. Adams et al. (2010) reported a strengthening of the association between obesity and all-cause mortality upon restricting

analyses to participants without pre-existing disease. This is despite this method being unable to account for undiagnosed illnesses.

Stratifying analyses by age has also been used to address reverse causation since younger patients are less likely to be afflicted with underlying chronic diseases (Pischon et al., 2008). Pischon and colleagues (2008) stratified age by <55 years, 55 to <65 years and ≥65 years and surprisingly reported a higher risk of mortality in the BMI categories below 23.5kg/m² amongst men aged below 55 years (HR for risk of mortality in 18.5 to <21kg/m² vs. 23.5 to <25kg/m²: 1.69). In contrast, risk of mortality in men with BMIs of 18.5 to < 21kg/m² aged 65 and above was 1.32. Interestingly, this finding was not demonstrated amongst women.

Using a slightly different approach to handle reverse causation, Adams and colleagues investigated the association between BMI and mortality by using patients recalled weight at the age of 50 (2010). In this analysis, the relative risks of death in the BMI categories below 23.5kg/m² based on recalled BMI were lower than observed in the analyses using the current BMI values.

In contrast, excess body fat is a well-established risk factor for premature mortality. For example, the Prospective Studies Collaboration demonstrated that mortality was approximately 30% higher for every increase in BMI of 5kg/m^2 above the optimal BMI range (22.5-25 kg/m^2), after adjustment for age at BMI determination, smoking and study. Most of the excess deaths associated with raised BMI were attributable to vascular diseases. This is a likely impact of the association between elevated adiposity and increased blood pressure, lipoprotein particles and type 2 DM incidence.

Nonetheless, whilst excess death is generally attributed to people classed as severely obese (BMI ≥35kg/m²), there remains uncertainty whether moderate elevations in BMI confer a survival advantage over normal weight, particularly amongst patients with chronic diseases such as type 2 DM and heart failure (Fonarow et al., 2007). This has been termed the obesity paradox. In the general population, a recent systematic review reported a potential survival advantage for being overweight but not obese relative to normal weight(HR overweight vs. normal weight: 0.92 [95% CI: 0.88/0.96]) (Flegal et al., 2013). The meta-analysis conducted in this review used data from over 80 papers, including the aforementioned studies and stratified analyses based on whether weight

was self-report or measured weight. The findings were the same even after stratification and therefore misclassification bias is unlikely to have contributed to this observation. It did not investigate the association between BMIs lower than 18.5kg/m² and mortality. Flegal and authors (2013) postulated that possible explanations for these findings include improved management of obesity-related chronic disease such as type 2 DM, earlier presentation, greater likelihood of receiving intensive treatment and potential benefits of raised metabolic reserves. Evidence which supports the notion that improved treatment may explain this trend comes from results from older cohort studies. In these studies, the association between obesity and mortality were much stronger than observed in more contemporary studies. Using data from National Health and Nutrition Examination Survey (NHANES) I, II and III, Flegal et al., (2005) demonstrated that relative risks of mortality were less strongly associated with obesity in NHANES II and III relative to NHANES I. These reduced associations possibly reflect better treatment for cardiovascular disease and its risk factors, particularly amongst obese patients (Preis et al., 2009).

Unfortunately, most of the aforementioned studies do not provide much detail regarding the approach to handling missing data utilised. Typically, CCA was applied without providing a justification for its selection (Adams et al., 2006; Corrada et al., 2006; Pischon et al., 2008; Berrington de Gonzalez et al., 2010; Bessonova et al., 2011; Zheng et al., 2011; Chen et al., 2012). For example, the EPIC study excluded over 100,000 (23% of sample) participants due to the presence of missing data (Pischon et al., 2008). None of the studies presented provided a comparison of included and excluded patients making it difficult to ascertain the risk of bias in each of these studies. Nevertheless, Hotchkiss and colleagues (2011) did investigate the differences in estimates generated from multiple imputation and CCA and reported similar estimates from both approaches. The authors demonstrated an increasing risk of missingness in BMI by survival and therefore used multiple imputation assuming MAR to handle the unobserved data.

3.3 BMI AND MORTALITY IN PATIENTS WITH TYPE 2 DM

Of the studies addressing this topic amongst patients with DM, only one study explored the potential impact of missing data on estimates of the association between BMI and mortality (Logue et al., 2013). In all other articles identified, patients with missing data were excluded regardless of the proportion of missing data. For example, Zoppini et al. (2003) excluded over 50% of their sample and did not provide any comparison of patients with and without missing data. It was therefore not possible to assess whether the use of CCA was likely to affect this study's findings in this study and many others addressing this topic.

The association between BMI at diagnosis and all-cause mortality remains uncertain amongst patients with type 2 DM. While findings have consistently observed a U-shaped relationship in patients with DM, there remains uncertainty whether an obesity paradox is present (Zoppini et al., 2003; Khalangot et al., 2009; Carnethon et al., 2012; Logue & Sattar, 2013; Thomas et al., 2014; Tobias et al., 2014).

Nevertheless, the vast majority report reduced mortality in overweight and obese individuals relative to normal weight patients. For example, using data from 37,000 patients with type 2 DM on the General Practice Research Database, Thomas et al. (2014) reported a HR of 1.47 (95% CI: 1.29, 1.69) for patients with normal weight (defined as $<25kg/m^2$) relative to obese patients (defined as $\geq 30kg/m^2$). This is despite the grouping together of moderately and severely obese patients. Similarly, findings from a pooled analysis of five US-based longitudinal studies indicated a considerably higher risk of mortality in the normal weight group at diagnosis relative to the obese group (HR: 2.08 [95% CI: 1.52, 2.85]) (Carnethon et al., 2012). The risk of mortality was similar between obese and overweight patients. Using data from the Scottish DM register, Logue and colleagues (2013) demonstrated the presence of an obesity paradox, primarily amongst men. Relative to overweight BMI status (25 to $<30kg/m^2$) men with obesity had better survival (HR: 0.97 [0.91-1.04]) whilst obese women had worse survival (HR: 1.04 [0.96-1.12]), though differences between these BMI categories were not statistically significant. Meanwhile, patients with normal weight had considerably worse all-cause mortality, relative to overweight patients and this was observed amongst men (HR 1.22 [1.13-1.32]) and women (HR: 1.32 [1.22-1.44]). In another study

conducted on over 80,000 Ukrainian, conflicting findings by sex were again observed (Khalangot et al., 2009). Amongst men, risk of mortality was higher in the 23 to 24.9kg/m² relative to the 25 to 29.9kg/m² group (HR: 1.11 [1.01-1.22]) but similar amongst women (0.99 [0.90 to 1.08]), after adjustment for age, smoking, alcohol consumption, systolic blood pressure, total cholesterol, history of cardiovascular disease, DM treatments and duration of DM. Evidence of an obesity paradox was also observed in a population of type 2 DM patients in Finland (Hu et al., 2005). However, the use of sex-specific tertiles hindered the interpretation of HRs in this study. Relative to BMI tertile 1 (<27.1kg/m² for men and <28.0 for women), risk of mortality was lower in patients with a BMI in tertile 2 (27.1 to <30.4kg/m² for men and 28.0 to 32.6kg/m² for women) (HR: 0.94 [0.80-1.10]). However, these estimates were not significant due to small sample sizes (n=3,708).

The authors of the above studies have offered multiple explanations for this obesity paradox. Firstly, failure to adequately addressing missing values may have led to biased estimates. Secondly, as described previously, failure to adequately control against reverse causation and confounding by pre-existing disease and smoking status, respectively. Thirdly, the paradox could reflect the poor performance of BMI as a general marker for adiposity. Finally, patients who are diagnosed with DM at a lower BMI may have a more damaging disease phenotype which ultimately leads to premature mortality. Should the paradox be real and not simply an artefact of poor study design, then there may be large implications for DM treatment. Currently, recommendations suggest newly-diagnosed type 2 DM patients should be encouraged to lose weight (Scottish Intercollegiate Guidelines Network, 2010). This advice may not be appropriate for thin patients with increased risk of mortality.

Conversely, a number of studies have reported no evidence to support the existence of an obesity paradox. Most recently, Tobias et al (2014) found no reduced mortality among overweight or obese patients at time of diagnosis compared to normal-weight patients using data from the Nurses' Health Study and the Health Professionals Follow-up Study. Likewise, an observational study using data from 13,000 patients with type 2 DM on the Swedish National Diabetes Register reported that obesity in increased the risk of mortality by 71% compared to normal weight, after adjustment for age, sex, type

of hyperglycaemic treatment, DM duration, smoking (Eeg-Olofsson et al., 2009). Explanations for these discrepant findings may include limited power and other practical differences such as inconsistent approaches for controlling for reverse causation and confounding.

3.3.1 CONFOUNDERS & EFFECT MODIFIERS

Along with smoking status and the presence of comorbidities, a number of other characteristics are typically controlled for in studies addressing the topic of BMI and mortality. Unsurprisingly, all studies typically adjust for age at diagnosis. Increasing age is associated with increasing mortality and patient BMI tends to increase with age, before gradually decreasing amongst elderly patients. Some studies have also identified potential effect modification of BMI by age at diagnosis. For example, using data from participants in the Verona Diabetes Study, Zoppini et al., (2003) demonstrated that amongst older patients (>65 years) obesity was protective against death, whilst obesity was related to poorer outcomes amongst younger patients. This observation was consistent with findings from other large-scale studies (Thomas et al., 2014; Tobias et al., 2014). This interaction effect has been suggested to reflect the inability of BMI to distinguish between lean muscle mass and adiposity. In elderly patients, decreased BMIs may reflect muscle wasting through sarcopenia rather than reductions in body fat. As a result, normal weight elderly patients may appear to have worse outcomes than obese patients. In the EPIC study conducted in the general population, this effect modification by age was stronger amongst men than women (Pischon et al., 2008). Nonetheless, this relationship was not observed in the study conducted by Carnethon et al. (2012).

Sex is also an important confounder of this relationship. Women tend to have higher BMIs at diagnosis than men whilst having lower mortality levels (Logue et al., 2011). Though the majority of studies have combined analyses with men and women there are some evidence to suggest effect modification by sex. For example, Logue al. (2013) observed a BMI of 30 to 35kg/m² to be protective against mortality amongst men but not women. Khalangot and colleagues reported similar findings, though no differences in relative risks of death by BMI were observed between men and women in the study conducted by Thomas and colleagues (2014). In contrast, Pischon et al. (2008) reported

increased more mortality in high BMIs and suggested this may be since BMI is more closely related to abdominal fat amongst men than in women. Deprivation status may also be an important confounder of this association. BMI tends to be higher in more deprived categories and risk of mortality (Connolly & Kesson, 1996). Treatment type is another potential confounder as some anti-hyperglycaemic agents have been found to affect weight changes and the use of these agents may also affect overall mortality.

A number of biochemical characteristics may confound or mediate the association between BMI and mortality. For example, raised BMIs are widely associated with raised total cholesterol and blood pressure values, whilst also being risk factors for a number of adverse health outcomes, particularly cardiovascular disease and mortality.

Glycated haemoglobin may affect the observed relationship. Logue et al. (2013) observed higher glycated haemoglobin levels amongst leaner individuals whilst poorly controlled glycated haemoglobin is associated with higher levels of DM complications. This may offer an additional explanation for the increased risk of mortality in leaner patients relative to overweight patients. However, upon additionally adjusting for glycated haemoglobin, total cholesterol, LDL-cholesterol and systolic blood pressure, the authors did not report any differences in estimates. Though these variables had a lot of incomplete data and therefore a large number of participants were excluded.

Meanwhile, Hu et al (2005) reported an attenuation in the risk of mortality in obese patients relative to normal weight patients upon additionally adjusting for systolic blood pressure and total cholesterol (obese vs normal HR: 1.24 [1.09, 1.41] vs. 1.14 [1.00-1.30] for fully adjusted model). A similar attenuation of the risk of total mortality in obese patients was observed in the Swedish National Diabetes Register study (Eeg-Olofsson et al., 2009).

However, the practice of adjusting for factors such as blood pressure and total cholesterol has been actively discouraged (Manson et al., 1987; World Health Organisation, 2000). These factors are direct effects of raised adiposity and are therefore on the causal pathway between BMI and mortality. Controlling for these factors will attenuate and potentially eliminate the relationship between body mass and mortality.

This may lead to the erroneous conclusion that BMI is not associated with increased mortality.

3.4 TRENDS IN BMI IN PATIENTS WITH TYPE 2 DM

As it was likely that measurement data recorded before and after diagnosis date would be used to fill-in missing BMI at diagnosis data, it was deemed worthwhile to investigate current knowledge regarding trends in BMI amongst patients with type 2 DM. Unfortunately, data describing patterns of BMI in patients with DM are less widely available and tend to originate from studies assessing the impact of various glycaemic control therapies. The UK Prospective Diabetes Study (1998) is a good example of this. In this study, changes in BMI in recently diagnosed type 2 DM patients were measured and related to different glycaemic control interventions. Patients on the most intensively treated arm gained considerably more weight than those who were not. In particular, patients assigned insulin-therapy were especially at risk of weight gain. This pattern of weight gain in patients prescribed insulin-therapy has been observed in numerous other studies (Adams et al., 2006; Nichols & Gomez-Caminero, 2007; Westphal & Palumbo, 2007; Janghorbani & Amini, 2009). One study even estimated that during the first year of insulin-treatment, a 1 per cent deduction in glycated haemoglobin came at the expense of a weight gain of approximately 2kg (Yki-Jarvinen, 2001). Meanwhile, treatment with metformin has been demonstrated to be the most weight-neutral glycaemic control therapy (UK Prospective Diabetes Study Group, 1998).

One observational study described trends in BMI prior to and after diagnosis of type 2 DM for a cohort of Pima Indians (Looker et al., 2001). The results from this study conflicted with results from intervention studies. Here, the authors observed a tendency for patients to lose weight post-diagnosis, with the mean change varying between -0.6 and $+0.2 kg/m^2$ per year. However, following stratification by glycaemia control medication type, those patients who were not receiving treatment lost weight most rapidly, whilst in keeping with other studies, patients prescribed insulin lost weight least rapidly or even gained weight. It is unlikely that these results are readily generalisable to other populations though. Overall weight losses (mean change -1.9%; 95% CI -2.1 to -

1.7) were also reported in an Iranian population with type 2 DM with a mean follow-up period of 9.1 years (Janghorbani & Amini, 2009).

Similarly, De Fine Olivarius et al. (2006) reported an average weight change of -3.3kg in the first 5 years of follow-up amongst 711 patients participating in the Diabetes Care in General Practice study. Most of this weight loss occurred within the first five months following diagnosis. Weight loss was largely dependent on treatment type. Patients receiving insulin gained on average 0.5kg whilst patients receiving metformin lost on average 4.3kg and patients receiving dietary advice only lost 5.4kg on average. A larger study conducted in Sweden (n=8,486) noted no change in BMIs (within 1 unit change) in 53% of newly-diagnosed patients with DM (Bodegard et al., 2013). Data from the Health and Retirement Study in the US indicated that 75% of newly-diagnosed DM patients exhibited stable weight trajectories and 18% lost weight before gaining weight during 10 years of follow-up, after adjustment for treatment type (Chiu et al., 2013). In contrast, a US-based study which followed patients for an average of 9.4 years identified small weight increases in patients with DM independent of age at baseline, though these increases were small (mean change 0.23 ± 0.2 kg/year) (Chaudhry et al., 2006). However, these analyses were restricted to only 205 newly diagnosed DM patients.

Data describing trends in BMI prior to diagnosis with type 2 DM are also limited and differences in methods for measuring weight changes make comparisons difficult. Generally studies have reported a steady rise in weight prior to diagnosis with DM (Looker et al., 2001; De Fine Olivarius et al., 2008; Vistisen et al., 2014). Though De Fine Olivarius and colleagues (2008) reported an average weight gain of 1kg per year prior to diagnosis, significant weight loss was reported in the year immediately preceding diagnosis, a possible consequence of hyperglycaemia. In their study of weight changes amongst Pima Indians with DM, Looker et al. (2001) reported an average increase of 0.43 and 0.71kg/m² per year before diagnosis. Meanwhile, using data from the Whitehall II study and latent class trajectory analyses, it was reported that 94% of patients exhibited an average weight gain of 2.3 BMI units between baseline and diagnosis (18 years on average) (Vistisen et al., 2014).

4. DATA CLEANING & MISSINGNESS IN THE SCI-DC REGISTER

4.1 INTRODUCTION

This chapter provides a detailed explanation of the datasets used throughout the entirety of this project. The extent of the missing data problem is documented and the demographic and clinical characteristics of the cohort are described.

4.2 DESCRIPTION OF DATASET

4.2.1 DATASETS

Data were obtained in the form of 13 separate datasets from a pseudonymised, 2008 research extract of the SCI-DC database. As previously described, this is a dynamic national register of people with diagnosed diabetes. The SCI-DC database captures demographic and key DM-related clinical data on a daily basis from the majority of hospital DM clinics (all adult clinics but a smaller proportion of paediatric clinics) and all but 5 of 1003 GPs in Scotland.

The datasets held information on 274,364 patients with any form of DM. Data extraction of these datasets took place in July 2008 and the generation of linked datasets took almost two years. These datasets held information on fixed demographic and repeat biochemical results, including dates of measurement. Datasets were combined using a unique patient identifier, the community health index. Patients' names, addresses, or dates of birth were not included in any dataset.

A hospital admissions dataset, the Scottish Morbidity Records (SMR-01), and mortality data were linked to the SCI-DC database by Information Services Division (ISD), a division of NHS Scotland using the Scottish unique health record identifier, the Community Health Index and probabilistic linkage. SMR-01 contains information of date of admission, date of discharge and the relevant International Classification of Disease (ICD) diagnosis codes. Date of death data were obtained in 2012 from the National Records of Scotland (NRS). This register captures all deaths occurring in Scotland following the completion of a death certificate.

Permission for the generation of the linked dataset was obtained from the Scottish multi-centre research ethics committee, the Caldicott guardians of each Health Board and ISD's Privacy Advisory Committee. Details of each dataset, including their respective number of observations and number of variables can be found in Table 4.1

4.2.2 DESCRIPTION OF VARIABLES

A description of available variables in the SCI-DC register is presented in Table 4.2. The first half of the table describes variables that are measured once in each person, whilst the second section describes variables that are measured multiple times. Notes are provided which outline each variable's method of collection and the recommended measurement practice for each clinical characteristic as described by the 2010/2011 QOF and the Scottish Intercollegiate Guidelines Network (SIGN) (2010).

Although the majority of these variables are unlikely to be of substantive interest in the final analysis model (which investigates the association between BMI and all-cause mortality), auxiliary variables may provide information regarding the potential reasons for missingness or may help to predict the missing values themselves. In missing data analyses where data are MAR, it has been recognised that the inclusion of auxiliary variables which are correlated with the variable with missing data will improve imputations, whilst including variables which are correlated with the probability of the variable being missing will also reduce bias (White et al., 2011)

Table 4.1: Details of 13 datasets extracted from SCI-DC from 2008 onwards.

Dataset name	Total no. observations	Available Variables
BMI	11,221,422	Body mass index kg/m²
		Height (m)
		Weight (kg)
		Smoking status
		Date of measurement
BP	8,980,146	Systolic blood pressure (mmHg)
		Diastolic blood pressure (mmHg)
		Date of measurement
Creat	3,472,238	Creatinine (umol/L)
		Date of measurement
Ethnic	273,690	Year of birth
		Sex
		Ethnic Group
		Date of Death
		Vital status (True/False)
		Health Board Area (HBA)
HbA1c	3,300,130	Percentage Glycated Haemoglobin
		Date of measurement
Hdl	1,576,348	High-density lipoprotein (HDL) cholesterol level (mmol/L)
		Date of measurement
Scottish Index of Multiple Deprivation (SIMD)	272,268	Rank on the SIMD
		SIMD score
		SIMD quintile
Trig	796,077	Triglyceride level (mmol/L)
		Date of measurement
TotalChol	2,856,858	Total cholesterol level (mmol/L)
		Date of measurement
Diagdate	274,363	Date of diagnosis
DMtypeMay10	273,756	DM type (generated from algorithm)
Scottish Morbidity Records (SMR01)	2,541,690	Admission date
		Discharge date
		Primary condition
		Secondary conditions (up to five conditions)
GROA	90,957	Date of death
		Primary cause/Secondary causes (up to three)

Table 4.2: Description of variables extracted from SCI-DC to be used in analyses.

	Variables	Coding	Notes
Fixed Measurements	Year of birth	Year	Used in conjunction with year of diagnosis to calculate patient age at diagnosis
	Date of death	Date	
	Sex	0 = female, 1 = male	
	Health Board Area (HBA)	A = Ayrshire & Arran, B = Borders, F = Fife, G = Greater Glasgow, H = Highlands, L = Lanarkshire, N = Grampian, R = Orkney, S = Lothian, T = Tayside, V = Forth Valley, W = Western Isles, Y = Dumfries and Galloway, Z = Shetland	Assigned to each patient by determining within which HBA the patient's registered GP is located. There are 14 regional NHS boards across Scotland responsible for the delivery of frontline healthcare services within that locality. Prior to 2006, Argyll and Bute was an additional HBA in operation. However, in 2006, the responsibilities of the Argyll and Bute Health Board were shared between the Highland and Greater Glasgow HBAs. This thesis will discuss the post-2006 HBAs only.
	Ethnicity	0 = Scottish, 1 = Caribbean, 2 = African, 3 = Indian, 4 = Pakistani, 5 = Bangladeshi, 6 = Chinese, 29 = Unknown, 30 = Any other ethnic origin, 31 = Irish, 32 = Other British, 33 = Other white background, 34 = Other black background, 35 = Other Asian background, 36 = Other mixed background, 40 = Not known, 50 = Not disclosed.	Patient ethnicity is traditionally, a poorly recorded demographic characteristic. Fortunately, over 95% of people in Scotland are White British and therefore failure to include this variable in analyses due to high rates of unobserved data should have a negligible effect (Office of the Chief Statistician, 2004) .
	Diabetes type	0 = Not diabetic, 1 = type 1, 2 = type 2, 3 = Maturity onset DM of the young, 4 = Other DM type	DM type was derived using an algorithm generated by Dr Shona Livingstone and Professor Helen Colhoun in Dundee which utilised data for age at diagnosis, the use of and the timing of treatment with oral hypoglycaemic agents and insulin. Presence of type 1 DM was ascertained by a diagnosis before 30 years of age, prescribing history showing continuous insulin prescription from diagnosis of DM and clinically defined type of DM.
	Scottish Index of Multiple Deprivation	1 = Most deprived → 5= Least deprived	The SIMD is an area-based measure of socio-economic status. It uses data from seven domains

	Variables	Coding	Notes
	(SIMD) quintile		considered important predictors of deprivation including; income, education, employment, health, housing, crime and geographic access to services. The index score for each data-zone is formed through the weighted summing of scores from each domain. There are 6,505 data-zones across Scotland which have a median population size of 769 and nest within local authorities. The data-zones are then divided into quintiles based upon their rank. Individuals are assigned a deprivation quintile by determining within which data-zone their postcode lies.
	Date of diagnosis		"Best-guess" date of diagnosis estimated from timing of treatment and earliest DM mention.
Repeated measurements	Body mass index (BMI)	Measured in kg/m^2	Calculated using standard approach of weight/height^2. The QOF recommends that BMI is measured and recorded in patients with DM at least every 15 months
	Height	Height in m (rounded to the nearest centimetre)	Recorded while patients are without shoes.
	Weight	Weight in kg	Recorded while patients are without shoes or outdoor clothing.
	Smoking Status	0 = Unknown, 1 = Current, 2 = Ex-smoker, 3 = Never smoker Recoded: 0 = Never, 1 = Ever	The QOF recommends that smoking status of DM patients is measured and recorded every 15 months.
	Diastolic and systolic blood pressure (BP)	mmHg	Resting blood pressure recorded using a sphygmomanometer. The QOF encourages general practices to record BP at least every 15 months. Practices are rewarded for achieving high percentages of patients with DM with a last blood pressure measurement of 145/85 or less. SIGN recommends patients should aim to achieve a diastolic BP of 80 or less and a systolic BP of less than 130.
	Glycated haemoglobin (HbA1c)	Percentage of total haemoglobin. Provides an estimate of average blood glucose level during the preceding 10-12 weeks.	High levels of HbA1c are indicative of poor control of blood glucose levels. The QOF recommends HbA1c is measured every 15 months in patients with

Variables	Coding	Notes
		DM. SIGN recommends patients with type 2 DM should aim to achieve a HbA1c value of 7.0% or below but good control is often defined as HbA1c <7.5%.
Serum creatinine	umol/L	Indicator of renal function. The QOF recommends patients with type 2 DM have serum creatinine measurements recorded every 15 months.
Serum total cholesterol	Mmol/L	Fasting total cholesterol value. The QOF recommends patients with type 2 DM have their total cholesterol recorded every 15 months and should aim to be at 5.0mmol/L or less.
Serum HDL-cholesterol	Mmol/L	Fasting HDL-cholesterol value. HDL-cholesterol is considered the 'good cholesterol' and therefore a HDL-cholesterol below 1.03mmol/L for men and below 1.29mmol/L for women is considered undesirable.
Serum triglycerides	Mmol/L	Fasting serum triglycerides level. A triglyceride level of below 1.7mmol/L is considered desirable.
Scottish Morbidity Records (SMR01)/ Hospital admissions	Diseases recorded using the International Classification of Diseases (ICD) – 9th revision codes for admissions prior to 1992 and ICD10 after this date.	

4.3 DATA CLEANING

A data cleaning exercise was undertaken with the aim of creating a single dataset upon which various missing data methods could be tested and applied. Analyses were restricted to patients with type 2 DM and who had a date of diagnosis prior to 01/07/2008.

The presence of type 2 DM was ascertained using an algorithm which used age at date of diagnosis (under 30 years) and the use and timing of treatment with oral hypoglycaemic agents and insulin to identify people with type 1 DM for exclusion. After exclusion of patients with any other form of DM (n=33,035), and who were diagnosed after study-endpoint (n=238) there were 241,091 patients with type 2 DM in the sample. A separate cohort of patients diagnosed with type 2 DM between 2004 and July 2008 was also generated (n=77,948).

Data cleaning was undertaken for the original cohort of 241,091 patients by checking for invalid readings. Within each dataset, each continuous variable was examined for implausible values (see Table 4.3), and these were set to missing. The upper and lower bounds were set in accordance with the SCI-DC data quality audit. Measurement dates were also examined for plausibility. All measurements taken prior to 02/01/1910 or after 01/07/2008 were considered implausible and were set to missing. Since data extraction for separate variables occurred at different times, the plausibility cut-off point was set to 01/07/2008 to obtain a snapshot of the data. Vital status was determined at 31/12/2012 to obtain a longer duration of follow-up. Measurements before 1910 were assumed implausible since these measurements were considered unlikely to still be available. Values taken before a patient's year of birth or after their date of death were also considered implausible.

The percentage of missing values was calculated for each variable to ascertain the extent of the missing data problem.

Table 4.3: Plausibility cut-offs for available continuous variables (obtained from SCI-DC data quality audit)

Variable	Lower Bound	Upper Bound
Age at diagnosis (years)	0	105
Diagnosis Date	02/01/1910	01/07/2008
Height (m)	1.0	2.1
Weight (kg)	15	200
BMI kg/m²	15	75
Systolic Blood Pressure (mm Hg)	80	400
Diastolic Blood Pressure (mm Hg)	40	300
Glycated Haemoglobin (%)	4	30
Serum Creatinine (umol/L)	40	1999
Total Cholesterol (mmol/L)	1.0	15
Triglycerides (mmol/L)	0.5	100
HDL-cholesterol (mmol/L)	0.5	6.0

4.3.1 FIXED VARIABLES

Datasets containing information on fixed demographic characteristics were examined for implausible or missing observations first. These variables were only measured once for each patient, e.g. age at diagnosis, sex, deprivation status and were likely to be key covariates in the final analysis model.

There were few implausible measurements in fixed variables but there were missing observations. Table 4.4 shows the level of completeness for each fixed variable after removal of implausible values in both the full sample and the 2004 to 2008 cohort. With the exception of ethnicity, all variables exhibited a completeness of over 96%. Ethnicity was poorly recorded (32% complete in full sample; 26% in the reduced 2004-2008 cohort) and so its inclusion in a CCA would substantially decrease sample size. However, since 95% of people in Scotland are White British and of the people with a recorded ethnicity, 95% were recorded as White British, it was deemed unlikely that the exclusion of this variable would have a large effect on final analyses (Office of the Chief Statistician, 2004).

Table 4.4: Data completeness of fixed variables in patients with type 2 DM

Variables	Full Cohort (n=241,091)		Reduced cohort, 2004-2008 (n=77,948)	
	Observed (%)	Missing (%)	Observed (%)	Missing (%)
Age at diagnosis	238,568 (98.95)	2,523 (1.05)	75,429 (96.77)	2,519 (3.23)
Sex	241,031 (99.98)	60 (0.02)	77,892 (99.93)	56 (0.07)
Ethnicity	78,006 (32.36)	163,085 (67.64)	20,587 (26.41)	57,361 (73.59)
Vital Status	241,091 (100.00)	0 (0.00)	77,948 (100.00)	0 (0.00)
Health Board Area	241,031 (99.98)	60 (0.02)	77,937 (99.99)	11 (0.01)
Deprivation status	239,425 (99.38)	1,666 (0.69)	77,407 (99.31)	541 (0.69)
Date of diagnosis	238,585 (98.96)	2,506 (1.04)	75,472 (96.82)	2,476 (3.18)

Age at diagnosis could not be calculated for 2,523 patients, predominantly because they were missing a DM diagnosis date. There were only 60 patients whose year of birth was unavailable, 56 of these patients were diagnosed with type 2 DM between 2004 and 2008. Upon further exploration, it was observed that the 60 patients with missing data for date of birth were the same 60 patients who had missing data for sex and Health board area. Of the 241,091 patients in the full sample, 236,973 (98%) had complete data for these fixed variables, excluding ethnicity. In the reduced sample this number equalled 74,933 (96%). Missingness in individual variables therefore did not combine to produce missingness in a large number of cases. There were 12 patients for whom all fixed variables except vital status were missing.

4.3.1.1 CHARLSON COMORBIDITY INDEX

To assess comorbidity levels and therefore relative 'healthiness' of each patient at date of diagnosis with type 2 DM, the Scottish Morbidity Records (SMR01) data were used to calculate a Charlson Comorbidity Index (CCI) for each patient on the DM register (Charlson et al., 1987). This score uses 19 pre-defined comorbid conditions to develop a weighted score which is then typically used to calculate the cumulative increased risk of one-year mortality (Table 4.5).

Table 4.5: Charlson Comorbidity Index (CCI) Scoring System by condition

Score*	Condition
1	Myocardial infarction
	Congestive heart failure
	Peripheral vascular disease
	Cerebrovascular disease
	Dementia
	Chronic pulmonary disease
	Connective tissue disease
	Peptic ulcer disease
	Mild liver disease (without portal hypertension, includes chronic hepatitis)
	DM without end-organ damage
2	Hemiplegia
	Moderate or severe renal disease
	DM with end-organ damage (retinopathy, neuropathy, nethropathy, or brittle DM)
	Tumour without metastases
	Leukemia
	Lymphoma
3	Moderate or severe liver disease
6	Metastatic solid tumour
	AIDS (not just HIV positive)

*Summed to calculate overall comorbidity index

To calculate a CCI at diagnosis for each patient the SMR-01 dataset, containing data on hospital admission, discharge dates and the patient's main condition and up to five other conditions was used. Hospital admissions occurring after date of diagnosis of type 2 DM were not considered. No restrictions on the length of look-back were in place since it was deemed unlikely that patients would be considered cured of these conditions without a significant amount of time passing – e.g. the majority of cancer patients require 10 years of remission to be adjudged cured. Admissions occurring prior to 1992 had disease codes corresponding to version 9 of the International Classification of Disease (ICD-9), whilst admissions after this date corresponded to ICD-10. Comorbidities were defined using ICD-9 and ICD-10 coding algorithms generated by Quan et al. (2005), which have been found to outperform their nearest rivals. Where a

patient had both a mild and a more severe form of a condition, only the most severe form was counted. This hierarchical structure ensures that a single type of comorbidity is not counted more than once in each patient. The weighted scores were summed to provide a final CCI, where higher scores reflect a more severe burden of comorbidity. Data were not available for people admitted to hospitals outside Scotland, though this number is likely to be small and should have a limited impact on scores.

Table 4.6: Frequency distribution of Charlson comorbidity index scores at diagnosis

Charlson Comorbidity Index (CCI)	Full Sample (n=238,585) Frequency (%)	Reduced Cohort (n=75,442) Frequency (%)
1	173,284 (72.63)	49,990 (66.26)
2	21,058 (8.83)	8,694 (11.52)
3	29,239 (12.26)	10,168 (13.48)
4	7,626 (3.20)	3,372 (4.47)
5	3,037 (1.27)	1,451 (1.92)
6	1,222 (0.51)	595 (0.79)
7	1,962 (0.82)	673 (0.89)
8	696 (0.29)	284 (0.38)
9	287 (0.12)	137 (0.18)
10	105 (0.04)	42 (0.06)
11	38 (0.02)	20 (0.03)
12	13 (0.01)	6 (0.01)
13	15 (0.01)	9 (0.01)
14	3 (0.00)	1 (0.00)

Of the 241,091 patients with DM, the hospital admissions dataset held records for 211,769 (93%) of them. For those patients without a hospital admission, they were assigned a CCI of 1 due to their diagnosis of DM. A CCI at diagnosis was not calculated for the 2,506 patients whose date of diagnosis was not available. Table 4.6 shows the frequency distribution of patient scores in both cohorts. In the full sample 72% of patients had a CCI of one, indicating that most patients had DM alone, with no other hospital record of a comorbid condition involved in the construction of the overall

score. This number reduced to 66% in the cohort of patients diagnosed between 2004 and 2008.

4.3.2 Repeat Variables

Datasets containing information on biochemical levels were examined next for problems and implausible measurements. A flow-chart of the data cleaning process for these repeatedly measured variables is presented in Figure 4.1. Within this process, each dataset was examined for duplicates, defined as either identical or non-identical. Identical duplicates were defined as readings that were identical to other readings in the dataset in terms of their identifier, date of measurement and measurement value. Non-identical duplicates occurred when there were multiple measurement values were available on a common date for the same individual. Where this was the case, the median value was used.

The data cleaning process outlined in Figure 4.1 was applied to each repeatedly measured variable, excluding BMI. Additional problems were encountered during data cleaning of the BMI variable. Details of the data cleaning process for the BMI variable are described in section 4.3.3.

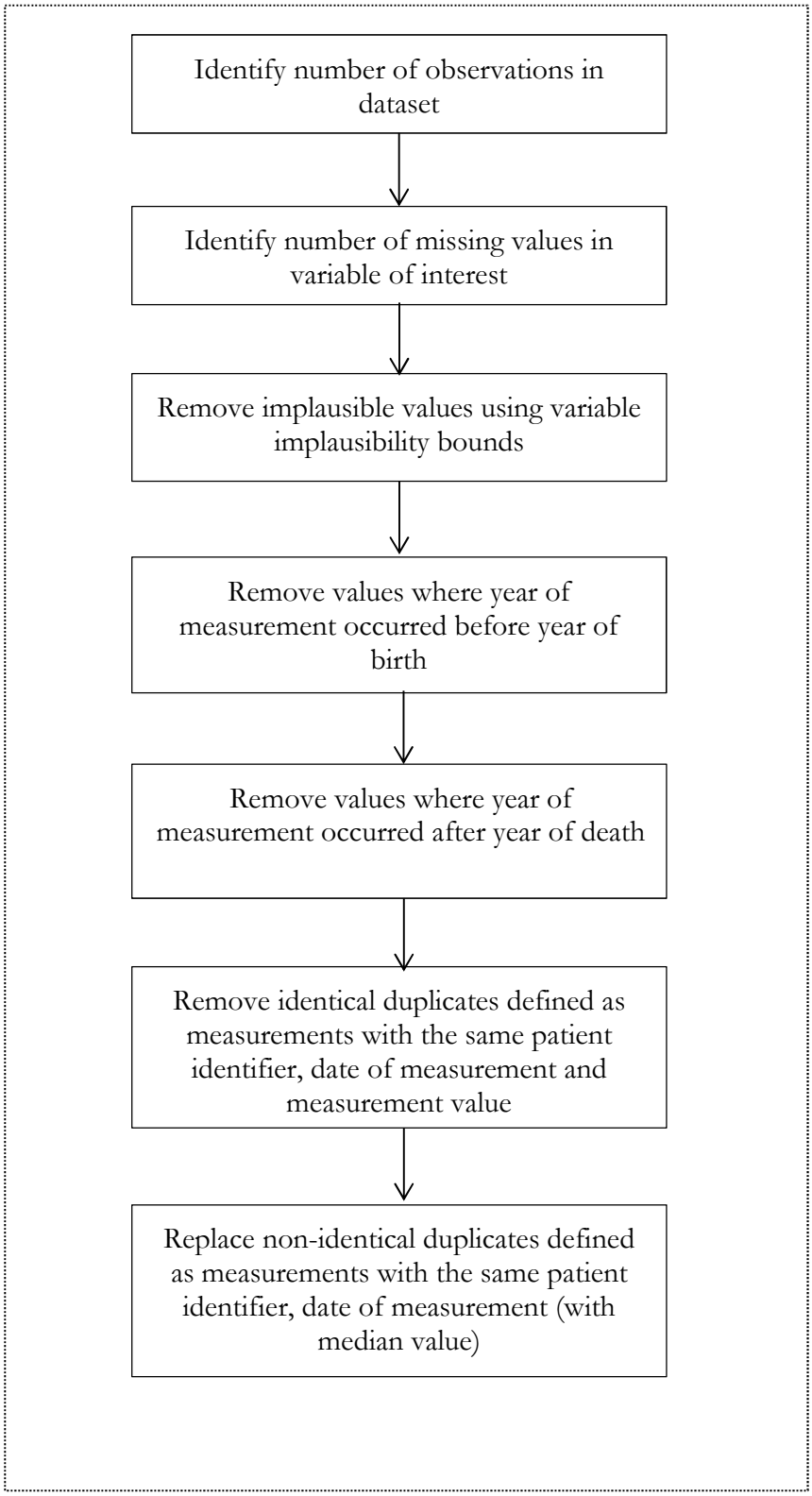

Figure 4.1: Data cleaning process for repeatedly measured biochemical characteristics in SCI-DC datasets

For each biochemical measurement, a baseline (at date of diagnosis) measurement variable was constructed. Baseline characteristics were defined as measurements taken closest to date of diagnosis rather than a measurement taken within a year of diagnosis. The decision to use this definition for baseline measurement was based upon balancing the likely harms of using either a measurement taken a considerable time after diagnosis to reflect baseline value or having a definition which was likely to result in large amounts of missing data. Table 4.7 presents the extent of the missing data problem when using the closest measurement compared to a measurement taken within one year of diagnosis. Table 4.8 presents the median time until measurement when the measurement taken closest to diagnosis with type 2 DM is used.

In general, completeness was considerably lower amongst repeatedly measured variables compared to fixed variables. Unsurprisingly, levels of completeness were highest amongst the traditional risk factors for CVD such as blood pressure and smoking status. In the full sample completeness of repeated measurements recorded within a year diagnosis was poor. Only systolic and diastolic blood pressure recordings surpassed 60% completeness. For example, only 53% of patients had a measurement of glycated haemoglobin recorded within a year of diagnosis. Amongst patients diagnosed between 2004 and 2008, completeness levels in measurements recorded within a year of diagnosis were much better than those observed in the full cohort. Variable completeness was typically over 85% between 2004 and 2008. Regardless of the defined time period, triglycerides and HDL-cholesterol values were relatively poorly recorded. In the full sample only 22% of patients had a recording of triglycerides within one year of diagnosis and this only improved to 33% in the 2004 to 2008 cohort.

In the full sample, 19% of patients had recordings for all biochemical measurements taken within one year of diagnosis. In contrast, 53% of patients had at least one recording of all repeatedly measured variables. Thirty five percent of patients diagnosed between 2004 and 2008 had complete data for all these repeated variables within a year of diagnosis, whilst 41% had complete data for closest measurements.

Table 4.7: Completeness of repeat variables based on measurements taken closest to date of diagnosis and within a year of diagnosis date

	Full sample (n=241,041)				2004-2008 sample (n=77,948)			
Variables	Closest measurements (%)		Measurements taken within 1 year of diagnosis (%)		Closest measurements (%)		Measurements taken within 1 year of diagnosis (%)	
	Observed	Missing	Observed	Missing	Observed	Missing	Observed	Missing
Smoking status	223,631 (93.75)	17,460 (6.25)	135,761 (56.32)	105,330 (43.68)	72,943 (93.58)	5,005 (6.42)	70,043 (89.56)	7,905 (10.14)
Systolic blood pressure	229,968 (96.41)	11,123 (3.59)	153,107 (63.52)	87,984 (36.48)	73,509 (94.31)	4,439 (5.69)	71,405 (91.61)	6,543 (8.39)
Diastolic blood pressure	230,012 (96.43)	11,079 (3.57)	153148 (63.54)	87,943 (36.46)	73,508 (94.30)	4,440 (5.70)	71,402 (91.60)	6,546 (8.40)
Glycated Haemoglobin	217,027 (90.98)	24,064 (9.02)	128,998 (53.52)	112,093 (46.48)	71,533 (91.77)	6,415 (8.23)	69,137 (88.70)	8,811 (11.30)
Serum Creatinine	218,265 (91.50)	22,826 (8.50)	125,49 (52.61)	115,601 (47.39)	72,644 (93.20)	5,304 (6.80)	70,125 (89.96)	7,823 (10.04)
Total cholesterol	215,379 (93.29)	25,712 (6.71)	126,274 (52.39)	114,817 (47.61)	72,838 (94.44)	5,110 (6.56)	70,635 (90.62)	7,313 (9.38)
Triglycerides	129,553 (54.31)	111,538 (55.69)	53,674 (22.27)	187,417 (77.73)	35,112 (45.05)	42,836 (54.95)	26,021 (33.38)	51,927 (66.62)
HDL-cholesterol	183,042 (76.73)	58, 049 (23.27)	86,771 (36.00)	154,320 (64.00)	61,802 (79.29)	16,146 (20.71)	54,721 (70.20)	23,227 (29.80)

Inspection of the median time between closest measurement and diagnosis revealed wide variations in length of time until measurement (Table 4.8). The median time until measurement was considerably longer amongst patients in the full cohort compared to the median times in the reduced cohort. For example, in the full cohort the median time until measurement of blood pressure was just over two months, whilst in the reduced cohort the median time was 15 days. The time until HDL-cholesterol and triglyceride measurement was over a year amongst the full cohort but again much reduced in the reduced cohort (<80 days). Clearly, a measurement taken a number of years before or after diagnosis with type 2 DM is unlikely to accurately reflect values at baseline. However, the median-time until measurement was deemed reasonable for the majority of the biochemical measurements and therefore these closest measurements were chosen to reflect these variables at diagnosis. Using these measurements also reduced the amount of incomplete data. Nonetheless, this definition of the time-frame for measurement is a potential source of bias which must be acknowledged when interpreting later analyses with these measurements.

Table 4.8: Median time between diagnosis and closest measurement date for repeatedly measured variables.

Variables (closest to date of diagnosis)	Median time between date of diagnosis and variable recording, days (IQR)	
	Full Sample	2004-2008 sample
Smoking status	144 (24, 943)	27 (6, 77)
Systolic blood pressure	74 (12, 714)	15 (2, 44)
Diastolic blood pressure	74 (13, 714)	15 (2, 44)
Glycated Haemoglobin	142 (22, 1,255)	22 (6, 72)
Serum Creatinine	161 (21, 1,344)	20 (6, 58)
Total cholesterol	141 (20, 1,324)	20 (6, 55)
Triglycerides	624 (81, 2,139)	79 (20, 381)
HDL-cholesterol	441 (42, 1,923)	35 (9, 128)

4.3.3 BMI

4.3.3.1 CLEANING BMI

Within the BMI dataset there were a large number of problems. This dataset had 9,925,067 rows of observations for patients with type 2 DM, with large proportions of missing data.

First, the dataset was checked for implausible height and weight values, defined in Table 4.3. A graph of height against weight was produced to check for further implausible values. In some instances, measurements of patient's respective height and weight on a single date were recorded on separate rows in the dataset. Accordingly, automated calculations of BMI had not been completed on these dates resulting in high missingness in the BMI variable (74%; 7,364,832 of 9,925,067), whilst missingness in height and weight was lower, 68% and 49%, respectively. The dataset therefore needed to be restructured so that all heights and weights recorded for a patient on a single date could be found on a single row. Once on a common row, height and weight measurements were used to calculate the patient's BMI for that date. Where patients had more than one measure of height or weight on a common date, the median value was used. For patients with weight measurements but no height measurement on a common date, the patient's median height from available height data was imputed, since height is unlikely to vary considerably over time in this adult population. BMI values falling below 15kg/m^2 or above 75kg/m^2 were considered implausible and set to missing, a process which reduced the number of rows to 3,254,193.

The number of measurements per patient is presented in Table 4.9. Of the 241,091 patients, there were 20,656 (9%) patients without a single measure of BMI. Over 35% of patients had more than ten measurements. Similar patterns were observed in the reduced cohort. However, fewer patients had more than ten measurements of BMI (14% vs. 36% in full cohort). This probably reflects a shorter follow-up time in the reduced cohort of patients diagnosed with type 2 DM between 2004 and 2008.

Table 4.9: Number of available BMI measurements per patient

Number of measurements	Frequency (%)	
	Full sample	2004-2008 sample
0	20,656 (8.57)	5,951 (7.63)
1	12,411 (5.15)	3,876 (4.97)
2	11,550 (4.79)	5,404 (6.93)
3-10	110,503 (45.83)	51,564 (66.15)
11-20	64,459 (26.74)	10,087 (12.94)
>20	21,512 (8.92)	1,066 (1.37)

4.3.3.2 BMI AT DATE OF DIAGNOSIS

The aim of this project was to illustrate the use of missing data on an analysis of the association between BMI at diagnosis with type 2 DM and all-cause mortality. Using BMI measurements recorded on diagnosis date was not viable since only 8% of patients in the full sample had a valid BMI at diagnosis. It was therefore necessary to create a time-window during which BMI measurements could be used which accurately reflected the predictor of interest at diagnosis. To define this BMI time-window, a balance between limiting the potential impact of weight changes and limiting the proportion of unobserved data was sought.

Definitions which included measurements taken a considerable time after diagnosis were likely to be biased by treatment effects and the potential development and progression of comorbid conditions. Following a diagnosis of type 2 DM, patients are encouraged to make lifestyle changes e.g. healthier eating, smoking cessation) and may be prescribed glycaemic control therapies that have been demonstrated to be associated with weight change. To restrict the impact of these biases, researchers investigating the relationship between BMI and all-cause mortality have used BMI recordings measured within one year prior to diagnosis. Unfortunately, using this definition would result in complete data for only 22% of patients in the full sample. To achieve greater data completeness, it was necessary to use measurements recorded after diagnosis to reflect patient BMI at diagnosis. Of the full sample, 83% of patients had a BMI recording after diagnosis thus representing a significant proportion of patients.

To assess how accurately BMI measurements recorded before or after diagnosis date reflected BMI at diagnosis, patient measurements recorded within 2 weeks of diagnosis were compared to the respective patients measurement outside this time-period (Table 4.10).. Mean differences presented in Table 4.10 indicate measurements recorded before diagnosis more accurately reflected BMI measurements at diagnosis than measurements recorded after diagnosis. However, the overall correlation between measurements recorded before or after diagnosis was high and generally, as time from diagnosis decreased the level of correlation increased.

There were two major limitations of this assessment. Firstly, only 18% of patients from the full sample had a BMI measurement within 2 weeks of diagnosis. Secondly, it is

likely that patients with available recordings within these time-intervals are different to those that do not. For example, patients with more recordings may have more variable BMI patterns due to weight loss attempts. Patient weight changes may therefore prompt increased frequency of measurements.

Table 4.10. Comparison of measurements recorded at diagnosis and measurements recorded before or after diagnosis

Measurement time-frames	N	Mean difference between measurements (SD)	Pearsons correlation
Diagnosis vs. 1 month before diagnosis	2,268	- 0.15 (1.52)	0.974
Diagnosis vs. 2 months before diagnosis	4,828	- 0.15 (1.64)	0.969
Diagnosis vs. 3 months before diagnosis	6,512	- 0.16 (1.86)	0.960
Diagnosis vs. 6 months before diagnosis	9,767	- 0.16 (1.95)	0.956
Diagnosis vs. 1 year before diagnosis	14,655	- 0.12 (2.04)	0.954
Diagnosis vs. 1 month following diagnosis	4,962	0.23 (1.66)	0.970
Diagnosis vs. 2 months following diagnosis	12,174	0.29 (1.71)	0.967
Diagnosis vs. 3 months following diagnosis	17,776	0.33 (1.75)	0.965
Diagnosis vs. 6 months following diagnosis	27,464	0.35 (1.84)	0.960
Diagnosis vs. 12 month following diagnosis	35,100	0.34 (1.90)	0.956

Due to the inherent limitations of this exploratory work, it was difficult to ascertain an appropriate time-frame to use as a definition for BMI at diagnosis. Furthermore, data from previous studies have reported mixed BMI trends in the year prior to diagnosis (see Section 3.4). Whilst Looker et al. (2001) suggested patient BMI generally increases within one year prior to diagnosis, Vistisen et al. (2014) and De Fine Olivarius et al. (2008) observed rapid weight losses in the year prior to diagnosis. Similarly, data from previous work for weight changes following diagnosis with type 2 DM do not show a consistent trend (Looker et al., 2001; de Fine Olivarius et al., 2006; Chiu et al., 2013). A suitable definition for BMI at diagnosis was therefore arbitrarily defined as a measurement taken 10 months prior to or 2 months after diagnosis with type 2 DM date. Where multiple measurements were available within this time-frame, the measurement taken closest to the date of diagnosis was used. This definition provided a year-long window for measurement whilst also limiting the amount of unobserved data and the potential for bias. The impact of using this definition for BMI at diagnosis is explored further in Chapter 8.

Of the 241,091 type 2 DM patients in the full sample, there were 100,979 (42%) with a valid BMI measurement at date of diagnosis, leaving 140,112 patients (58%) without a valid measurement of BMI. In the reduced cohort 56,990 (73%) patients had an observed BMI at diagnosis and 20,958 (27%) patients were without.

4.3.3.3 *PREDICTED BMI AT DATE OF DIAGNOSIS*

The majority of patients had 2 or more measurements and therefore it was possible to use these measurements to obtain an extrapolated or interpolated estimate BMI at diagnosis. This new variable could be used to ascertain the potential for missingness in BMI at diagnosis being MNAR, dependent on the missing BMI values themselves. To create the variable individual level-linear regressions were applied using individual BMI measurement values and timing of measurement from diagnosis with DM. The value of this variable was the intercept of the regression equation estimated for each patient. Figure 4.2 presents this process for a patient with 11 available measurements.

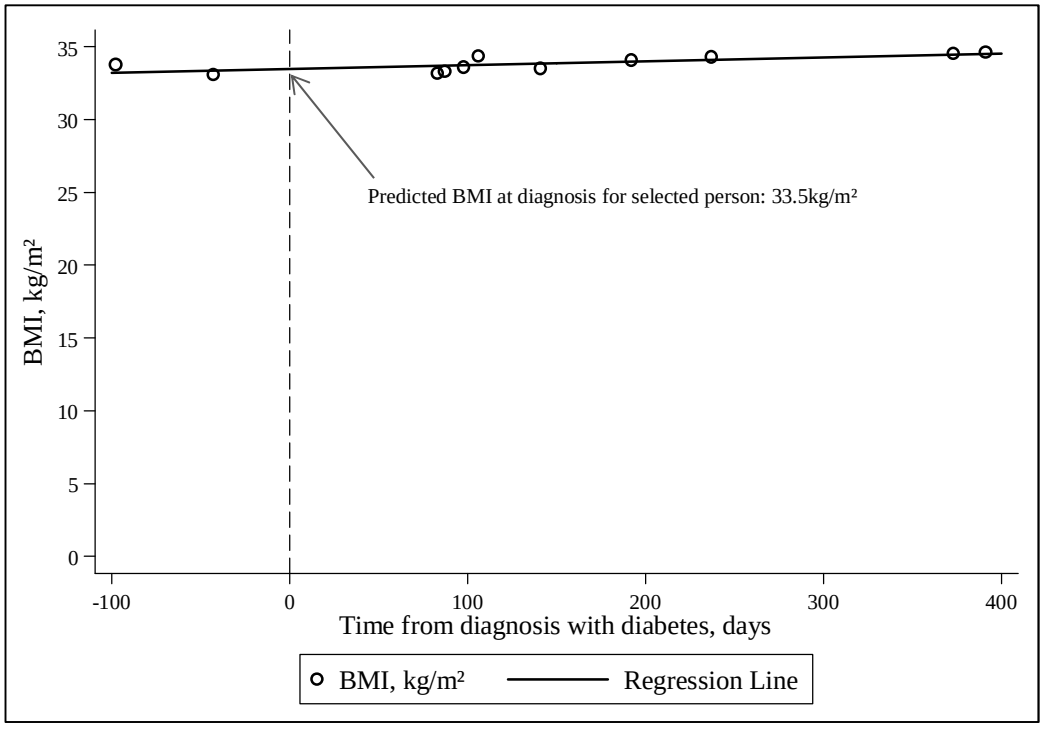

Figure 4.2: Production of predicted BMI in a patient with type 2 DM

This method had a number of problems. For example, this approach led to the development of implausible BMI estimates, especially where measurements were extrapolated from earlier or later measurements. This issue was particularly severe amongst patients whose only available measurements were recorded a long time before

or after diagnosis date. If these measurements showed a slight incline or decline in BMI, the estimated BMI value at date of diagnosis was considerably different from the observed values. Indeed, some predicted BMI values were negative. Figure 4.3 presents this problem for a patient whose three measurements were taken over six months after diagnosis with DM. For this patient, the predicted BMI at diagnosis (23.9kg/m²) was considerably lower than the patient's observed measurements. This figure also presents a large amount of variability in BMI measurements over a short period of time, suggesting possible measurement error.

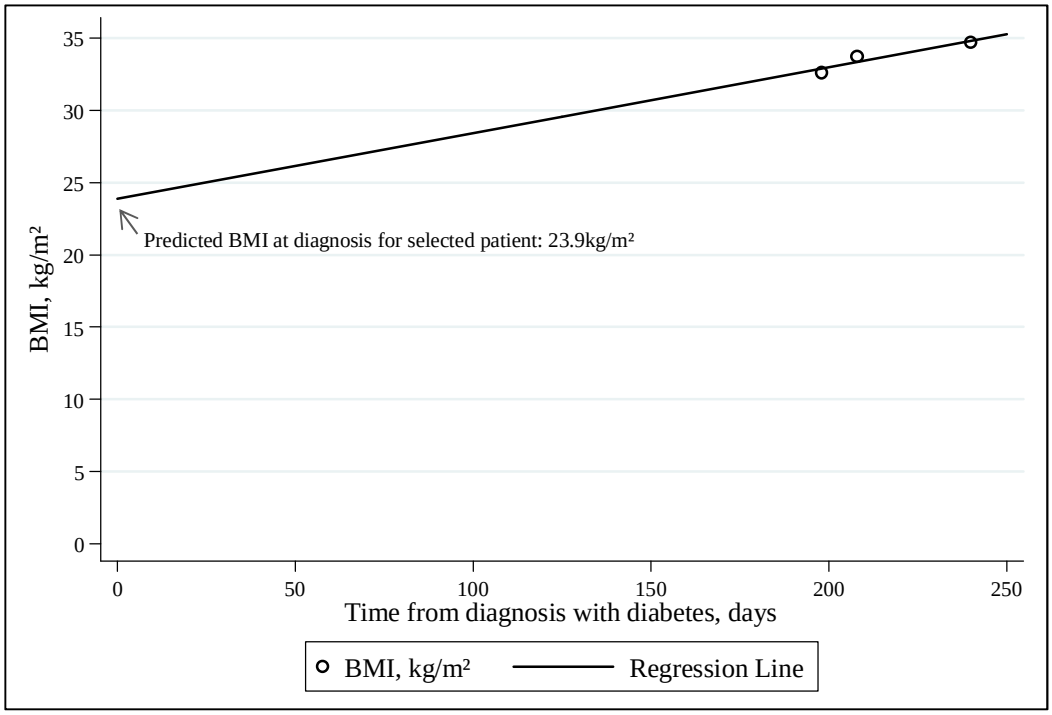

Figure 4.3: Illustration of potential problems associated with using individual level linear regression to predict BMI at date of diagnosis.

There were also 20,656 patients in the full cohort that were without any BMI measurements and therefore a BMI at diagnosis with DM could not be estimated. A further 9,273 patients only had one BMI measurement and so their predicted BMI value was the same as their observed value. For people with an observed BMI at diagnosis with DM, the mean difference between observed BMI values at date of diagnosis and predicted BMI from the individual-level regression models was 0.23kg/m^2 (SD: 1.56). This difference is relatively small and therefore this method appears to provide reasonably valid estimates of BMI at date of diagnosis of DM. Nonetheless, it is

uncertain whether this difference would be the same for patients without a BMI measurement at diagnosis.

4.3.3.4 CREATION OF MEDIAN BMI

Since there were numerous problems with the creation of a predicted BMI using linear regression, a variable which used the median BMI value across measurements was created. This variable had the advantage of creating plausible measurements. However it could not reflect changes in BMI over time. This variable will be described as Median Ever-recorded BMI in later discussions.

4.4 COHORT CHARACTERISTICS

Table 4.11 provides a description of the baseline characteristics of the full cohort and for those patients diagnosed between 2004 and 2008 after data cleaning. The mean age at diagnosis with DM was 61.5 years in the full sample, but slightly higher in the 2004 - 2008 cohort. This finding in the most up-to-date cohort is surprising given that the typical age at which patients are diagnosed with DM is declining (Alberti et al., 2004). There were a higher proportion of males (56%) with DM than females (46%), a trend that has been observed in several different DM populations (Lipscombe & Hux, 2007; Sarwar et al., 2010). A higher prevalence of DM was observed in the most deprived groups. The mean BMI at diagnosis was 31.6kg/m² in the full cohort, while it was slightly higher in the reduced cohort (31.9kg/m²). This is perhaps an indication that people are being diagnosed at heavier weights in recent times. The prevalence of ever-smokers at time of diagnosis was high at 57% in the full cohort and 42% in the reduced cohort. The distribution of DM prevalence by HBA was similar to the general population distribution (Table 4.12) (National Records of Scotland, 2013). However, the number of people with DM in Lothian was slightly smaller (13%) than its population distribution (16%).

Table 4.11: Characteristics of people within full sample and 2004 to 2008 cohort

Characteristic		Full sample	2004-2008 sample
Mean age at diagnosis, yrs (SD)		61.48 (12.57)	62.21 (12.65)
Male sex (%)		129,177 (53.59)	42,410 (54.45)
Deaths prior to 31/12/11 (%)		62,339 (25.86)	14,317 (18.37)
SIMD (%)	1 (Most deprived)	56,953 (23.79)	18,435 (23.82)
	2	54,914 (22.94)	17,621 (22.76)
	3	49,519 (20.68)	15,990 (20.66)
	4	43,670 (18.24)	14,053 (18.15)
	5 (Least deprived)	34,369 (14.35)	11,308 (14.61)
Year of diagnosis, (%)	Before 1995	52,419 (21.74)	-
	1995 – 1999	45,648 (18.93)	-
	2000 – 2004	86,576 (35.91)	-
	After 2004	56,448 (23.41)	-
BMI at date of diagnosis, kg/m^2 (SD)		31.60 (6.43)	31.92 (6.52)
Mean of (median) ever-recorded BMI, kg/m^2 (SD) †		30.75 (6.11)	31.48 (6.27)
Mean of predicted BMI, kg/m^2 (SD)		30.89 (29.48)	31.57 (7.99)
Median Follow-up time, years (IQR)		5.92 (2.91, 10.44)	5.55 (4.41, 6.77)
With additional comorbidity (%)			27,958 (35.87)
Measurements taken at closest time to diagnosis with DM			
Ever-smokers (%)		128,159 (57.31)	30,513 (41.83)
Mean Systolic BP, mmHg (SD)		143 (21.04)	139 (18.89)
Mean Diastolic BP, mmHg (SD)		82 (12.19)	80 (10.98)
Mean Glycated Haemoglobin, % (SD)		7.93 (2.00)	8.00 (2.10)
Mean serum creatinine, umol/L (SD)		94.58 (34.56)	91.11 (27.77)
Mean total cholesterol, mmol/L (SD)		5.25 (1.27)	5.15 (1.30)
Mean triglycerides, mmol/L (SD)		2.58 (2.51)	2.52 (2.60)
Mean HDL cholesterol, mmol/L (SD)		1.24 (0.42)	1.24 (0.40)

Table 4.12: Distribution of patients with DM in Scotland by HBA

Health Board Area	Full sample	Reduced cohort (2004 – 2008)
Ayrshire & Arran	18,814 (7.81)	7,031 (9.02)
Borders	5,388 (2.24)	1,613 (2.07)
Dumfries & Galloway	7,971 (3.31)	2,297 (2.95)
Fife	17,795 (7.38)	5,423 (6.96)
Forth Valley	13,980 (5.80)	4,291 (5.51)
Grampian	22,567 (9.36)	8,029 (10.30)
Greater Glasgow	56,189 (23.31)	18,303 (23.48)
Highlands	13,259 (5.50)	4,224 (5.42)
Lanarkshire	26,114 (10.83)	9,580 (12.29)
Lothian	32,784 (13.60)	10,183 (13.07)
Orkney	945 (0.39)	334 (0.43)
Shetland	977 (0.41)	298 (0.38)
Tayside	23,111 (9.59)	5,977 (7.67)
Western Isles	1,137 (0.47)	354 (0.45)

5. PLAUSIBILITY OF MISSING DATA MECHANISMS

5.1 INTRODUCTION

In large, routinely-collected clinical databases there is often limited information available on the mechanisms giving rise to missing data; especially when data are collected on a national scale. Geographical variations in recording practices across large areas make it difficult to surmise or predict reasons for missingness. Changing recording practices and motivations for recording information add additional complexity to this problem. Furthermore, unlike in studies where data collection is conducted by members of a research team, gathering specific reasons for missingness, such as lack of patient motivation is untenable in studies which use routinely collected health records. There are also few opportunities to collect additional covariates which may later be useful for predicting missing values.

The mechanism of missingness dictates the appropriateness of any given method for handling partially observed datasets. From a practical perspective, it is not possible to verify with any certainty the mechanism giving rise to missing data; as to do so requires knowledge of the unobserved values. Despite this, prior knowledge of the data collection methods and data exploration exercises are useful in identifying the likely mechanisms by which missing data occurred. This chapter presents an assessment of the potential mechanisms giving rise to missing BMI. Findings from this chapter are then used in the assessment of the appropriateness of different missing data methods applied in later chapters.

In addition, this chapter also presents findings from investigations into the possible predictors of the BMI at diagnosis variable. These analyses were used to identify important variables to be included in missing data models to help impute plausible values for BMI.

5.2 MECHANISMS OF MISSINGNESS & THE SCI-DC DATA

In this population-based register, it is likely that all three mechanisms of missingness were present to varying degrees. Definitions of each mechanism are provided in section 1.1.4. This section elucidates initial notions regarding the likelihood of each missing data

mechanism. Here, examples of potential missing data mechanisms specific to the BMI at diagnosis variable within the SCI-DC dataset are presented.

5.2.1 MCAR DATA

For BMI data at diagnosis with DM to be MCAR, defined as missingness which does not depend on observed or unobserved measurements, missingness in BMI would have been dependent on chance alone. Although it was probable that there may have been instances whereby BMI at diagnosis was not recorded through chance occurrences, such as broken weighing scales, it seemed unlikely that these occurrences accounted for such high levels of incomplete data (58% of BMI records around time of diagnosis of type 2 DM were missing).

5.2.2 MNAR DATA

For data to have been MNAR, missingness in BMI at diagnosis would have been dependent on some unobserved data; either the missing BMI value itself or on other unobserved variables. For example, it seemed plausible that patient's whose weight was deemed normal would not have been weighed or that even if they were weighed, the measurement was not recorded. Meanwhile, people who were perceived to be under- or overweight may have been monitored more closely. In some cases the patient's weight may even have been above the maximum weight recorded on scales. Nonetheless, the availability of multiple repeated measures of BMI could be protective against the risk of missingness at diagnosis with DM being dependent on the missing value itself. For example, BMI values before or after the date of diagnosis may be used to indicate whether patients with 'healthier' weights before or after diagnosis were more likely to have missing BMIs at the diagnosis date. Data would therefore be MAR dependent on BMI at earlier or later time-points.

The risk of MNAR data was likely to have been reduced through the implementation of the QOF in 2004, when GPs began receiving financial incentives for the improved recording of risk factors. Prior to 2004, BMI data was primarily measured and recorded when it was clinically relevant. Accordingly this data were more likely to be MNAR or

MAR. For patients diagnosed after this time, the mechanism of missingness may have been different.

It was also possible that unobserved variables may have played a role in missingness in the BMI at date of diagnosis variable. For example, missing BMI at diagnosis data may have been dependent on the GP to which the patient attended as different practices are likely to have different procedures. Since we only have data on health board areas and not individual practices, these data would be MNAR. In addition, when the dataset was set up in 2000, data were retrieved retrospectively, where they were available. Different general practices/health board areas may have had different look-back periods. Consequently, the mechanisms of missingness may be different before and after SCI-DC's development. For patients diagnosed prior to the development of SCI-DC, missingness may be related to GP surgery; whilst for those diagnosed after its development, other factors may have governed missingness.

5.2.3 MAR DATA

Given the number of variables available in this linked dataset, it seemed plausible that data were MAR, whereby missingness in the BMI variable was related to the observed data. As an example, the proportion of data on BMI at diagnosis that were missing could have varied with patient age, sex or socio-economic status. Previous studies have identified associations between failure to attend appointments and younger age and lower socio-economic status (Waller & Hodgkin, 2000). If these relationships were found to hold in the type 2 DM cohort therefore resulting in higher proportions of missing data in these groups, the MAR assumption would hold. In addition health status at date of diagnosis could impact on missingness. Patients who are too ill at the time of diagnosis may be less likely to attend appointments. Information on the burden of comorbid conditions was available and therefore the likelihood of MAR dependent on 'healthiness' could be crudely assessed using this. It could not, however, account for mild and transient illness, such as a cold or a bad back or people with limited mobility whose height or weight could not be measured with the available equipment.

An additional variable that missingness in BMI could be dependent on was year of diagnosis. More specifically, missingness may have been related to whether diagnosis

occurred before or after the implementation of the QOF. Missingness may be highest amongst patients who were diagnosed prior to its implementation.

5.3 CLINICIANS' VIEWS ON EXPLANATIONS FOR MISSING DATA

5.3.1 INTRODUCTION & METHODS

In an effort to understand the potential reasons for missing BMI data at diagnosis in the SCI-DC register, I carried out five informal interviews with four clinicians and a data-entry clerk within primary and secondary care clinics. These interviews were undertaken within a GP in West Lothian and within an out-patient DM clinic at the Western General Hospital, Edinburgh. The interviews were structured with two aims: firstly to obtain an understanding of the processes by which clinicians collected and recorded key DM information, and secondly to understand why some patients may be missing data. To achieve these aims I conducted five face-to-face semi-structured interviews ranging from 15 minutes to 1 hour long in duration. Interviews were recorded and were later transcribed.

5.3.2 RECORDING PROCEDURES

From the interviews it was apparent that strict recording procedures are in place to ensure patients with DM are seen regularly and important risk factors are measured.

Within each setting, patients are seen at least annually. During these review appointments, important health indicators specified by the QOF are measured. This ensures that patients have at least one recording of each important risk factor every year.

In both settings, patients whose glycated haemoglobin is considered to be poorly controlled (defined by QOF as above 7.5% in the 2011/2012 revision, although individual targets are frequently set) may be recalled regularly until their glycated haemoglobin lies at a satisfactory level. At the GP, the practice nurse stated that, at the patient's follow-up, no further readings of other recordings are taken unless a computer prompt informs them of due measurements. In comparison, each time a patient attends the out-patient clinic they are weighed and blood and urine samples are taken.

To guarantee patients are seen by a clinician at least annually, each setting has a recall system which ensures letters are sent out to patients one month prior to their review date. These letters specify an appointment time or advise patients to call the surgery to arrange an appointment. If patients miss an appointment three reminders are sent out. In the outpatient clinic, the consultant will send a letter to the patient's GP if there is still no response. The practice nurse stated that patients rarely miss their appointments and there are no discerning characteristics of those that do.

Within each setting, a defined process of data recording is undertaken when a patient is given a diagnosis of DM. When patients are initially diagnosed within the GP setting, they are asked to return to the surgery within 2 weeks to measure important risk factors. If a patient's glycated haemoglobin level is considerably higher than the recommended level, the patient would be seen within 1 to 2 days of diagnosis.

To measure BMI, the patient is weighed and this recording is entered into the patient's electronic health record. In neither setting would a self-reported weight be entered. BMI is calculated automatically using height. According to guidelines, weight measurement should be standardised; shoes and outdoor clothing should be removed. However, upon observation of this process it seems there are wide variations in measurement practices, increasing the risk of substantial measurement error. Within the out-patient clinic height is only repeatedly recorded when there is a clear shrinking in size or when a patient is under 21 years of age. Each setting had a feature on the electronic system to highlight implausible BMI values.

There had been no noticeable changes in recording practices (even after the introduction of the QOF) whilst the interviewees had been employed in their roles, a period which stretched to 12 years for one interviewee.

5.3.3 REASONS FOR MISSING DATA

Some clinicians surmised that potential reasons for missing data around date of diagnosis could reflect uncertainty surrounding the accuracy of the date of diagnosis, rather than the failure to record important key indicators at this time point. As previously discussed, a strict recording process for each newly diagnosed patient is in place which should ensure important measurements, including BMI, are recorded within

2 weeks of date of diagnosis. Indeed, the practice nurse stated that upon diagnosis the vast majority of patients are quite concerned about their new diagnosis and are typically very compliant.

The practice nurse could not think of any reason for patients to have missing BMI recordings. In hindsight, this response may have been due to a misunderstanding of the question posed. The aim of this question was to attempt to understand why missing data may occur in general, whilst the practice nurse may have interpreted it as why she, personally may not have recorded BMI data. To offer suggestions may have been perceived to be an admission of her not doing her job correctly. Other clinicians were unable to provide possible explanations for the presence of missing data.

5.4 Plausibility of MCAR & MAR assumptions - Findings from data

5.4.1 Introduction

An assessment of the likely mechanisms of missingness in the SCI-DC register was conducted using all available data. Through an exploration of the data, this subsection aimed to identify the most likely mechanisms of missingness for BMI at diagnosis.

5.4.2 Methods

To assess the plausibility of the MCAR assumption, comparisons between patients with and without observed BMI measurements at diagnosis with type 2 DM were made. Systematic differences in demographic and clinical characteristics between these groups would indicate a likely violation of the MCAR assumption. Analyses based on patients with complete data would therefore be undertaken on a sample which would not be representative of the true population.

To compare groups, a missing indicator variable was created which was equal to zero if the patient had a BMI recording at diagnosis and one if the patient did not. Means are presented for normally distributed continuous variables and percentages for categorical variables by missing BMI status. Where continuous variables were not normally distributed, medians and interquartile ranges (IQR) are presented. Formal significance

tests were not undertaken for these comparisons due to large sample sizes. Even small changes between the groups would be statistically significant but not necessarily clinically relevant.

To assess the plausibility of the MAR assumption, univariate and multivariate logistic regression analyses were carried out using a missing indicator variable. The aim of these regressions was to determine which variables were associated with missingness in the BMI variable. Identification of such variables would then inform their likely inclusion in missing data models. To assess the plausibility of data being MNAR, within-patient summary measures of available BMI measurements (median ever-recorded BMI and linear regression predicted BMI) were included in univariate and multivariate logistic regression analyses. These variables were used to indicate possible BMI ranges at which rates of missingness were the highest.

Putting age into models as a continuous linear effect was considered potentially misleading as treatment patterns and healthcare-seeking practices were likely to differ across age in a non-linear fashion. For this reason, it was decided to include age as a categorical variable. Age at diagnosis of DM was divided into the following 5 age groups: under 50 years, 50 to 59 years, 60 to 69 years, 70 to 79 years and 80 years or over. There were 43,746, 58,959, 69,307, 49,085 and 17,471 patients in each respective age group. The under 50 years age group was chosen as the reference category because it was the extreme value with the largest numbers of people in it.

Year of diagnosis was also categorised. The introduction of the QOF in 2004 and the formation of SCI-DC in 2000 onwards were likely to produce non-monotonic changes in risk of missingness by year of diagnosis. Categories were formed to ensure odds ratios effectively described these distinct time-points whilst ensuring group sizes were not overly unbalanced. These categories were: Before 1995, 1995 to 1999, 2000-2004 and after 2004.

Charlson comorbidity index (CCI) is presented as both a continuous and a dichotomised variable. To form the binary variable patients were divided into two groups: CCI=1 'No comorbidities', CCI > 1 'Comorbidities'. Patients in the 'no comorbidities' category therefore had DM only, whilst patients in the other category had at least one other important comorbid condition. By grouping patients into these risk

categories, we were able to assess the risk of missingness in relatively unhealthy DM patients and a comparatively healthy DM population.

Reference categories for other categorical variables were chosen on the basis of having large numbers and/or lowest perceived risk of missingness in BMI at date of diagnosis.

STATA, version 11.2, College Station, Texas, StataCorp LP was used to run all analyses (StataCorp, 2009).

The variables that were included in the multivariate model were:

- Age at diagnosis (categorical)
- Sex (binary)
- Vital status (binary)
- Year of diagnosis (categorical)
- Within-patient Median ever-recorded BMI (continuous)
- Diastolic blood pressure around date of diagnosis (continuous)
- Glycated haemoglobin around date of diagnosis (continuous)
- Smoking status (binary: never/ever)
- Health Board Area (HBA) (categorical)
- Deprivation status (categorical)
- CCI (categorical)
- HDL-cholesterol around date of diagnosis (continuous)
- Total cholesterol around date of diagnosis (continuous)

Since the logistic regression models used CCA as the method for handling missing data, triglycerides and serum creatinine were not included in the multivariate model since their inclusion would result in a sample size reduction of almost 50%. It was deemed unlikely that these variables would be meaningful in terms of their ability to affect missingness in the BMI variable. Diastolic blood pressure (BP) and predicted BMI were also excluded from the multivariate model due to concerns regarding collinearity with systolic BP and median ever-recorded BMI, respectively.

The following interactions were also investigated graphically and by the inclusion of interaction terms in logistic regression models.

- Age at diagnosis vs. Sex
- Age at diagnosis vs. Deprivation status

- Deprivation status vs. Sex
- Charlson comorbidity index vs. Age at diagnosis
- Charlson comorbidity index vs. Vital status

The relationship between survival and missing BMI data was also explored using a Cox proportional hazards model, adjusted for age, sex, smoking status and deprivation status. The validity of the proportional hazards assumption was checked using Kaplan-Meier and log-log plots. Censoring was assumed to be non-informative. Informative censoring occurs when the censoring process is dependent on the event occurrence leading to biased estimates of survival. Censored patients in this study were deemed to have similar survival to those that were not censored. Age standardised mortality rates by BMI missingness were calculated using the Standard European population.

5.4.3 RESULTS I: PREDICTORS OF MISSINGNESS

In this cohort of patients with type 2 DM, 140,112 (58.12%) of the 241,091 patients did not have a BMI measurement recorded around date of diagnosis with DM. Table 5.1 presents some important differences between patients with and without a BMI recorded at diagnosis of type 2 DM.

Table 5.1. Factors influencing missingness in BMI at diagnosis in patients with type 2 DM in Scotland. Odds ratios (OR) represent odds for having missing BMI data at diagnosis

Variable		Observed BMI (n=100,979, 41.88%)	Missing BMI (n=140,112, 58.12%)	Univariate models OR (95% CI)	Multivariate model OR (95% CI)
Age at diagnosis, years (%)	< 50	17,252 (37.29)	29,017 (62.71)	1 (Ref)	1 (Ref)
	50 – 59	24,366 (41.33)	34,53 (57.36)	0.84 (0.82, 0.87)	0.93 (0.90, 0.96)
	60 – 69	30,420 (43.89)	38,887 (56.11)	0.76 (0.74, 0.78)	0.85 (0.82, 0.88)
	70 – 79	22,286 (45.40)	26,799 (54.60)	0.71 (0.70, 0.73)	0.86 (0.82, 0.90)
	≥80	6,655 (38.09)	10,816 (61.91)	0.97 (0.93, 1.00)	1.01 (0.96, 1.07)
Sex (%)	Female	45,611 (40.78)	66,243 (59.22)	1 (Ref)	1 (Ref)
	Male	55,357 (44.98)	73,820 (55.02)	0.91 (0.89, 0.93)	0.93 (0.91, 0.95)
Vital status on 31/12/2011 (%)	Alive	78,508 (51.11)	75,087 (48.89)	1 (Ref)	1 (Ref)
	Dead	22,471 (25.68)	49,437 (74.32)	3.03 (2.97, 3.08)	2.22 (2.14, 2.30)
SIMD (%)	1 (Most deprived)	24,429 (42.89)	32,524 (57.11)	1.01 (0.98, 1.04)	1.03 (0.99, 1.07)
	2	23,256 (42.35)	31,658 (57.65)	1.03 (1.01, 1.06)	1.04 (1.01, 1.08)
	3	20,133 (43.25)	29,192 (58.95)	1.09 (1.06, 1.12)	1.06 (1.03, 1.11)
	4	17,586 (40.27)	26,084 (59.56)	1.13 (1.09, 1.16)	1.09 (1.06, 1.14)
	5 (Least deprived)	14,829 (43.15)	19,540 (56.85)	1 (Ref)	1 (Ref)

Table 5.1 continued…

Variable		Observed BMI (n=100,979, 41.9%)	Missing BMI (n=140,112, 58.1%)	Univariate models OR (95% CI)	Multivariate model OR (95% CI)
Year of diagnosis (%)	Before 1995	2,824 (5.39)	49,595 (94.61)	57.96 (55.54, 60.49)	58.74 (55.20, 62.50)
	1995 – 1999	9,607 (21.05)	36,041 (78.64)	12.38 (12.02, 12.76)	13.08 (12.49, 13.69)
	2000 – 2004	45,227 (52.24)	41,350 (47.76)	3.02 (2.95, 3.09)	3.43 (3.29, 3.57)
	After 2004	43,322 (76.75)	13,126 (23.25)	1 (Ref)	1 (Ref)
Within-patient mean ever-recorded BMI, kg/m^2 (SD) †		31.31 (6.13)	30.27 (6.05)	0.97 (0.97, 0.98)	0.99 (0.98, 1.00)
Mean predicted BMI at diagnosis, kg/m^2 (SD)		31.40 (6.13)	30.87 (6.97)	0.99 (0.986, 0.989)	-
Median Follow-up time, years (IQR)		6.72 (4.88, 8.98)	11.00 (6.95, 16.00)	1.184 (1.181, 1.186)	1.32 (1.31,1.33)
Measurements recorded around date of diagnosis					
Presence of comorbid conditions, derived from CCI) (%)	None	69,225 (39.95)	104,059 (60.05)	1 (Ref)	1 (Ref)
	One or more	31,754 (46.83)	36,053 (53.17)	0.76 (0.74, 0.77)	0.99 (0.96, 1.01)
Smoking status (%)	Never smoker	40,521 (42.44)	54,951 (57.56)	1 (Ref)	1 (Ref)
	Ever smoker	59,810 (46.67)	68,349 (53.33)	0.84 (0.83, 0.86)	0.91 (0.89, 0.93)
Mean Systolic BP, mmHg (SD)		141.37 (19.88)	144.33 (21.81)	1.01 (1.006, 1.007)	1.001 (1.001, 1.002)
Mean Diastolic BP, mmHg (SD)		81.54 (11.35)	82.04 (12.79)	1.003 (1.002, 1.004)	-
Mean Glycated haemoglobin, % (SD)		8.02 (2.09)	7.85 (1.92)	0.96 (0.95, 0.96)	0.94 (0.93,0.95)
Mean serum creatinine, umol/L (SD)		91.74 (28.71)	96.06 (35.88)	1.01 (1.00, 1.01)	1.001 (1.000, 1.002)
Mean total cholesterol, mmol/L (SD)		5.26 (1.28)	5.24 (1.26)	0.98 (0.98, 0.99)	0.93 (0.92, 0.94)
Mean triglycerides, mmol/L (SD)		2.60 (2.53)	2.56 (2.50)	0.99 (0.98, 1.00)	-
Mean HDL cholesterol, mmol/L (SD)		1.24 (0.41)	1.25 (0.43)	1.08 (1.05, 1.10)	-

Table 5.1 continued…

Health Board Area				
Ayrshire & Arran	8,766 (46.44)	10,111 (53.56)	0.85 (0.82, 0.88)	0.89 (0.85, 0.93)
Borders	1,896 (35.19)	3,492 (64.81)	1.36 (1.28, 1.44)	1.41 (1.31, 1.53)
Dumfries & Galloway	3,190 (39.92)	4,801 (60.08)	1.11 (1.06, 1.16)	1.06 (0.99, 1.13)
Fife	7,255 (40.64)	10,598 (59.36)	1.08 (1.04 1.11)	1.05 (1.00, 1.10)
Forth Valley	5,745 (41.02)	8,259 (58.98)	1.06 (1.02, 1.10)	1.00 (0.95, 1.05)
Grampian	9,235 (40.94)	13,320 (59.06)	1.06 (1.03 1.10)	1.02 (0.97, 1.07)
Greater Glasgow	23,732 (42.44)	32,190 (57.56)	1 (Ref)	1 (Ref)
Highlands	4,321 (32.47)	8,987 (67.53)	1.53 (1.47, 1.60)	1.41 (1.32, 1.50)
Lanarkshire	12,103 (46.18)	14,108 (53.82)	0.86 (0.83, 0.88)	0.88 (0.84, 0.91)
Lothian	13,733 (41.94)	19,009 (58.06)	1.02 (0.99, 1.05)	0.96 (0.92, 0.99)
Orkney	308 (32.25)	647 (67.75)	1.55 (1.35, 1.76)	1.61 (1.33, 1.95)
Shetland	277 (28.27)	703 (71.73)	1.87 (1.63, 2.15)	2.07 (1.73, 2.47)
Tayside	9,934 (42.91)	13,217 (57.09)	0.98 (0.95, 1.01)	0.59 (0.57, 0.62)
Western Isles	483 (42.33)	658 (57.67)	1.00 (0.89, 1.13)	0.86 (0.73, 1.01)

† Calculated by averaging median patient BMI. SD: Standard Deviation, IQR: Interquartile Range

Missingness in BMI at diagnosis was strongly associated with year of diagnosis, with missingness being highest amongst patients who were diagnosed prior to 1995 (5.4% with data available if diagnosis before 1995 compared with 76.8% if diagnosis was after 2004). In the multivariate logistic regression model, earlier year of diagnosis remained independently associated with the probability of having missing BMI at diagnosis than later diagnoses (Before 1995 vs. After 2004 multivariate (mv) odds ratio (OR): 58.74 [95% CI: 55.20-62.50])

People who had died prior to the date of data extraction in January 2012 were considerably more likely to be missing a BMI at diagnosis than patients who were still alive at that time-point (deceased vs. alive: univariate (uni) OR 3.03 [95% CI: 2.62, 3.80]). This association reduced considerably after adjustment (deceased vs. alive: mv OR 2.22 [2.14, 2.30]) suggesting possible confounding by year of diagnosis or age. Men were slightly less likely to be missing a BMI at diagnosis than women (males vs. females: mv OR 0.93 [0.91, 0.95]).

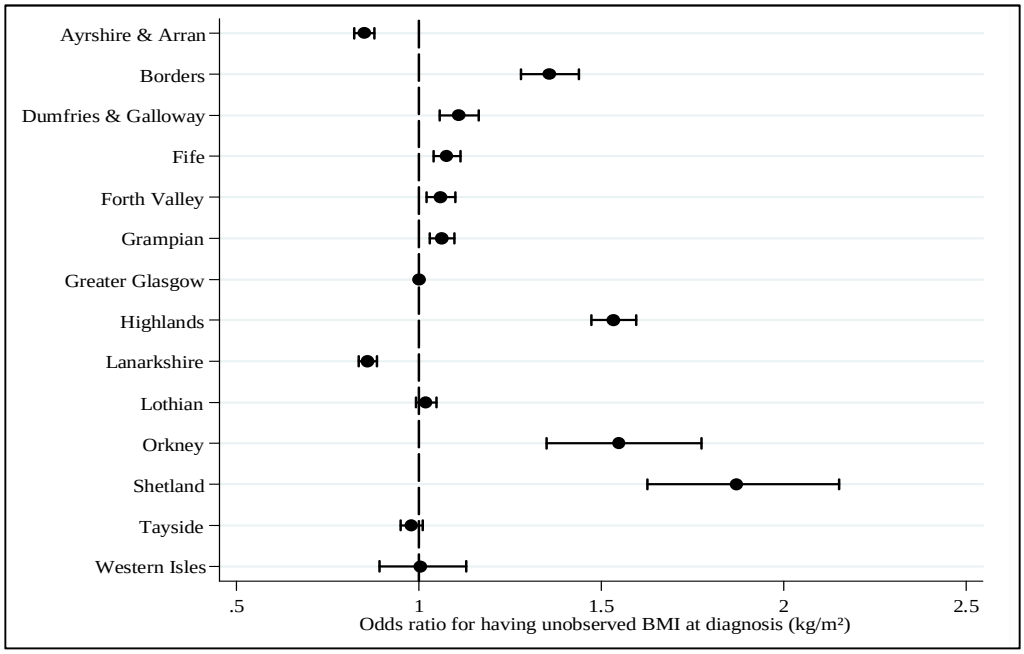

Figure 5.1: Unadjusted odds ratios (& 95% confidence intervals) of having missing BMI at diagnosis by HBA.

Missingness in BMI at diagnosis was related to the Health Board Area (HBA) in which the patient resided (Table 5.1 & Figure 5.1). With the exception of the Western Isles, patients attending healthcare services within the most sparsely populated HBAs were more likely to be missing a BMI at date of diagnosis measurement. Patients within the Highlands or the Shetland HBA, whose population density were 10 & 15 persons per square kilometre respectively, were considerably more likely to have missing data than attending patients within the Greater Glasgow HBA (1,052 persons per square kilometre) (Highlands vs. Greater Glasgow: mv OR 1.41 [95% CI: 1.32-1.50], Shetland vs. Greater Glasgow: mv OR 2.07 [1.73, 2.47]), though this was based on smaller numbers. In comparison, patients in Tayside were much less likely to have a missing

BMI at diagnosis than patients from Greater Glasgow (Tayside vs. Greater Glasgow: mv OR 0.59 [0.57, 0.62]).

Median ever-recorded patient BMIs were lower amongst patients with missing BMI at diagnosis than for patients without missing BMI (observed 31.31kg/m^2 vs. missing: 30.27kg/m^2). Upon categorisation of ever-recorded median BMI using the World Health Organisation's BMI classification system (BMI below 18.5kg/m^2 underweight, 18.5-24.99kg/m^2 normal weight, 25-29.99kg/m^2 overweight, 30-34.99kg/m^2 moderately obese, 35-39.99kg/m^2 severely obese, 40kg/m^2 or over very severely obese), the highest rates of missingness of BMI at diagnosis was observed in the underweight and normal weight categories (Figure 5.2). Patients with severe or very severe obesity were more likely to have observed recordings of BMI at diagnosis than patients who had a ever-recorded median BMI in the moderately obese category.

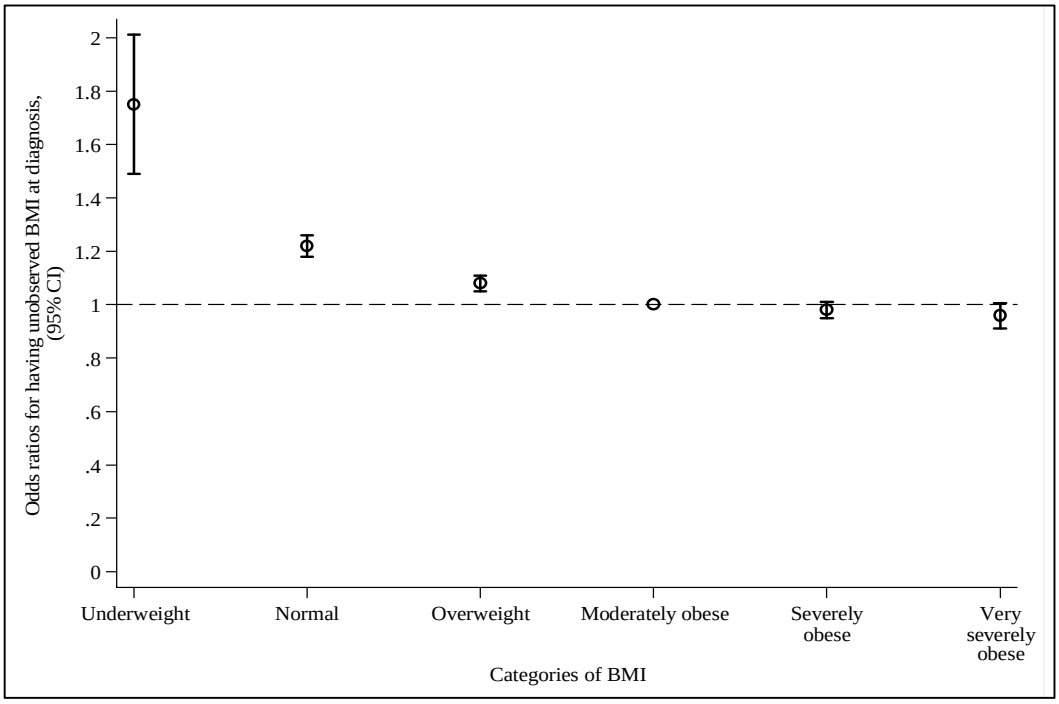

Figure 5.2: Adjusted ORs for having missing BMI values at date of diagnosis by categories of BMI group (from median ever-recorded BMI). Estimates adjusted for age, sex and smoking status

Though the median level of comorbidity was the same for patients with and without complete data, a greater proportion of patients with missing data had no other comorbid conditions defined in the Charlson Comorbidity Index. The proportion of missing data was 60% amongst patients without comorbid conditions compared to 54%

of patients with comorbidities. Ever smokers had lower odds of having a missing BMI measurement at diagnosis than never smokers (Ever vs. Never mv OR 0.91 [95% CI: 0.89, 0.93]). Lower glycated haemoglobin, lower total cholesterol and higher HDL-cholesterol were all associated with increased odds of missing BMI at diagnosis.

No evidence of multiplicative interactions was observed in the multivariable model.

5.4.4 Results II: Survival and missingness

Patients with missing data had a median follow-up time of 11.00 years compared with 6.72 years for patients without missing BMI data (Table 5.1). Patients without a date of diagnosis were not included as follow-up time could not be calculated (n=2,511).

A Kaplan-Meier plot showed some further interesting features of patient follow-up and survival (Figure 5.3). Firstly, while the maximum follow-up time amongst patients with missing data was just below 60 years, maximum follow-up time was less than 37 years amongst patients with an observed BMI. Nonetheless, median survival times were comparatively similar for patients with and without missing data with 17.72 years and 16.90 years, respectively.

During the course of follow-up the survival curves for the missingness groups cross over. In the early years of follow-up, patients without missing data have better survival than patients with missing BMI data. In the later years of follow-up, survival seems to be poorest amongst patients with observed data; however this apparent difference was based on small numbers.

The crossing-over of the curves in the Kaplan-Meier plot indicate a likely violation of the proportional hazards assumptions made by the Cox proportional hazards model. This notion correlates with the log-log plot, shown in Figure 5.4 whereby the curves do not run parallel to one another.

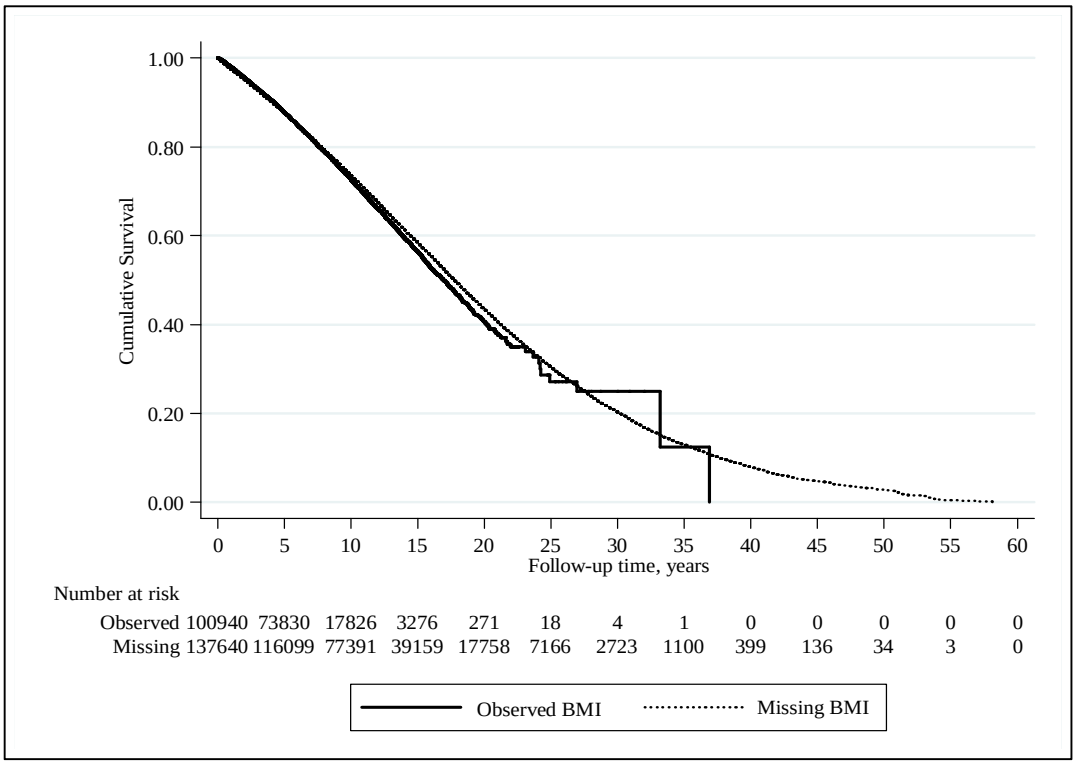

Figure 5.3: Kaplan Meier plot of the probability of survival by groups of missingness

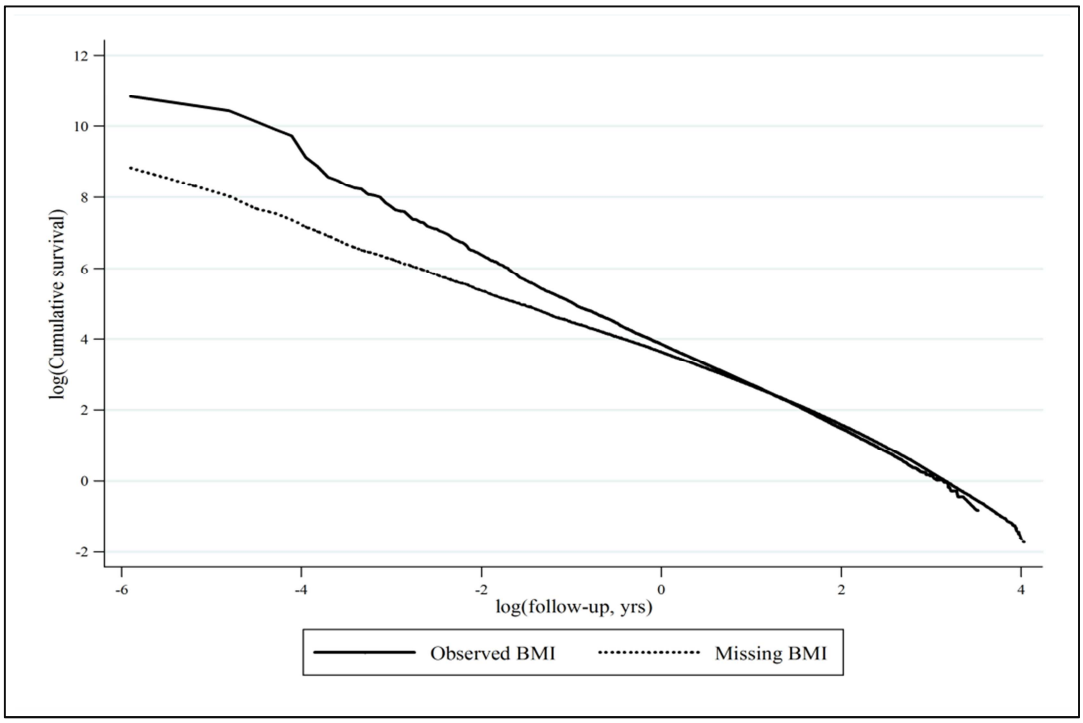

Figure 5.4: Log-log plot for testing the proportional hazards assumption

Table 5.2 presents the unadjusted and age and sex adjusted HRs for all-cause mortality by missingness, as well as crude and age-standardised mortality rates. Absolute and

relative measures for mortality contradict one another; relative measures indicate that patients with missing BMI had better survival (missing vs. observed mvHR: 0.95 [0.93. 0.97]), whilst absolute measures suggest patients with missing BMIs had worse survival compared with patients with observed BMI data (22.5 deaths per 1,000 person-years vs 33.6 deaths). These contradictory results remained after stratification by sex (See Appendix Table A5.1). These findings can be assumed to have arisen through the violation of the proportional hazards assumption. Accordingly, relative measures of survival calculated using the Cox proportional hazards model did not accurately describe the true survival experience of the groups.

Table 5.2: Relative and absolute measures of survival by groups of missingness

	Complete BMI	Missing BMI
Crude HR (95% CI)	1	0.97 (0.95, 0.98)
Age and sex adjusted HR (95% CI)	1	0.95 (0.93, 0.97)
Deaths	22,383	62,574
Person-years	716,608	1,619,412
Crude mortality, deaths per 1,000 person-years	31.2	38.6
Age-standardised mortality, per 1,000 person years (95% CI)	22.5 (22.2, 22.8)	33.7 (33.4, 33.9)

To investigate this problem further, time was divided into discrete time periods. The first five years in which the majority of patients were either censored or died were divided into yearly intervals whilst the remaining follow-up time was divided into 5-year intervals. The proportionality assumption was checked within each discrete time interval and deemed reasonable in all intervals.

Absolute and relative measures of survival for each time interval by missingness are presented in Table 5.3. Age-standardised mortality rates could not be calculated in the last two categories for patients with an observed BMI due to low numbers. Across each time-interval, absolute and relative measures of survival were in agreement. In the first year of follow-up, patients with missing data had worse survival than patients without, though this pattern was reversed from one year of follow-up onwards. In both groups of missingness, risk of mortality increased considerably as follow-up time lengthened.

During the exploration of this problem of contradictory findings, an irregularity in the dataset was observed that had not before been considered. Upon investigating mortality by missingness, it was observed that in the early years of the dataset there is artificially low mortality. Figure 5.5 presents this problem. Numbers of deaths gradually increased as calendar year increased; a pattern indicative of uneven retrospective look back. The majority of patients diagnosed with DM prior to the formation of SCI-DC only appeared on the register if they were still alive at the time of its formation. This feature was likely to introduce considerable survival bias to the study and therefore needed to be resolved (see Section 5.5).

Table 5.3: Absolute and relative measures of survival within defined time-intervals by missingness in BMI at diagnosis

		Intervals for follow-up time, years									
		0 to < 1	1 to < 2	2 to <3	3 to < 4	4 to <5	5 to< 10	10 to <15	15 to <20	20 to < 25	25 to <30
Observed	Deaths	2,086	2,441	2,503	2,520	2,566	7,756	2,054	426	30	1
	Person-years	98,729	96,381	93,915	90,883	80,177	208,672	40,849	6,587	382	30
	Crude mortality	21.1	25.3	26.7	27.7	32.0	37.1	50.3	64.7	78.5	33.3
	Age-standardised mortality	13.7 (13.0, 14.4)	**16.7 (16.0, 17.5)**	**18.1 (17.3, 18.9)**	**19.0 (18.2, 19.9)**	**22.5 (21.5, 23.4)**	**29.1 (28.3, 29.8)**	**49.1 (46.0, 52.1)**	**74.4 (71.6, 77.2)**	-	-
Missing	Deaths	3,588	3,254	3,350	3,401	3,550	17,021	12,983	8,118	4,163	1,866
	Person-years	133,491	130,173	126,884	123,351	117,384	483,513	278,840	134,724	57,022	21,857
	Crude mortality	26.9	25.0	26.4	27.6	30.2	35.2	46.6	60.3	73.0	85.4
	Age-standardised mortality	**16.5 (16.0, 17.1)**	16.2 (15.6, 16.8)	17.7 (17.1, 18.4)	18.2 (18.0, 19.4)	21.4 (20.8, 22.1)	29.5 (29.1, 30.0)	48.4 (47.1, 49.6)	76.3 (60.0, 92.5)	97.2 (91.7, 102.6)	123.4 (110.2, 136.6)
Crude HR, missing vs. observed (95% CI)		1.27 (1.20, 1.34)	0.99 (0.94, 1.04)	0.99 (0.94, 1.04)	0.99 (0.94, 1.07)	0.94 (0.90, 0.99)	0.92 (0.90, 0.95)	0.91 (0.86, 0.95)	0.90 (0.82, 1.00)	0.90 (0.63, 1.30)	1.61 (0.09, 4.34)
Age and sex adjusted HR, missing vs. observed (95% CI)		1.26 (1.20, 1.33)	0.97 (0.92, 1.02)	1.01 (0.96, 1.07)	1.01 (0.93, 1.09)	1.00 (0.95, 1.05)	0.99 (0.96, 1.01)	0.95 (0.91, 0.99)	0.96 (0.87, 1.06)	0.96 (0.67, 1.38)	1.77 (0.12, 5.46)

HR: Hazard ratio; CI: Confidence interval. Bold digits indicate worse age-standardised mortality rates than comparison group.

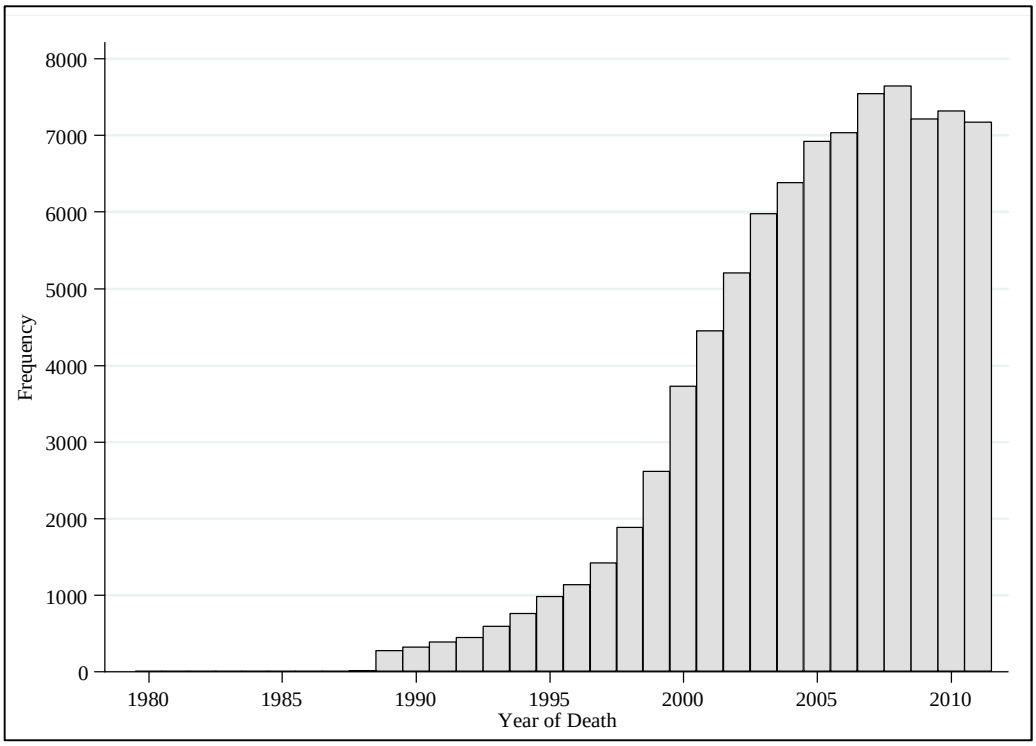

Figure 5.5: Frequency of deaths amongst patients with DM by calendar year

5.4.5 DISCUSSION OF RESULTS FROM FULL DATASET

This section was able to discern some interesting features of the completeness of the SCI-DC register. Generally it was clear that the dataset's completeness improved through time. The largest improvements occurred in BMI completeness after 2004 following the introduction of the QOF. Improvements such as these have been reported in a number of routine healthcare databases in the UK (Bhaskaran et al., 2013; Marston et al., 2014). For example, Bhaskaran et al., (2013) reported that between 2000 and 2004, 34% of patients on the Clinical Practice Research Datalink (CPRD) had a BMI measurement recorded in the previous three years and this increased to 51% between 2005 and 2011.

The purpose of this thesis was to investigate methods for dealing with missing data. Early years of the dataset were missing deceased patients and therefore it made sense to restrict our efforts to incident cases between 2004 and 2008, when death data was more complete. Though this process acted to reduce the sample size considerably, unbiased, but less precise results are of more value than more precise but biased estimates. Methods could be repeated on earlier data, but there was not sufficient time to do this in the current piece of work.

Aside from the elucidated problems in the dataset, interesting findings were observed in relation to the missingness mechanism that may be a useful guide in later analyses of missing data mechanisms. In the full cohort and indeed in the reduced sample, it is likely that the three missing data mechanisms coexist. That is, some of the missing data were MCAR, some were MAR and others were MNAR. Nonetheless, systematic differences between patients with and without a BMI at diagnosis were suggestive of a possible violation of the MCAR assumption.

The finding that missingness was highest amongst patients diagnosed before 1995 is unsurprising and probably reflects widespread improvements in risk factor recording through initiatives such as the QOF. Through the introduction of financial incentives, type 2 DM care and risk factor recording has become more standardised. As a result, BMI should now be recorded at least annually in patients with type 2 DM, regardless of its clinical relevance.

Patients residing in sparsely populated areas were also more likely to have missing data. This finding could be a reflection of the increased rurality of these areas and the subsequent issues surrounding patient access to and the availability of healthcare services. Meanwhile, completeness in the Tayside register was considerably better than elsewhere, possibly because SCI-DC was initially developed and implemented in this health board area.

A number of crude indicators of relative healthiness, such as total cholesterol, HDL-cholesterol and glycated haemoglobin suggested that healthier patients were more likely to have missing BMI at diagnosis measurements than unhealthy patients. Lower total cholesterol values and higher HDL-cholesterol values typically reflect lower cardiovascular disease risk and these were found to be associated with an increased risk of missingness of BMI. Similarly, lower glycated haemoglobin levels, which are indicative of better DM control, were associated with lower completeness of BMI recording around date of diagnosis of DM. This finding coincides with what was discovered in discussions with clinicians. Patients with poorly controlled glycated haemoglobin are prioritised and recalled regularly, thus providing more opportunities to record patient measurements. To further support this notion, patients with missing data

were also more likely to have no comorbidities and be never smokers. This finding could be due to lower attendance rates of comparatively healthier patients.

Nonetheless, analyses investigating the survival experience of the groups indicated that patients with observed data had better short-term (≤1 year) survival than patients with missing data. The recording of patient BMI upon the diagnosis of DM may not be relevant or indeed possible in dying patients. Long-term survival was better amongst patients with missing data compared with patients with observed data, reflecting a change in the factor driving missingness in BMI.

Median ever-recorded BMI appeared to be a strong predictor of missingness. Patients whose median-ever recorded BMI fell within the underweight or normal categories were found to be more likely to be without a BMI at diagnosis measurement. This may relate to a lower inclination of clinicians to measure patients whose weight was deemed to be 'healthy'. Meanwhile obese and severely obese patients whose weight was of clinical concern may have been monitored closely. Failure to include further BMI recordings in imputation models could therefore lead to data being MNAR.

Some clinicians indicated that missing BMI data around diagnosis date could reflect incorrect recording of diagnosis date. This is an interesting notion but is difficult to validate. In order to ascertain DM type an algorithm is used which involves validating the date of diagnosis. This process uses timing of first DM mention, timing of treatment and age at diagnosis. Using these different sources of information should ensure a reliable "best-guess" of DM diagnosis date. Although this algorithm has not been validated, it is likely that the performance of it has improved with the introduction of QOF and the subsequent increased recording of DM-related indicators. Where there are errors in the diagnosis date, there is no reason to believe these are differential and biased in a particular direction. The use of a BMI value recorded within a one-year time interval (10 months before and 2 months after) to reflect BMI at diagnosis should provide some protection should diagnosis date be incorrect and the true diagnosis date lies within this one-year time-interval.

The mechanisms by which missing BMI data came about in this large cohort are potentially numerous. This section has been able to elucidate a number of potential

reasons by which missing data came about, though many more possibly exist. These possibilities will need to be considered when applying missing data methods.

5.5 2004 - 2008 COHORT

5.5.1 INTRODUCTION

As discussed in the previous section, despite the register's introduction in 2001, universal coverage of the SCI-DC register was not achieved until 2004. Consequently, completeness of the register before this time was poor. Whilst patients who were alive at the time of the formation of SCI-DC would be included in the register, patients who died before this time may or may not have been included in the register, depending on the availability of retrospective data. Accordingly, this produces artificially low mortality in the early years of the register. It was therefore necessary to divide the cohort into sections to deal with differential survivor bias. This section reports on the group diagnosed between 2004 and 2008.

Limiting the cohort to incident cases within this time period was likely to have a profound impact on both the amount of missing data and the likely mechanisms by which missing data came about. For example, 2004 was when the QOF was introduced and therefore risk factor recording was intensified and became more universal. Failure to record BMI should therefore have been primarily due to chance alone after 2004, meaning that data are MCAR. Prior to QOF, risk factor recording was at least partly clinically driven.

5.5.2 METHODS

The methods used to identify potential reasons for missingness were the same as those applied in section 5.4.2, but applied to the smaller sample of people who were diagnosed with type 2 DM more recently.

5.5.3 RESULTS

Of the 241,041 patients in the cohort, 77,948 patients were diagnosed with type 2 DM between 2004 and 2008. The proportion of missingness in the BMI at diagnosis variable was considerably lower (26.9%) in this subsample than was observed for the whole sample (58.1%).

Table 5.4 presents the characteristics of patients with and without missing data for this subsample. Odds ratios from univariate and multivariate models for the risk of missingness in BMI are presented. Trends in missingness were similar to those observed in Section 5.4.3.

People with type 2 DM who were deceased at the end of follow-up were considerably more likely to be missing BMI at diagnosis than patients who were still alive in 2008 (deceased vs. alive: mv OR 1.65 [1.52, 1.78]). Older age at diagnosis was protective against missingness until the over 80 years of age group. In this age-group, the odds of missingness were 9% higher than that observed in the under 50 year old age group (Over 80s vs under 50s: mv OR 1.09 [1.01, 1.19]). Male sex and ever-smoking were protective against missingness. The presence of comorbid conditions also conferred an increased risk of missingness in BMI. Of the clinical characteristics, only elevated glycated haemoglobin levels were moderately associated with a lowered risk of missingness in BMI at diagnosis (mv OR: 0.96 [0.95, 0.97]).

Table 5.4: Factors influencing missingness in BMI at diagnosis with type 2 DM between 2004 and 2008 in Scotland.

Variable		Observed BMI (n=56,990, 73.11%)	Missing BMI (n=20,958, 26.89%)	Univariate models OR (95% CI)	Multivariate model OR (95% CI)
Age at diagnosis, years (%)	< 50	9,807 (75.50)	3,182 (24.50)	1 (Ref)	1 (Ref)
	50 – 59	13,859 (76.48)	4,263 (23.52)	0.95 (0.90, 1.00)	0.90 (0.85, 0.96)
	60 – 69	16,566 (77.59)	4,785 (22.41)	0.89 (0.85, 0.94)	0.79 (0.74, 0.84)
	70 – 79	12,701 (76.34)	3,937 (23.66)	0.96 (0.91, 0.73)	0.77 (0.72, 0.82)
	≥80	4,027 (63.63)	2,302 (36.37)	1.76 (1.62, 1.88)	1.09 (1.01, 1.19)
Sex (%)	Female	31,388 (74.01)	11,022 (25.99)	1 (Ref)	1 (Ref)
	Male	25,592 (72.13)	9,890 (27.87)	0.91 (0.88, 0.94)	0.93 (0.91, 0.95)
Vital status (%)	Alive	48,869 (76.80)	16,895 (23.20)	1 (Ref)	1 (Ref)
	Deceased	8,121 (56.72)	6,196 (43.28)	2.53 (2.43, 2.62)	1.65 (1.52, 1.78)
SIMD (%)	1 (Most deprived)	13,672 (74.16)	4,763 (25.84)	1.02 (0.96, 1.07)	1.02 (0.95, 1.09)
	2	12,910 (73.26)	4,911 (26.74)	1.06 (1.01, 1.12)	1.06 (1.00, 1.14)
	3	11,636 (72.77)	4,354 (27.23)	1.09 (1.03, 1.15)	1.11 (1.04, 1.19)
	4	10,011 (71.24)	4,042 (28.76)	1.18 (1.11, 1.25)	1.16 (1.08, 1.24)
	5 (Least deprived)	8,421 (74.47)	2,887 (25.53)	1 (Ref)	1 (Ref)
Year of diagnosis (%)	2004	13,648 (71.85)	5,346 (28.15)	1 (Ref)	1 (Ref)
	2005	13,364 (76.36)	4,137 (23.64)	0.79 (0.75, 0.83)	0.81 (0.77, 0.85)
	2006	13,444 (77.84)	3,828 (22.16)	0.73 (0.69, 0.76)	0.76 (0.73, 0.80)
	2007	12,945 (76.77)	3,917 (23.23)	0.77 (0.74, 0.81)	0.83 (0.79, 0.87)
	2008	3,569 (74.15)	1,244 (25.85)	0.89 (0.83, 0.96)	0.97 (0.90, 1.04)
Mean ever-recorded BMI, kg/m^2 (SD) †		31.61 (6.21)	30.99 (6.48)	0.984 (0.981, 0.987)	0.986 (0.985, 0.991)

Table 5.4 continued…

Variable		Observed BMI (n=56,990, 73.11%)	Missing BMI (n=20,958, 26.89%)	Univariate models OR (95% CI)	Multivariate model OR (95% CI)
Mean predicted BMI at diagnosis, kg/m^2 (SD)		31.67 (6.27)	31.18 (12.73)	0.991 (0.988, 0.994)	-
Median Follow-up time, years (IQR)		5.58 (4.49, 6.74)	5.45 (4.17, 6.86)	0.94 (0.93, 0.95)	1.02 (0.96,1.08)
Measurements recorded around date of diagnosis					
Presence of comorbid conditions (derived from CCI) (%)	None	38,273 (76.56)	11,717 (23.44)	1 (Ref)	1 (Ref)
	One or more	18,717 (66.95)	9,241 (33.05)	1.38 (1.36, 1.40)	0.87 (0.82, 0.91)
Smoking status (%)[a]	Never smoker	23,313 (76.40)	7,200 (23.60)	1 (Ref)	1 (Ref)
	Ever smoker	33,418 (78.76)	9,012 (21.24)	0.85 (0.81, 0.89)	0.91 (0.89, 0.93)
Mean Systolic BP, mmHg (SD)		139.29 (18.65)	139.48 (19.68)	1.000 (0.999, 1.001)	-
Mean Diastolic BP, mmHg (SD)		80.59 (10.94)	80.13 (11.14)	0.996 (0.995, 0.998)	-
Mean Glycated haemoglobin, % (SD)		8.04 (2.13)	7.89 (2.03)	0.96 (0.95, 0.97)	0.965 (0.956,0.975)
Mean serum creatinine, umol/L (SD)		90.87 (26.99)	91.95 (30.34)	1.001 (1.000, 1.002)	-
Mean total cholesterol, mmol/L (SD)		5.16 (1.30)	5.12 (1.31)	0.98 (0.96, 0.99)	0.98 (0.97, 1.00)
Mean triglycerides, mmol/L (SD)		2.53 (2.61)	2.54 (2.54)	0.99 (0.98, 1.00)	-
Mean HDL cholesterol, mmol/L (SD)		1.24 (0.40)	1.26 (0.42)	1.12 (1.07, 1.18)	-
Health Board Area					
Ayrshire & Arran		4,950 (70.40)	2,081 (29.60)	1.33 (1.24, 1.42)	1.07 (0.99, 1.16)
Borders		1,212 (75.14)	401 (24.86)	1.03 (0.91, 1.16)	1.00 (0.86, 1.56)
Dumfries & Galloway		1,709 (74.40)	588 (25.60)	1.08 (0.98, 1.20)	1.21 (1.08, 1.36)
Fife		3,949 (72.82)	1,474 (27.18)	1.17 (1.09, 1.25)	1.18 (1.08, 1.36)
Forth Valley		3,196 (74.48)	1,095 (25.52)	1.07 (0.99, 1.17)	1.02 (0.93, 1.12)
Grampian		5,902 (73.60)	2,127 (26.40)	1.13 (1.06, 1.20)	0.94 (0.86, 1.02)

Table 5.4 continued…

Variable	Observed BMI (n=56,990, 73.11%)	Missing BMI (n=20,958, 26.89%)	Univariate models OR (95% CI)	Multivariate model OR (95% CI)
Greater Glasgow	13,755 (75.15)	4,548 (24.85)	1 (Ref)	1 (Ref)
Highlands	2,747 (65.03)	1,477 (34.97)	1.72 (1.57, 1.87)	1.44 (1.29, 1.60)
Lanarkshire	6,998 (73.05)	2,582 (26.95)	1.16 (1.09, 1.23)	0.97 (0.90, 1.04)
Lothian	7,409 (72.76)	2,774 (27.24)	1.18 (1.11, 1.25)	1.21 (1.13, 1.30)
Orkney	205 (61.38)	129 (38.62)	2.04 (1.63, 2.56)	1.69 (1.26, 2.25)
Shetland	212 (71.14)	86 (28.86)	1.26 (0.98, 1.63)	1.30 (0.97, 1.75)
Tayside	4,491 (75.14)	1,486 (24.86)	1.04 (0.97, 1.12)	1.06 (0.97, 1.75)
Western Isles	255 (72.03)	99 (27.97)	1.24 (0.98, 1.56)	1.12 (0.84, 1.49)

† Calculated by averaging median patient BMI. SD: Standard Deviation, IQR: Interquartile Range

According to both absolute and relative measures, all-cause mortality risk was significantly greater amongst patients with missing data than amongst patients without missing data (Complete vs. Missing age and sex adjusted HR: 1.47 [1.42, 1.53]) (Table 5.5). The proportional hazards assumption was checked using log-log plots and deemed reasonable.

Table 5.5: Relative and absolute measures of survival by groups of missingness in cohort of patients diagnosed with DM between 2004 and 2008.

	Complete BMI	Missing BMI
Crude HR (95% CI)	1	1.61 (1.55, 1.67)
Age and sex adjusted HR (95% CI)	1	1.47 (1.42, 1.53)
Deaths	8,068	4,072
Person-years	307,816	95,739
Crude mortality, deaths per 1,000 person-years	26.2	42.5
Age-standardised mortality, per 1,000 person years (95% CI)	17.8 (17.3, 18.2)	26.3 (25.3, 27.1)

5.5.4 DISCUSSION OF RESULTS FROM 2004-2008 SUBSAMPLE

In this subsample of patients diagnosed between 2004 and 2008, there were still important differences between patients with and without missing data, despite the introduction of the QOF. Patient survival was perhaps the biggest predictor of missingness, though the presence of comorbid conditions and being an ever-smoker were protective against missingness.

Patients with missing BMI data had considerably worse survival than patients without missing BMI at diagnosis. This finding has large implications for missing data analyses since failure to include patients with missing data would markedly over estimate overall survival in the population.

The reason behind poorer survival in patients with missing data is unknown; it may be that it is not a priority to record BMI in individuals who are approaching end of life. Alternatively, there may be fewer opportunities in which to record this measurement before death.

The observation that raised glycated haemoglobin levels conferred a reduced risk of having missing BMI data probably reflected the need for patients with poorly controlled DM to attend their healthcare provider on a regular basis. Nonetheless, the observed effect was quite small in this sample.

HBA also remained a considerable predictor of missingness in BMI at diagnosis suggesting that improvements in recording practices were not universal.

To summarise, within this subsample the MAR assumption seems relatively plausible for at least some of the data. The use of missing data methods which make the MAR assumption should therefore yield valid results. However, it would be beneficial to explore the sensitivity of these methods to departures from this assumption.

5.6 PREDICTORS OF BMI AT DATE OF DIAGNOSIS

5.6.1 INTRODUCTION

Many missing data approaches require robust predictive models to identify 'best-guesses' of the missing values. For example, to produce unbiased estimates using multiple imputaiton, it is necessary to include variables which reliably predict the missing values in the imputation model. This section outlines the process of identifying variables that are predictive of BMI at diagnosis. These predictors were later included in models for handling partially observed data.

5.6.2 METHODS

For continuous variables, patterns of BMI at diagnosis were examined visually using scatterplots with added regression lines. BMI at diagnosis was grouped into 5kg/m^2 categories to enable the observation of non-linear trends. Patients with a BMI between 15kg/m^2 (the lower plausibility bound) and 25kg/m^2 were grouped together as were those with a BMI above 45kg/m^2 to 75kg/m^2 (upper plausibility bound). Means and standard deviations by BMI categories were calculated. Significance tests for mean differences were not carried out due to large numbers.

Linear regression was used to assess the association between BMI at diagnosis (continuous variable) and available demographic and clinical characteristics. Crude and adjusted regression coefficients are presented. Diastolic BP and serum creatinine were log-transformed due to evidence of non-linear relationships with BMI around diagnosis of DM. Triglycerides and HDL-cholesterol were not included in multivariate models due to high proportions of incompleteness. Since initial analyses were attempting to assess the performance of missing data methods in the absence of repeat recordings of BMI, neither within-patient median nor regressed BMI were included in the final model.

5.6.3 RESULTS

In the study population of patients diagnosed with DM between 2004 and 2008, there were 56,990 patients who had an available BMI at diagnosis. Figures 5.6, 5.7 and 5.8 present scatterplots of BMI at diagnosis by available continuous variables. Table 5.6

presents binary and categorical characteristics by categories of BMI at diagnosis. Univariate and multivariate regression coefficients for the association between BMI at diagnosis and available variables are also presented in Table 5.6.

According to the scatterplots and summary statistics, few variables were strongly associated with BMI at diagnosis. Nonetheless, patient BMI appeared to decrease with increasing age at diagnosis and there were a lower proportion of deceased patients in the highest BMI categories. Men tended to have a lower BMI at diagnosis than women (men vs.women: mv. regression coefficient -2.01 [95% CI: -2.11, -1.91]) and patients in the most deprived quintile were more likely to have higher BMIs at diagnosis than patients in the least deprived quintile (Q1 vs. Q5: mv. Regression coefficient 1.62 [95% CI: 1.44, 1.79]) (Table 5.7).

The number of patients with comorbid conditions decreased with increasing BMI and there were a smaller proportion of ever smokers in the higher BMI categories. Patients residing in certain HBAs were also more likely to fall into the highest BMI categories (Table 5.8). For example, patients in Shetland (33.0 kg/m^2) and Orkney (32.11kg/m^2) had much higher BMIs than those from the referent category, Greater Glasgow (31.7kg/m^2).

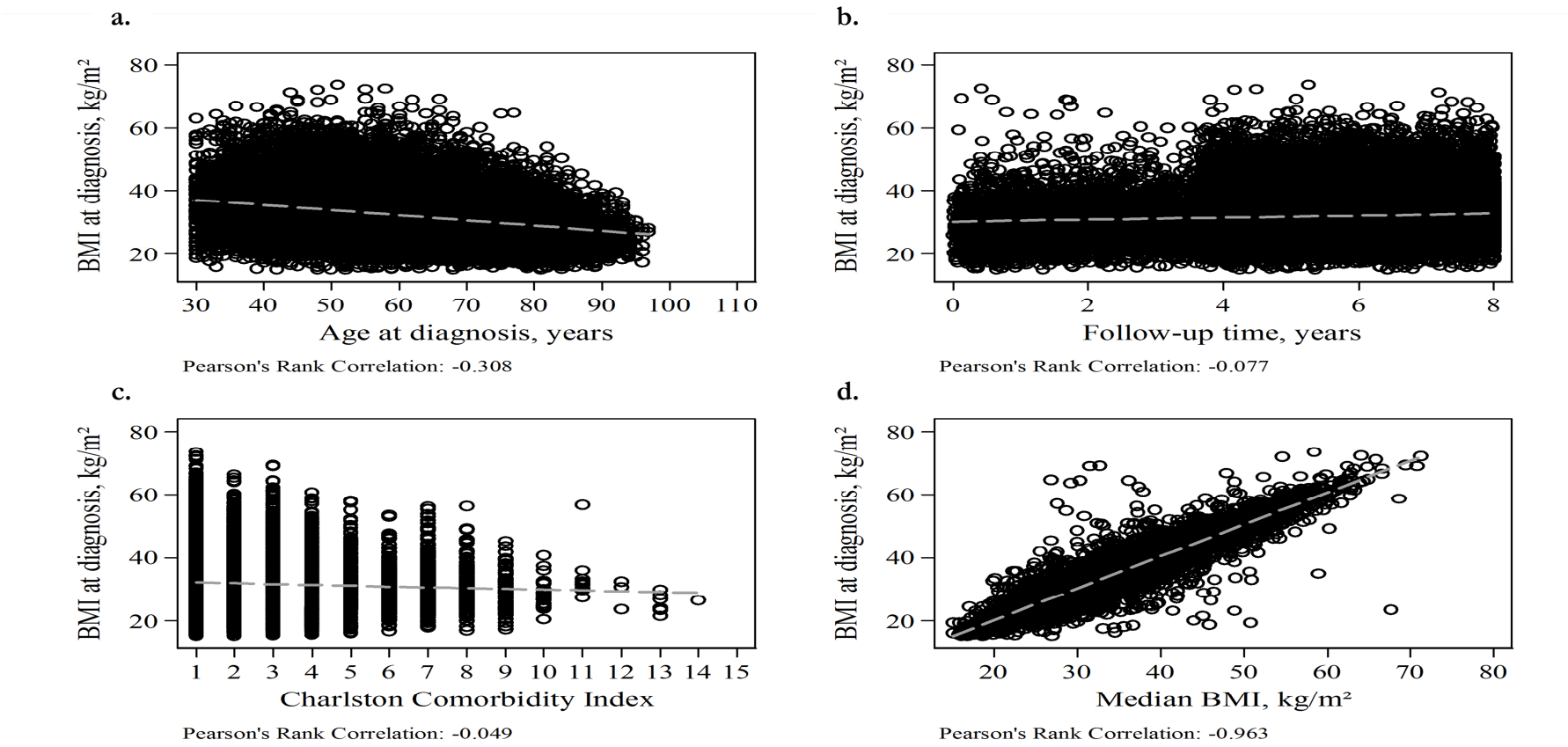

Figure 5.6. Scatterplots of BMI at diagnosis with DM against (a) Age at diagnosis, (b) Follow-up time, (c) Charlson Comorbidity Index, (d) Individual-median BMI Regression lines and Pearson's Rank correlations are also presented.

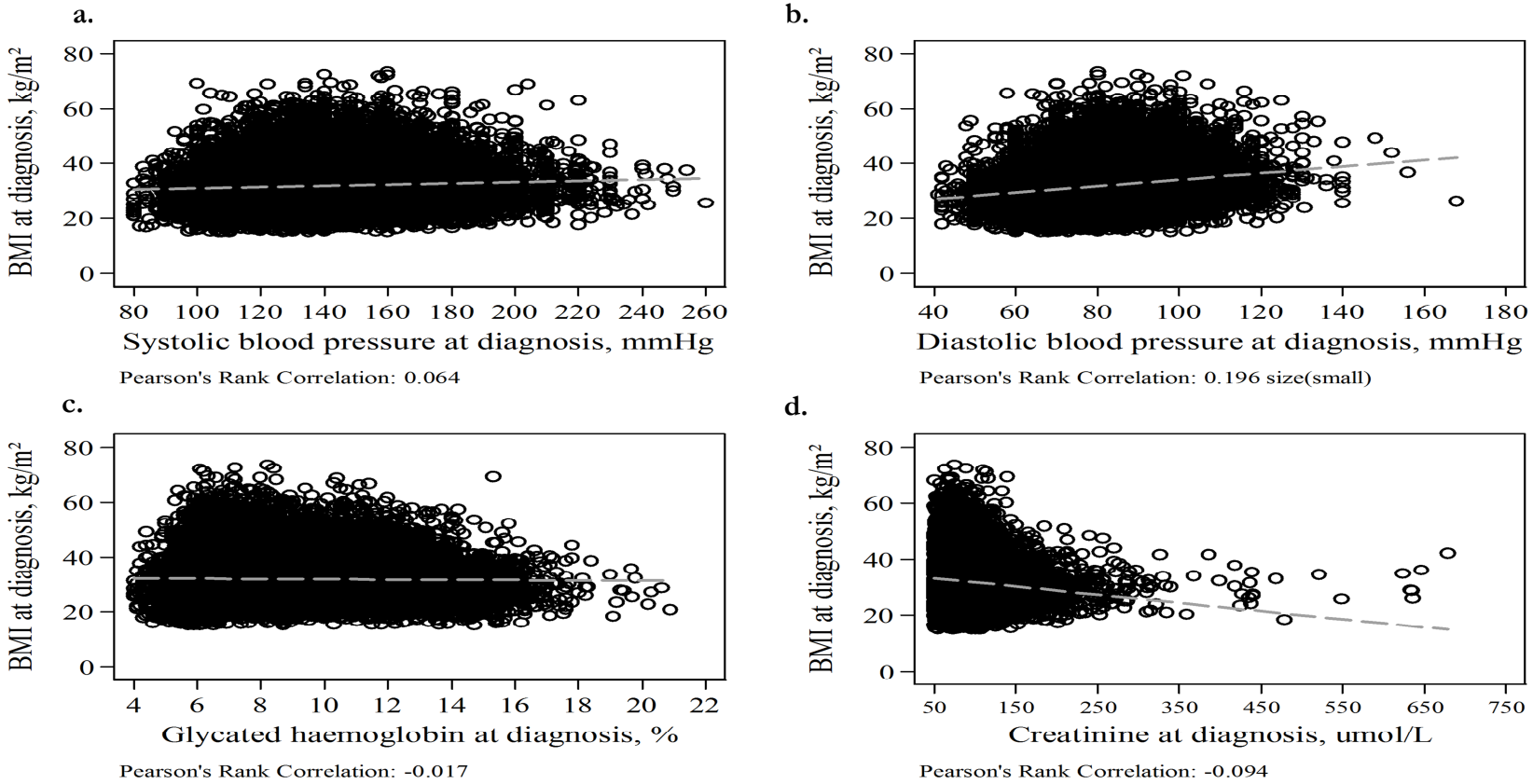

Figure 5.7: Scatterplots of BMI at diagnosis with DM with (a) Systolic BP (b) Diastolic BP (c) Glycated Haemoglobin (d) Creatinine

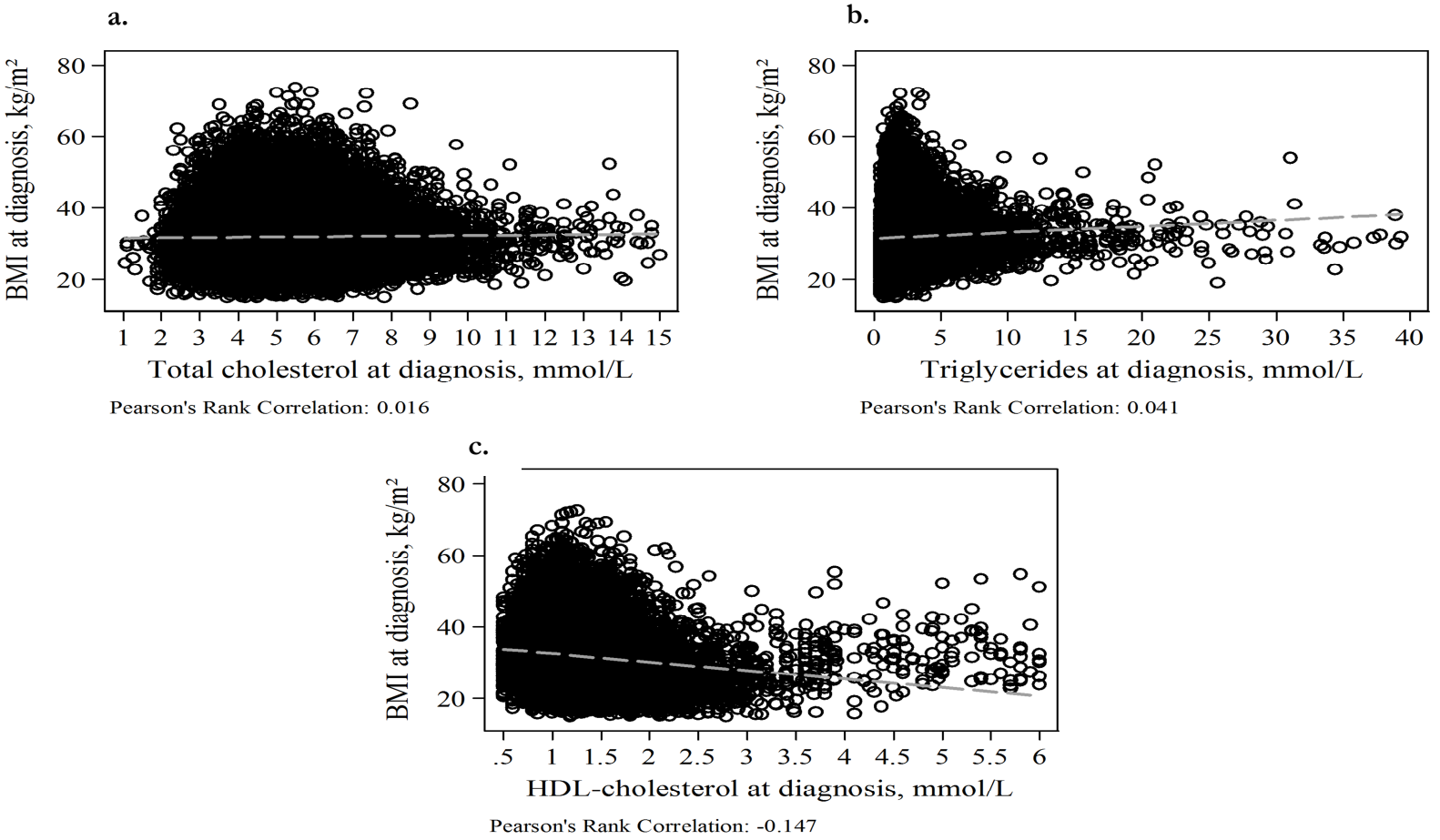

Figure 5.8: Scatterplot of BMI at diagnosis with DM against (a) Total cholesterol (b) Triglycerides (c) HDL-cholesterol

Table 5.6: Clinical and demographic characteristics of study population by categories of BMI at diagnosis with DM

	BMI category, kg/m²							Crude Linear regression coefficients (95% CI)	Multivariate linear regression coefficients (95% CI)
	<25	25 to < 30	30 to < 35	35 to < 40	40 to < 45	≥ 45	ALL		
N (%)	6,561 (11.51)	17,674 (31.01)	17,539 (30.78)	9,060 (15.90)	3,841 (6.74)	2,315 (4.06)	56,990 (100.00)	-	-
Mean age at diagnosis (SD)	67.10 (12.85)	64.78 (11.86)	61.58 (11.54)	58.33 (11.48)	55.35 (11.20)	52.72 (10.36)	61.91 (12.31)	-0.16 (-0.17, -0.15)	-0.163 (-0.167, -0.158)
Male (%)	3,409 (51.97)	10,862 (61.47)	10,356 (59.06)	4,455 (49.17)	1,560 (40.64)	746 (32.24)	31,388 (55.09)	-1.49 (-1.60, -1.38) vs. female	- 2.01 (- 2.11, - 1.91) vs. female
Deceased at study-end (%)	1,834 (27.95)	2,793 (15.80)	2,094 (11.94)	864 (9.54)	325 (8.46)	211 (9.11)	8,121 (14.25)	-2.51 (-2.66, -2.36) vs. alive	-0.89 (-1.12, -0.66) vs. alive
Median year of diagnosis (IQR)	2005	2006	2006	2006	2006	2006	2006	-	-
With comorbid conditions (%)	2,363 (36.02)	5,925 (33.58)	5,804 (33.09)	2,844 (31.39)	1,108 (28.85)	663 (28.64)	18,717 (32.84)	-0.54 (-0.66, - 0.43) vs. no additional	-0.37 (-0.48, -0.26) vs. no additional
Ever smokers (%)	3,939 (60.37)	10,543 (59.92)	10,589 (60.63)	5,140 (56.97)	2,106 (55.12)	1,103 (47.83)	33,418 (58.91)	-0.65 (-0.76, -0.55) vs. Never smoker	-0.37 (-0.47, -0.26) vs. Never smoker
Median follow-up, years(IQR)	5.32 (4.08, 6.57)	5.60 (4.49, 6.75)	5.58 (4.53, 6.75)	5.62 (4.56, 6.78)	5.69 (4.63, 6.80)	5.66 (4.62, 6.75)	5.55 (4.41, 6.77)	-0.05 (-0.07, -0.04)	-
Mean Systolic BP, mmHg (SD)[a]	137 (20.07)	139 (18.85)	140 (18.21)	140 (18.29)	141 (18.06)	142 (18.60)	139 (18.65)	0.02 (0.01, 0.03)	-
Mean Diastolic BP, mmHg (SD) [a]	77 (10.77)	79 (10.54)	81 (10.67)	83 (10.93)	84 (10.82)	85 (11.51)	81 (10.94)	0.12 (0.11, 0.12)	0.085 (0.080, 0.089)
Mean Glycated haemoglobin, % (SD) [a]	8.19 (2.52)	8.08 (2.20)	7.96 (2.04)	7.99 (1.98)	8.00 (1.88)	8.12 (1.92)	8.04 (2.13)	-0.05 (-0.08, -0.03)	-0.17 (-0.20, -0.15)

	BMI category, kg/m²							Crude Linear	Multivariate linear
Mean serum creatinine, umol/L (SD) [a]	92.27 (30.41)	93.00 (29.04)	91.77 (26.52)	88.31 (21.95)	85.85 (25.98)	82.14 (18.69)	90.87 (26.99)	-0.023 (-0.025, -0.209)	-
Mean Total cholesterol, mmol/L (SD) [a]	5.12 (1.35)	5.15 (1.33)	5.17 (1.31)	5.18 (1.24)	5.21 (1.19)	5.18 (1.17)	5.16 (1.30)	0.08 (0.04, 0.12)	- 0.30 (-0.33, -0.27)
Mean Triglycerides, mmol/L (SD) [a]	2.04 (1.74)	2.48 (2.73)	2.71 (2.90)	2.67 (2.59)	2.59 (2.30)	2.36 (1.73)	2.53 (2.61)	0.10 (0.07, 0.13)	-
Mean HDL-cholesterol, mmol/L (SD) [a]	1.42 (0.49)	1.26 (0.39)	1.21 (0.38)	1.19 (0.37)	1.17 (0.34)	1.19 (0.38)	1.19 (0.38)	-2.39 (-2.53, -2.24)	-

[a] Recorded around date of diagnosis. IQR: Interquartile Range. SD: Standard deviation

Table 5.7: Mean BMI at diagnosis with DM by deprivation status (from SIMD)

SIMD quintile (%)	Mean BMI at diagnosis, kg/m² (SD)	Crude Linear regression coefficients (95% CI)	Multivariate linear regression coefficients (95% CI) [a]
1 (most deprived)	32.56 (6.98)	1.70 (1.52, 1.87)	1.62 (1.44, 1.79)
2	32.15 (6.62)	1.28 (1.11, 1.46)	1.22 (1.05, 1.39)
3	32.03 (6.44)	1.16 (0.98, 1.34)	1.11 (0.94, 1.29)
4	31.51 (6.21)	0.65 (0.46, 0.84)	0.65 (0.47, 0.83)
5 (least deprived)	30.86 (5.89)	Ref	Ref

[a] Adjusted for age, sex, vital status, comorbidity status, smoking status, diastolic BP, glycated haemoglobin and total cholesterol and HBA

Table 5.8: Mean BMI at diagnosis with DM by HBA

Health Board Area	Mean BMI at diagnosis, kg/m² (SD)	Crude Linear regression coefficients (95% CI)	Multivariate linear regression coefficients (95% CI) [a]
Ayrshire & Arran	31.73 (6.55)	0.05 (-0.17, 0.27)	0.32 (0.11, 0.53)
Borders	31.62 (6.33)	-0.09 (-0.48, 0.30)	0.90 (0.52, 1.29)
Dumfries & Galloway	32.35 (6.56)	0.65 (0.31, 0.98)	0.83 (0.61, 1.25)
Fife	32.31 (6.55)	0.59 (0.35, 0.83)	0.72 (0.48, 0.94)
Forth Valley	32.21 (6.64)	0.52 (0.26, 0.78)	0.81 (0.57, 1.06)
Grampian	31.95 (6.53)	0.25 (0.04, 0.46)	0.84 (0.63, 1.04)
Greater Glasgow	31.69 (6.49)	Ref	Ref
Highlands	31.55 (6.34)	0.00 (-0.32, 0.32)	0.48 (0.17, 0.79)
Lanarkshire	32.00 (6.55)	0.30 (0.10, 0.50)	0.31 (0.12, 0.50)
Lothian	32.04 (6.67)	0.33 (0.14, 0.53)	0.63 (0.45, 0.82)
Orkney	32.11 (6.34)	0.29 (-0.63, 1.21)	0.91 (0.00, 1.82)
Shetland	33.00 (6.17)	1.20 (0.31, 2.09)	1.47 (0.63, 2.32)
Tayside	31.95 (6.38)	0.25 (0.02, 0.48)	0.74 (0.52, 0.96)
Western Isles	32.16 (5.60)	0.50 (-0.31, 1.31)	1.06 (0.29, 1.83)

[a] Adjusted for age, sex, vital status, deprivation status, comorbidity status, smoking status, diastolic BP, glycated haemoglobin and total cholesterol

Both systolic and diastolic BP increased with increasing BMI group, though at the highest DBP values, this association disappears (Figure 5.6b). Similarly Figure 5.6d indicates that BMI tends to decrease with increasing blood creatinine level up until a creatinine level of 350umol/L. After this level there does not appear to be a relationship between BMI at diagnosis and creatinine, though this was based on small numbers.

A U-shaped relationship was observed between glycated haemoglobin (Table 5.6) and BMI. HDL-cholesterol increased with increasing BMI, though this association was not apparent in patients with a HDL-cholesterol above 4mmol/L. This relationship was possibly confounded by sex. Women have on average, higher HDL-cholesterol values than men (1.33mmol/L vs. 1.17mmol/L) and are over-represented in the very high BMI groups. A similar pattern was observed for total cholesterol. No relationship was observed between BMI and follow-up time nor year of diagnosis. Mean BMIs increased slightly by year of diagnosis (Figure 5.9)

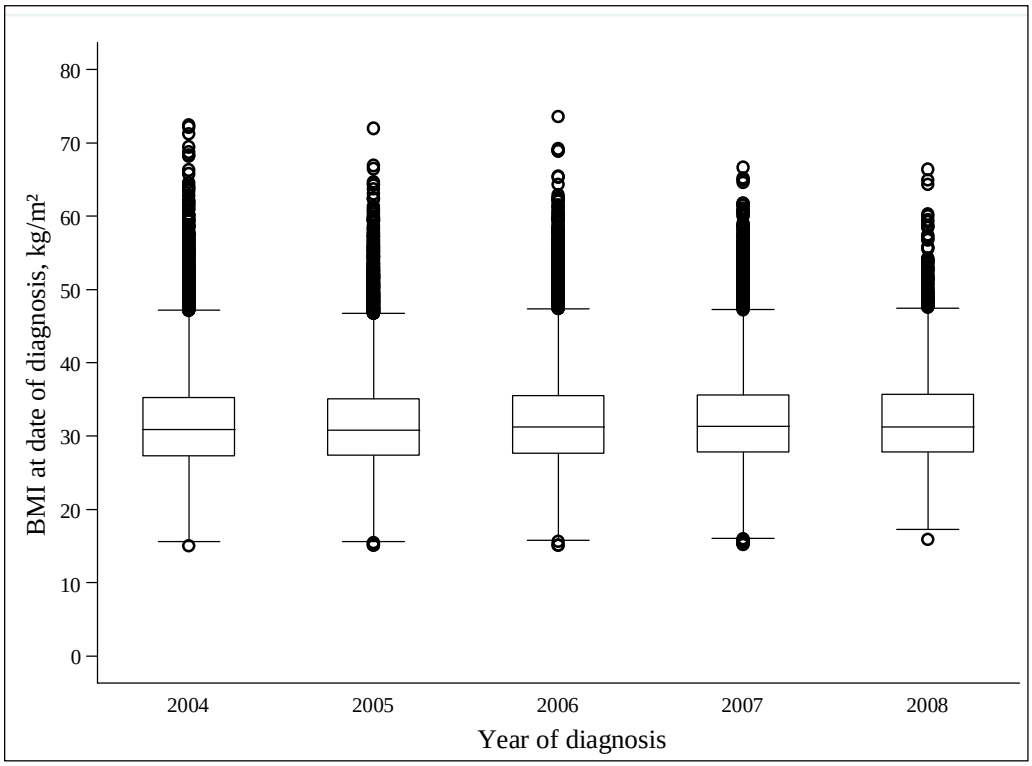

Figure 5.9: Box and whiskers plot of BMI at diagnosis by year of diagnosis

The R-squared statistic from the multivariate model was 0.1578, thus indicating that less than 16% of the variance in BMI at date of diagnosis was explained by available characteristics.

5.6.4 SUMMARY

This section identified several important predictors of BMI at diagnosis, including age at diagnosis, sex, health board area, year of diagnosis and deprivation status. These findings correspond well with what has been observed elsewhere.

The tendency for people to lose weight when they become elderly is well-documented and reflects, amongst other things, losses in muscle mass and changes in metabolic rates. The finding that older patients had lower BMIs may have been slightly overstated through a cohort effect. Younger people tend to be more overweight than currently older people were at the same age (Roubenoff, 1999). Since these data are not describing weight changes amongst individuals caution needs to be used when interpreting these findings. Nonetheless, this relationship may also go some way in explaining the finding that deceased patients tended to have lower BMIs than still-alive patients. Patients approaching the end of their life are more likely to be older, thus explaining their lower BMIs at diagnosis. However, this relationship remained after adjustment for age at diagnosis and other variables indicating that vital status may be independently associated with BMI. Indeed, this finding signals the potential for reverse causality in which the disease process that increases the risk of death also causes weight loss. The finding of a higher prevalence of comorbid conditions in the lowest BMI categories may also reflect this process.

Patients in the most deprived category were more likely to be heavier than patients in the least deprived category. This finding corresponds well with previously conducted studies on both general and DM populations in developed countries (McLaren, 2007; Wild et al., 2008). Indeed, this association is considered to be an important factor in determining the known association between deprivation and type 2 DM prevalence. It might be expected that this relationship may also account for the finding that patients residing in certain HBAs were more likely to be heavier than patients residing elsewhere. However, places such as Shetland and the Orkney have some of the lowest levels of deprivation yet had the highest mean BMIs at diagnosis. Indeed, no patients residing in Shetland or Orkney lived in an area found to be in the most deprived category so factors other than deprivation are clearly important.

There were a lower proportion of ever smokers in the highest BMI category, thus reflecting the fact that smokers tend to weigh less than non-smokers. This association may have been diluted due to our classification system which does not take into account that amongst ever-smokers some patients are former smokers. Many previous studies have identified a tendency for patients to gain weight following smoking cessation and

therefore the classification system used here may have diluted the effect of smoking on BMI at diagnosis (Cao et al., 2012).

There was a slight indication that newly-diagnosed patients were heavier, or perhaps shorter in later years than they were in previous years. This, in part, may reflect secular trends in BMI distribution. In samples of the general population in Scotland participating in Health Surveys, mean BMI has increased from 26.8kg/m² in 1998 to 27.9kg/m² in 2008 (Lean et al., 2012). In addition, there is now a growing trend for much younger patients to be diagnosed with type 2 DM and these patients tend to have much higher BMIs than people whose DM is diagnosed at a later age (Logue et al., 2011).

Many of the observed trends in biochemical characteristics must be interpreted carefully. The measurements used were those recorded closest to diagnosis with DM and therefore may have been recorded a considerable time after or before diagnosis. These measurements may not accurately reflect the true reading at diagnosis. It is difficult to interpret what impact this may have on observed associations. However, the association between increasing BMI and increasing BP is well known (Kannel et al., 1967).

Unfortunately, common to most epidemiological studies, the prediction of BMI was poor when cross-sectional data was used (van Buuren et al., 1999; Hardt et al., 2012). This feature of the dataset may be detrimental to later imputation models which will attempt to use these variables to provide a best-guess of the missing values. Where fewer predictors of BMI are available, the fraction of missing information increases. This describes the proportion of missing data values and the correlation between the missing data and the observed data. As the fraction of missing information increases, the uncertainty in the imputed values increases, thus resulting in longer imputation model convergence times and wider confidence intervals. Furthermore, when the fraction of missing information is large, problems associated with misspecified imputation models are amplified (White et al., 2011). To overcome this problem and generate robust parameter estimates, Graham et al. (2007) recommend increasing the number of imputations.

5.7 CONCLUSIONS

The aim of this section was firstly to assess the plausibility of the separate missing data mechanisms and secondly to identify variables that were predictive of the BMI variable. This process was complicated by the finding that mortality was erroneously low in the full cohort and therefore the cohort was reduced to incident DM cases between 2004 and 2008.

In this newly formed cohort, systematic differences between persons with and without missing BMI data were observed. In consequence, the MCAR assumption made by numerous approaches for handling missing data is unrealistic and any resulting estimates from these approaches are likely to be biased. More specifically, this section has indicated that the probability of BMI at diagnosis measurements being observed is dependent on survival suggesting data are MAR. Persons with shorter-survival times were more likely to have missing data than those with longer survival times. This factor must therefore be included in imputation models when correcting for missingness. The probability of missingness in BMI was also found to be conditional on year of diagnosis and health board area. These variables are unlikely to be important in the model of substantive interest but will need to be included in the missing data model to improve imputations.

Despite these findings, it remains possible that a considerable amount of data were MNAR. Alternative methods will therefore need to account for the possibility that data are MNAR. To achieve this, ad-hoc sensitivity analyses will be displayed to investigate the robustness of estimates to departures from the MAR assumptions.

Table 5.9 summarises the observed relationships between available variables and BMI at diagnosis and also missingness in BMI. It also presents the variables that will be used in the missing data model. Selection of these variables was based on their ability to predict BMI and/or missingness in BMI. Despite its identification as a variable capable of potentially describing the variance in BMI at diagnosis, HDL-cholesterol was not included on the basis of having too many missing values.

Table 5.9: Summary of relationship of baseline characteristics with BMI at diagnosis and missingness in BMI at date of diagnosis

Variable	Risk of missingness in BMI	Strength of relationship	BMI at diagnosis	Strength of relationship	Included in missing data model
Age at diagnosis	↓ with ↑ age	Quite strong	↓ with ↑ age	Quite strong	✓
Sex	↑ for women	Quite strong	↑ for women	Quite strong	✓
Vital Status	↑ with death	Strong	↓ with death	Strong	✓
Deprivation status	None	-	↑ with ↑ deprivation	Strong	✓
Year of diagnosis	↓ with later year of diagnosis	Strong	↑ with later year of diagnosis	Weak	✓
Follow-up time	↑ with ↑ follow-up	Strong	None	-	✓
Smoking status	↑ in never-smokers	Quite strong	↑ in never-smokers	Quite strong	✓
Charlson comorbidity index	None	-	↓with additional comorbidity	Quite strong	✓
Systolic BP	None	-	↑ with ↑ SBP	Quite strong	✓
Diastolic BP	None	-	↑ with ↑ DBP	Weak	×
Glycated haemoglobin	↓ with ↑ HbA_{1c}	Quite strong	None		✓
Creatinine	None	-	↓ with ↑ creatinine	Weak	×
Total cholesterol	↓ with ↑ total cholesterol	Quite strong	↑ with ↑ total cholesterol (weak)	Weak	✓
Mean triglycerides	None	-	↑ with ↑ triglycerides	Weak	×
HDL-cholesterol	↑ with ↑ HDL-cholesterol	Quite strong	↓ with ↑ HDL-cholesterol	Strong	×
HBA	↑ in sparsely populated areas	Strong	↑ in certain HBAs	Strong	✓

6. DEALING WITH MISSING DATA IN SINGLE BMI MEASUREMENTS

6.1 INTRODUCTION

This chapter describes the association between BMI at diagnosis and all-cause mortality in patients diagnosed with type 2 DM between 2004 and 2008. Several approaches for handling unobserved BMI data were applied. Analyses were restricted to patients with missing data in a single variable (BMI) to illustrate available methodology when there is a univariate pattern of missingness. At this stage, repeated BMI variables are not used to estimate plausible values of BMI at diagnosis. More complicated missing data scenarios are explored in later chapters.

Information regarding the likely missing data mechanism is used to explain observed estimates and to inform a discussion of the problems associated with selecting an inappropriate method for handling missing data.

For each missing data mechanism, at least one approach for handling missing data is presented. The chosen methods were selected on the basis of their availability in popular statistical packages and their accessibility to non-statisticians. The following methods were chosen:

1. For MCAR data - CCA and population mean imputation;

2. For MAR data - the single imputation approach, stochastic imputation and two varieties of multiple imputation, namely multiple imputation using chained equations (MICE) and predictive mean matching (PMM).

3. For MNAR data – a pattern mixture sensitivity analysis were applied to investigate the robustness of estimates to departures from the MAR assumption.

6.2 METHODS

This methods section is divided into three subsections; the first describes the initial exclusion criteria, the second outlines the methods used for handling missing data and the third describes the final analysis model.

6.2.1 Exclusion Criteria

Initial analyses were performed upon a sub-sample of patients with a univariate pattern of missingness. It was therefore necessary to restrict the sample to patients with complete data for all variables (other than BMI) to be included in both the missing data model and the final model of interest. Variables to be included in the final model were chosen on the basis of being known to influence and confound the relationship between BMI and all-cause mortality. These variables were age, sex, Charlson comorbidity index, smoking status and deprivation status. The variables to be included in the missing data model were selected from the exploratory analyses detailed in Chapter 5. These variables were included in the missing data models for two reasons. First, the inclusion of variables which are associated with the probability that BMI data are missing will reduce bias and will make the MAR assumption more plausible. Second, the inclusion of variables which are at least moderately correlated with BMI at date of diagnosis will help to reduce the uncertainty of the imputations. The selected variables to be included in the missing data model were age at diagnosis, sex, vital status, year of diagnosis, follow-up time, Charlson comorbidity index, HBA, systolic BP, glycated haemoglobin and total cholesterol. Patients with missing data in any of the above variables were therefore removed.

Patients who had died within two years of follow-up were also excluded from analyses to limit the risk of biases through reverse causation. Inclusion of these individuals could bias associations in the lowest BMI categories as their low BMIs may be in response to chronic disease. Risk of mortality in these categories would therefore be exaggerated.

The final, full case-flow is presented in section 6.3.1.

6.2.2 Methods for handling missing data

Six methods for handling missing BMI data were applied. This section describes the application of each approach.

6.2.2.1 MCAR: CCA

This method analysed data on patients with observed BMI only.

6.2.2.2 MCAR: POPULATION MEAN ANALYSIS

Missing values were replaced with the calculated mean BMI of the non-missing observations at diagnosis (32kg/m²).

6.2.2.3 MAR: STOCHASTIC IMPUTATION

Stochastic imputation is a single imputation approach which involves the replacement of missing values with a predicted value from a regression model with an added error term. The addition of this random error term gives this single imputation approach its name.

The *uvis* command in the ICE suite (imputation by chained equations) for STATA was used to predict missing values for BMI at date of diagnosis (Royston, 2004). This user-written program is typically used for multiple imputation models but in its application here was used to create a single complete dataset only. Briefly, ICE is an iterative procedure which uses a series of conditional distributions to impute missing data. For each variable with missing data, a regression model is formulated conditional upon other selected variables. Therefore binary variables with missing data are modelled using logistic regression, continuous variables are modelled using linear regression and so on. Random draws are then taken from the resulting posterior distribution of the missing data to form the final imputations. Where multiple variables have missing data, this procedure is iterated a number of times (the default is 10 in ICE) to obtain final estimates of the missing values. In this analysis only the BMI at diagnosis variable required imputing and therefore only one regression model was used and no iterations were required. Only one complete dataset was produced using this approach.

The imputation model comprised of the following variables: age at diagnosis, sex, year of diagnosis, deprivation status, smoking status, Charlson comorbidity index, diastolic BP, total cholesterol and HBA. To avoid biasing relationships towards the null, it is crucial to include all variables whose associations will be assessed in the later analysis model. It was therefore necessary to include the outcome variables, survival time and vital status in the imputation model to help predict the missing covariate.

Previous studies have indicated a non-linear relationship between BMI and survival in both the general population and populations of DM patients. The presence of a non-linear relationship between the covariate to be imputed (BMI) and other variables could lead to the biasing of relationships towards the null. Typically, non-linear relationships

are resolved using polynomials, splines or the categorisation of the continuous variable. In accordance with previous studies which categorised BMI to overcome non-linearity issues, our final model did the same. To ensure the imputation model and the final model were compatible, BMI was also categorised in the imputation model. Though the categorisation of this quantitative exposure is inefficient, there remains no general consensus on which approach is best for handling non-linearity problems in imputation models (See section 2.3.1.2 for more details). Available methods are currently under-developed and largely untested on real data. Categorisation of BMI is therefore a straightforward technique which will also enable us to present easily-interpretable absolute and relative estimates by BMI category. In addition, it will reduce the risk of obtaining implausible values for BMI (below 15kg/m^2 or above 75kg/m^2).

BMI was grouped into 5kg/m^2 categories between 20kg/m^2 and 40kg/m^2. Those patients whose BMI fell outside this range were included in the lowest and highest BMI categories.

6.2.2.4 MAR: MULTIPLE IMPUTATION USING CHAINED EQUATIONS

uvis for STATA was used to create 25 multiply imputed data sets. Despite early research indicating 5-10 imputed datasets were sufficient, recent studies have indicated that as a rule of thumb, the number of imputed datasets should be at least as large as the proportion of missing data (White et al., 2011).

The multiple imputation using chained equations approach was chosen since it does not assume multivariate normality and was therefore more able to efficiently handle the incomplete ordinal BMI variable (This was further examined in Chapter 8).

Initial estimates were generated using an imputation model comprising of all variables included in the previous stochastic imputation model.

Following the creation of 25 complete data sets, ICE's *mim* command was used to analyse each dataset separately and to combine parameter estimates to form a single pooled estimate.

6.2.2.5 MAR: MULTIPLE IMPUTATION USING PREDICTIVE MEAN MATCHING

The *ice* command in STATA with the additional *ppm* option was used to create 25 multiply imputed datasets. PMM involves imputing an observed value which is amongst the closest values to the predicted value from the respective regression model. In *ice*, the default number of neighbours from which imputed values are randomly chosen from is 3. PMM is capable of maintaining the observed relationships in variables included in the imputation model and has therefore been proposed as a valuable means of overcoming problems surrounding non-linear relationships and non-normality.

The imputation model was specified in the same way as in the above multiple imputation approaches.

6.2.2.6 MNAR: SENSITIVITY ANALYSIS USING A PATTERN-MIXTURE BASED APPROACH

A pattern mixture sensitivity analysis was used to assess the robustness of estimates obtained from multiple imputation, when the MAR assumption is violated. This MNAR approach was proposed by Donald Rubin (1987) at a time when software options for MNAR data were severely limited.

It was hypothesised that 50% of patients with missing data were without a BMI measurement because their BMI was not considered clinically relevant and therefore was not recorded. This would occur if the practitioner did not believe the patients BMI was of concern, i.e. it lay in the 'normal' BMI range. Accordingly, a random selection of 50% of patients with imputed data had their BMI category reduced by one unit in each of the 25 datasets formed using MICE. Therefore if a patient was coded in the 35 to <40kg/m² category they would be placed into the 30 to <35kg/m². Where patients were already in the lowest BMI category they remained in this category. At the time of writing, there was no accessible guidance for conducting more advanced approaches to exploring the sensitivity of MAR assumptions.

6.2.3 ANALYSIS MODEL

Associations between BMI at date of diagnosis and all-cause mortality were investigated using Cox proportional hazards regression. BMI at diagnosis, defined as a BMI measurement recorded 10 months prior to or 2 months after date of diagnosis was the

predictor of interest. Given the findings of previous studies which have convincingly demonstrated that mortality has a J- or U-shaped pattern with BMI, it was deemed inappropriate to summarise the effect of BMI using a single hazard ratio. For this reason, BMI was grouped into 5kg/m² categories between 20kg/m² and 40kg/m². Those patients whose BMI fell outside this range were included in the lowest and highest BMI categories as appropriate.

Hazard ratios (HR) were estimated relative to a reference category of 25 to< 30kg/m^2 due to containing a large number of people and having the lowest perceived risk of mortality (Logue et al., 2013). The timescale was calculated from the time of diagnosis until death or the end of follow-up (31/12/2012), whichever came first. Models were adjusted for age at diagnosis, sex, smoking status at diagnosis, deprivation status and Charlson comorbidity index. Previous studies have indicated that these variables are potential confounders of the relationship between BMI and mortality. The inclusion of these variables was therefore important to ensure the true relationship of interest was preserved. Violations to the proportional hazards assumption were investigated graphically using Kaplan Meier plots and log-log survival plots. Evidence of effect modification by sex was also investigated by stratifying analyses.

Age-standardised mortality rates for each category of BMI were calculated using the European standard population so that both absolute and relative measures could be used to compare the effect of missing data methods on mortality estimates.

6.3 RESULTS

6.3.1 EXCLUSION CRITERIA

The initial data cleaning procedure reduced the full cohort from 274,363 to 77,948 patients. The exclusion criteria for these analyses further reduced the sample size to 66,372 patients. Figure 6.1 presents the full case-flow schedule.

The main reason for exclusion of patients was the presence of incomplete data for variables included in the missing data or final analysis models (n=8,487). Proportions of missing data (Table 6.1) and the top ten missing data patterns (Table 6.2) revealed that

the foremost reason for exclusion amongst the 8,487 patients was the unavailability of smoking status, glycated haemoglobin or total cholesterol values. There were 12 patients who did not have data for any variable of interest (not shown in Table 6.2). A further 3,089 patients were excluded due to having died within two years of the date of diagnosis of DM. Amongst the final sample, 13,775 (20%) patients were missing a BMI record at diagnosis. There were a total of 7,592 deaths occurring more than two years after diagnosis of DM in the sample between 2004 and 2012.

Table 6.1. Proportions of missing data for variables to be included in missing data models (n=77,948)

Variable	Observed (%)	Missing (%)
Age at diagnosis	75,429 (96.77)	2,519 (3.23)
Sex	77,892 (99.93)	56 (0.07)
Vital status	77,948 (100.00)	0 (0.00)
Deprivation status	77,407 (99.31)	541 (0.69)
Year of diagnosis	75,442 (96.79)	2,506 (3.21)
Follow-up time	75,472 (96.82)	2,476 (3.18)
Smoking status	72,943 (93.58)	5,005 (6.42)
Charlson comorbidity index	75,442 (96.79)	2,506 (3.21)
Systolic BP	73,509 (94.31)	4,439 (5.69)
Glycated haemoglobin	71,533 (91.77)	6,415 (8.23)
Total cholesterol	72,838 (93.44)	5,110 (6.56)
HBA	77,937 (99.99)	11 (0.01)

Table 6.2. Top 10 patterns of data completeness in variables to be included in imputation models and analysis models (n=77,948)

Pattern of missingness												No. missing variables	Frequency (%)
Age at diagnosis	Sex	Vital Status	Deprivation status	Date of diagnosis	Follow-up time	Smoking status	Charlson Comorbidity Score	Systolic BP	Glycated Haemoglobin	Cholesterol Status	HBA		
✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	0	69,461 (89.12)
.	✓	✓	✓	.	.	.	.	.	.	.	✓	8	2,348 (3.01)
✓	✓	✓	✓	✓	✓	✓	✓	✓	.	✓	✓	1	1,748 (2.24)
✓	✓	✓	✓	✓	✓	.	✓	.	.	.	✓	4	1,348 (1.73)
✓	✓	✓	✓	✓	✓	.	✓	✓	✓	✓	✓	1	642 (0.82)
✓	✓	✓	✓	✓	✓	✓	✓	✓	.	.	✓	2	552 (0.71)
✓	✓	✓	.	✓	✓	✓	✓	✓	✓	✓	✓	1	430 (0.55)
✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	.	✓	1	419 (0.54)
.	✓	✓	✓	.	✓	.	.	.	.	.	✓	7	212 (0.27)
✓	✓	✓	✓	✓	✓	.	✓	.	✓	✓	✓	2	186 (0.24)

✓ = observed, . = missing

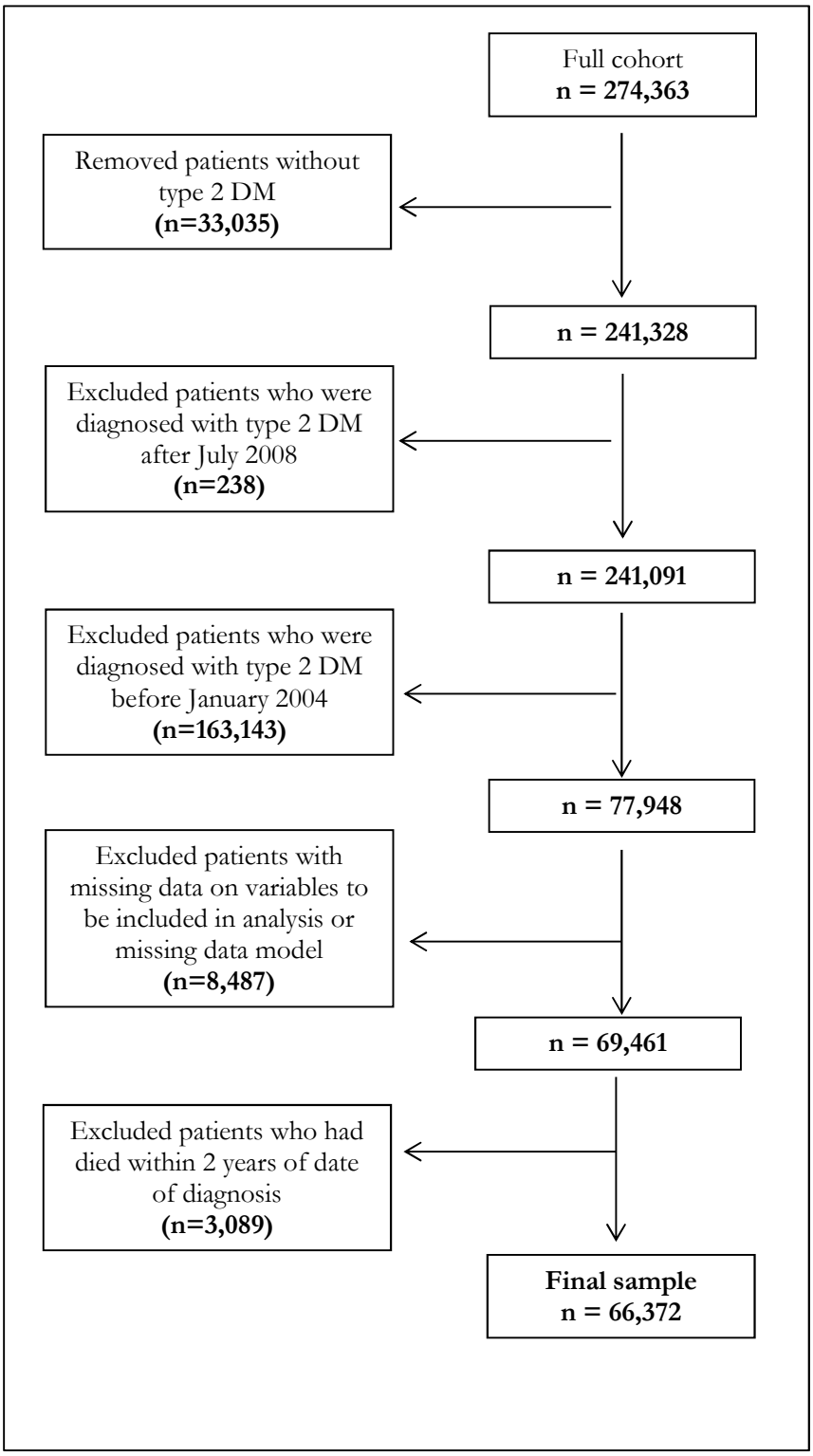

Figure 6.1: Full case-flow of excluded patients

6.3.2 MCAR: COMPLETE CASE ANALYSIS (CCA)

Analyses were based on 52,597 eligible patients (67% of patients diagnosed with type 2 DM between 2004 and 2008) having excluded patients with missing BMI data (n=13,775). The final sample included 28,930 men and 23,667 women, among whom a total of 5,491 deaths from all causes occurred. The median follow up time was 5.71 years, with a maximum follow-up time of 7.99 years (2004-2012).

Baseline characteristics by BMI category for the study sample are presented in Table 6.3. Characteristics by category of BMI were similar to those observed in the original whole sample. Age at diagnosis and the proportion of male patients decreased with increasing BMI. The proportion of patients alive at end of follow-up increased with increasing BMI, though this may be a result of a younger age at diagnosis amongst patients in the highest BMI categories. The proportion of patients in the most deprived category increased as did the median number of days of follow-up with increasing BMI. The prevalence of smoking also decreased with increasing BMI. As expected, elevated total cholesterol and BP levels were associated with higher BMI values.

Table 6.3. Baseline characteristics of patients included in final CCA by BMI category (n=52,597)

		BMI category, kg/m²					
		<25	25 to < 30	30 to < 35	35 to < 40	≥ 40	Total
N (%)		5,653 (10.75)	16,288 (30.97)	16,368 (31.12)	8,503 (16.17)	5,785 (11.00)	52,597 (100.00)
Mean age at diagnosis of diabetes, years (SD)		66.23 (12.69)	64.42 (11.73)	61.32 (11.43)	58.08 (11.37)	54.28 (10.87)	61.51 (12.13)
% Male		52.22	61.38	59.00	49.08	37.18	55.00
% Alive at end of follow-up		80.49	88.22	90.96	92.80	93.45	89.56
Median no. of years follow-up (IQR)		5.65 (4.53, 6.74)	5.75 (4.70, 6.84)	5.70 (4.68, 6.81)	5.74 (4.69, 6.83)	5.77 (4.75, 6.84)	5.72 (4.68, 6.83)
SIMD quintile (column %)	1 (Most deprived)	23.03	21.83	23.46	26.07	29.99	24.05
	2	21.76	22.27	22.32	23.99	24.32	22.74
	3	19.07	19.93	21.16	21.24	20.31	20.48
	4	18.29	18.56	18.46	16.21	15.23	17.75
	5 (Least deprived)	17.85	17.41	14.60	12.49	10.15	14.99
Median Charlson Comorbidity Score (IQR)		1 (1,2)	1 (1,2)	1 (1,2)	1 (1,2)	1 (1,2)	1 (1,2)
% Never smoker		40.60	40.32	39.64	43.30	47.68	41.43
Mean Systolic BP, mmHg (SD)[a]		137 (20.08)	139 (18.69)	140 (18.23)	141 (18.29)	141 (17.28)	139 (18.62)
Mean Diastolic BP, mmHg (SD) [a]		77 (10.69)	79 (10.49)	81 (10.65)	83 (10.90)	84 (10.80)	81 (10.88)
Mean glycated haemoglobin, % (SD) [a]		8.24 (2.55)	8.08 (2.20)	7.96 (2.03)	7.99 (1.98)	8.05 (1.90)	8.04 (2.12)
Mean serum creatinine, umol/L (SD) [a]		90.91 (28.65)	92.63 (28.01)	91.49 (26.06)	88.07 (21.74)	84.24 (23.36)	90.44 (26.19)
Mean Total cholesterol, mmol/L (SD)[a]		5.16 (1.35)	5.16 (1.32)	5.17 (1.31)	5.19 (1.24)	5.20 (1.19)	5.18 (1.29)
Mean Triglycerides, mmol/L [a]		2.04 (1.72)	2.50 (2.78)	2.71 (2.90)	2.68 (2.63)	2.50 (2.12)	2.54 (2.64)
Mean HDL-cholesterol, mmol/L (SD) [a]		1.42 (0.49)	1.26 (0.39)	1.21 (0.39)	1.18 (0.37)	1.17 (0.34)	1.18 (0.35)

[a] Closest recorded measurement to date of diagnosis, SD: Standard Deviation, IQR: Interquartile Range

HR estimates described a U-shaped relationship between BMI at diagnosis and all-cause mortality, after adjustment for age at diagnosis, sex, smoking status, deprivation status and Charlson comorbidity index (Table 6.4 & Figure 6.2). Patients with a BMI within the desirable weight category (< 25kg/m^2), as defined by the World Health Organisation

cut-off points, were found to be at an elevated risk of mortality compared to patients in the 25 to <30kg/m^2 group (HR: 1.51 [1.40, 1.63]). Compared with a BMI of 25 to < 30kg/m^2, risk of mortality was also higher amongst patients with a BMI of 40kg/m^2 or above. Standard errors were largest amongst the highest and lowest BMI categories reflecting smaller numbers in these groups. Obese patients (30 to <35kg/m^2) had a slightly lower relative risk of mortality than overweight patients (25 to <35kg/m^2) (HR: 0.98 [0.92, 1.06]). Crude mortality rates indicated that those patients in the highest BMI groups had the lowest rates of death of all BMI categories (11.6 deaths per 1,000 person-years), a trend that was reversed when results were age-standardised. Even after age-standardisation, mortality rates in the lowest BMI category remained highest (19.1 deaths per 1,000 person-years) compared with mid-range BMI categories (around 11 per 1,000 person years).

Table 6.4: Estimates of relative hazard of all-cause mortality and age-standardised mortality by categories of BMI using CCA.

	BMI category (kg/m^2)				
	< 25	25 to < 30 (REF)	30 to < 35	35 to < 40	≥ 40
N (%)	5,653 (10.75)	16,288 (30.97)	16,368 (31.12)	8,503 (16.17)	5,785 (11.00)
Multivariable adjusted HR [ab]	1.51 (1.40, 1.63)	1.00	0.98 (0.92, 1.06)	1.04 (0.94, 1.14)	1.37 (1.22, 1.53)
Standard Error	0.057	-	0.035	0.049	0.080
No. deaths	1,101	1,918	1,478	610	374
Person years	31,218	92,509	92,471	47,925	32,461
Crude mortality [c]	35.3	20.7	16.0	12.7	11.6
Age-standardised mortality [b c]	19.1 (17.3, 20.9)	11.6 (11.0, 12.3)	11.2 (10.5, 11.9)	11.4 (10.3, 12.5)	14.3 (11.5, 17.0)

[a] Adjusted for age, sex, smoking status, deprivation status and Charlson comorbidity index, [b] with 95% CI, [c] deaths per 1,000 person-years

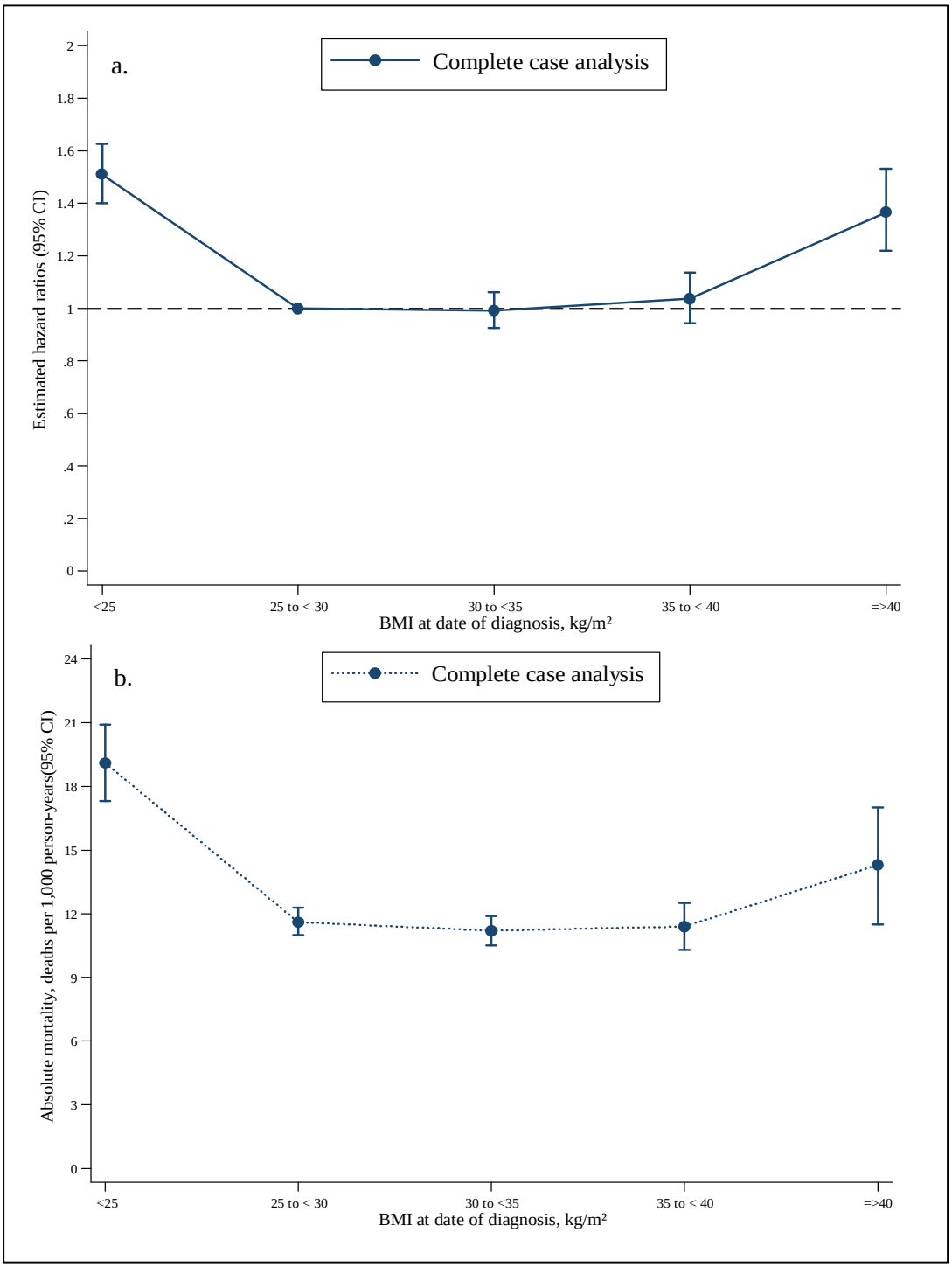

Figure 6.2. a) Hazard ratio estimates for risk of all-cause mortality and b) deaths per 1,000 person-years by categories of BMI at diagnosis of type 2 DM in CCA after exclusion for deaths in the first two years after diagnosis. Reference category is 25 to less than 30kg/m² in a.

When analyses were stratified separately by sex there was no change in the observed relationships (see Appendix Table A6.2).

6.3.3 MCAR: POPULATION MEAN IMPUTATION

Within this sample, 13,775 patients had their missing BMI value replaced with the mean BMI at date of diagnosis, 32 kg/m². Therefore, 45 per cent of patients were now grouped into the 30 to <35kg/m² category. Amongst the cohort of 66,372 patients included in these analyses, there were 36,223 men and 30,149 women. There were 7,592 deaths from all-causes during follow-up and the median follow-up time was 5.76 years.

Baseline characteristics of patients included in the analysis are presented in the appendix (Appendix Table A6.3). Briefly, the addition of these patients into the 30 to < 35kg/m^2 BMI category had little impact on the patterns of clinical and demographic characteristics by categories of BMI, but did increase the follow-up time, lower the proportion of males and increase the proportion of smokers in the 30 to 35kg/m² group.

Table 6.5 presents both the HR estimates for risk of mortality and age-standardised mortality rates by BMI category after adjustment for age, sex, smoking status, deprivation status and Charlson comorbidity index. Figure 6.3 displays a comparison of HR estimates and age-standardised mortality rates obtained after the implementation of population mean-imputation and CCA. Adding patients with missing data into the 30 to <35kg/m^2 category had the effect of raising mortality in this group (HR 1.23 [1.17, 1.31]) compared to when CCA was used (HR: 0.98 [0.92, 1.06]). In agreement with this finding, the age-standardised mortality rate for this group also increased from 11.2 deaths per 1,000 person-years to 14.1 deaths per 1,000 person-years using population mean imputation. Estimates in all other categories of BMI remained the same.

Table 6.5: Estimates of relative hazard of all-cause mortality and age-standardised mortality by categories of BMI using population mean imputation.

	BMI category (kg/m²)				
	< 25	25 to < 30 (REF)	30 to < 35	35 to < 40	≥ 40
N (%)	5,653 (8.52)	16,288 (24.54)	30,143 (45.42)	8,503 (12.81)	5,785 (8.72)
Increase in N from CCA (%) [a]	.	.	13,775 (100.00)	.	.
Multivariable adjusted HR (95% CI) [b c]	1.49 (1.39, 1.61)	1.00	1.23 (1.17, 1.31)	1.04 (0.95, 1.14)	1.37 (1.23, 1.54)
Standard Error	0.057		0.035	0.049	0.079
No. deaths	1,101	1,918	3,577	610	375
Person years	31,218	92,509	171,538	47,925	32,461
Crude mortality rate [d]	35.3	20.7	20.9	12.7	11.6
Age-standardised mortality rate [c d]	19.1 (17.3, 20.9)	11.6 (11.0, 12.3)	14.1 (13.5, 14.6)	11.4 (10.3, 12.5)	14.3 (11.5, 17.0)

[a] Percentage of missing data, [b] Adjusted for age, sex, smoking status, deprivation status and Charlson comorbidity index [c] with 95% CI [d] Deaths per 1,000 person-years. CCA: CCA

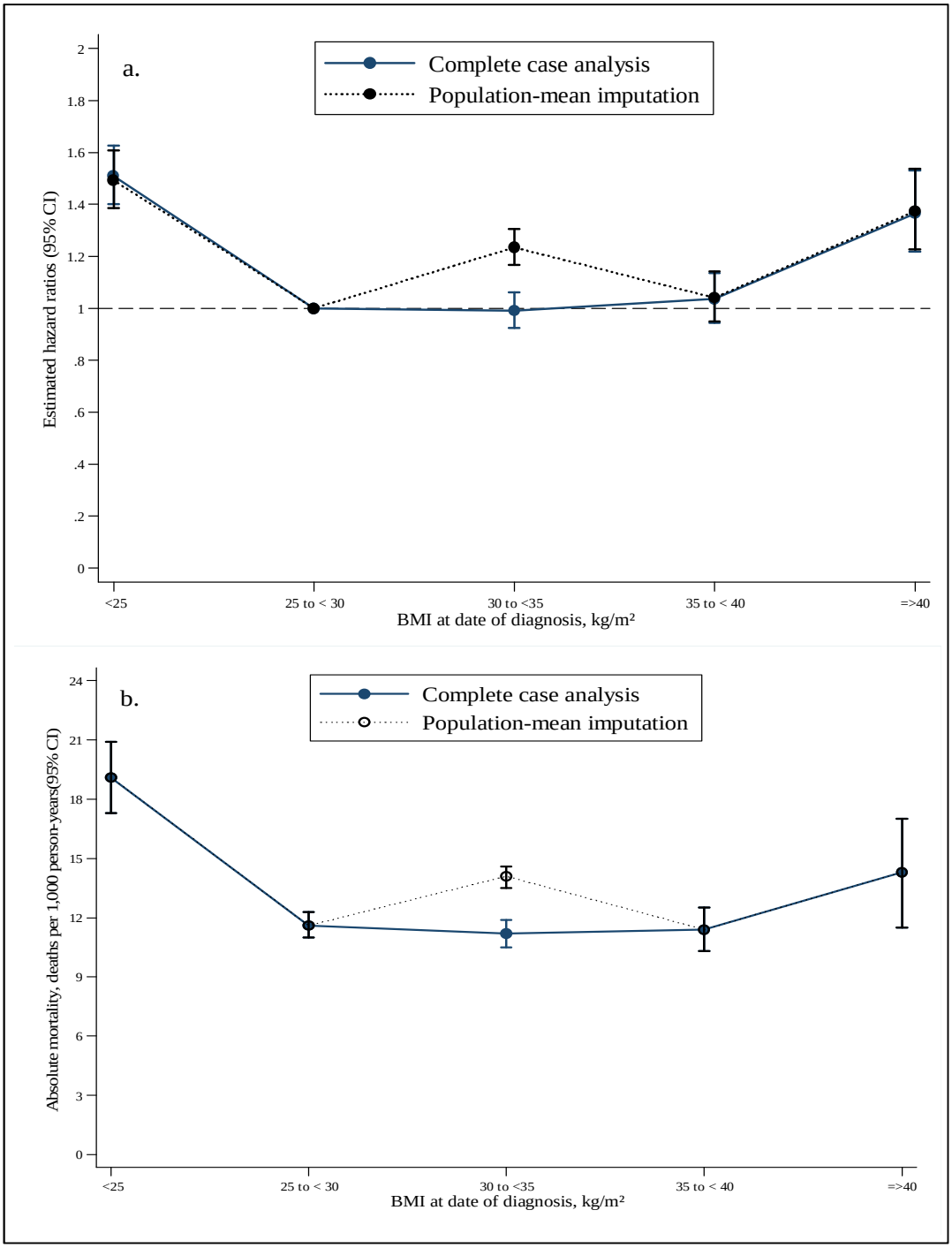

Figure 6.3: Comparison of a) adjusted hazard ratio estimates for risk of all-cause mortality and b) deaths per 1,000 person-years by categories of BMI using CCA and population mean imputation. Deaths in first two years of after diagnosis of type 2 DM are excluded. Reference category is 25 to less than 30kg/m² group in a.

6.3.4 MAR: STOCHASTIC IMPUTATION

Using data from the 52,597 patients with complete observations, stochastic imputation replaced missing BMI categories in 13,775 patients. Estimates obtained from the adjusted Cox proportional hazards model and age-standardised mortality rates are presented in Table 6.6 and Figure 6.4. The U-shaped relationship observed using CCA was flattened with relative estimates of mortality decreasing across all BMI categories when stochastic imputation was applied.

HR estimates from stochastic imputation indicated that patients in the 30 to <35kg/m² and 35 to 40kg/m² had similar mortality to patients in the referent category. Confidence intervals were also narrower than observed when CCA was the method for handling missing data.

As previously stated, relative risk of mortality decreased in all categories of BMI upon the use of stochastic imputation and this is despite the inclusion of patients with missing data whose survival is known to be poorer. Absolute mortality values indicated that stochastic imputation placed a large number of comparatively unhealthy patients with missing data into the reference category and therefore had the effect of raising mortality risk in this group (Age-standardised mortality rate in 25 to less than 30kg/m² category: CCA 11.2 deaths per 1,000 person-years vs. 12.4 using stochastic imputation). Relative comparisons of estimates from stochastic imputation are therefore made between this reference category with higher mortality risk to other BMI categories for which absolute mortality has not increased to the same extent.

Interestingly, absolute measures of mortality for the highest BMI category were identical for CCA and stochastic imputation suggesting stochastic imputation placed individuals with similar survival experiences into this category.

Table 6.6: Estimates of relative hazard of all-cause mortality and age-standardised mortality by categories of BMI using stochastic imputation.

	BMI category (kg/m²)				
	< 25	25 to < 30	30 to < 35	35 to < 40	≥ 45
N (%)	7,201 (10.85)	20,475 (30.85)	20,642 (31.10)	10,781 (16.24)	7,273 (10.96)
Increase in N from CCA (%) [a]	1,548 (11.25)	4,187 (30.44)	4,274 (31.07)	2,278 (16.56)	1,488 (10.82)
Multivariable adjusted HR [b c]	1.42 (1.34, 1.52)	1.00	0.97 (0.91, 1.02)	0.98 (0.91, 1.06)	1.22 (1.10, 1.34)
Standard Error	0.046		0.029	0.040	0.060
No. deaths	1,527	2,696	2,032	832	505
Person years	40,122	117,321	118,676	62,246	42,244
Crude mortality rate [c d]	38.1	23.0	17.1	13.4	12.0
Age-standardised mortality rate [c d]	20.3 (18.6, 22.0)	12.9 (12.3, 13.6)	12.4 (11.7, 13.0)	12.2 (11.2, 13.2)	14.3 (12.9, 15.9)

[a] Percentage of missing data, [b] Adjusted for age, sex and smoking status, deprivation status and Charlson comorbidity index [c] with 95% CI, [d] deaths per 1,000 person-years. CCA: CCA

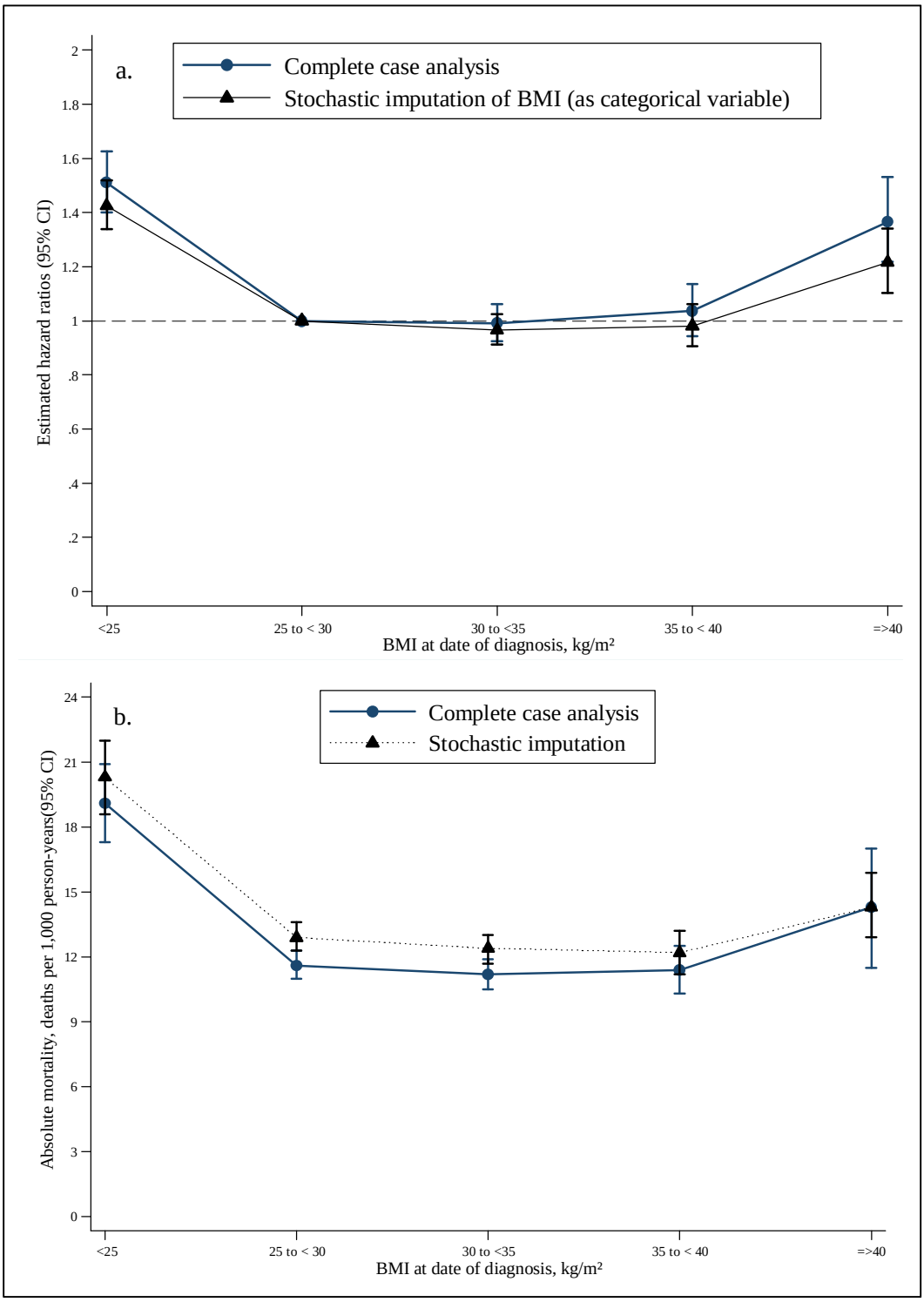

Figure 6.4: Comparison of a) adjusted hazard ratio estimates for risk of all-cause mortality and b) deaths per 1,000 person-years by categories of BMI using CCA and stochastic imputation. Deaths in first two years of after diagnosis of type 2 DM are excluded. Reference category is 25 to less than 30kg/m² group in a.

6.3.5 MAR: MULTIPLE IMPUTATION USING CHAINED EQUATIONS (MICE)

Twenty five datasets were created using observed data from the 52,969 patients with complete data. The process of imputing 25 datasets was computationally intensive and took considerable time.

Table 6.7: Estimates of relative hazard of all-cause mortality and age-standardised mortality by categories of BMI using MICE.

	BMI category (kg/m²)				
	< 25	25 to < 30	30 to < 35	35 to < 40	≥ 40
Average N (%)	7,173 (10.81)	20,481 (30.86)	20,641 (31.10)	10,757 (16.21)	7,319 (11.03)
Average increase in N from CCA (%)[a]	1,520 (11.04)	4,193 (30.43)	4,273 (31.02)	2,254 (16.36)	1,534 (11.14)
Multivariable adjusted HR[bc]	1.38 (1.28, 1.48)	1.00	0.97 (0.91, 1.04)	1.00 (0.90, 1.09)	1.21 (1.08, 1.34)
Standard Error	0.049		0.032	0.046	0.07
No. deaths	1,483	2,713	2,049	842	505
Person years	40,100	117,256	118,615	62,098	42,541
Crude mortality rate [d]	37.0	23.1	17.3	13.6	11.9
Age-standardised mortality rate [c d]	19.3 (17.7, 20.9)	12.9 (12.4, 13.7)	12.4 (11.7, 13.0)	12.4 (11.4, 13.4)	14.2 (12.6, 15.9)

[a] Percentage of missing data, [b] Adjusted for age, sex and smoking status, deprivation status and Charlson comorbidity index [c] with 95% CI, [d] deaths per 1,000 person-years. CCA: CCA

Table 6.7 and Figure 6.5 display both relative and absolute measures of mortality risk by BMI category at time of DM diagnosis. To calculate average numbers in each BMI group, along with age-standardised mortality in the presence of multiple datasets, estimates from each individual dataset were pooled according to Rubin's Rules.

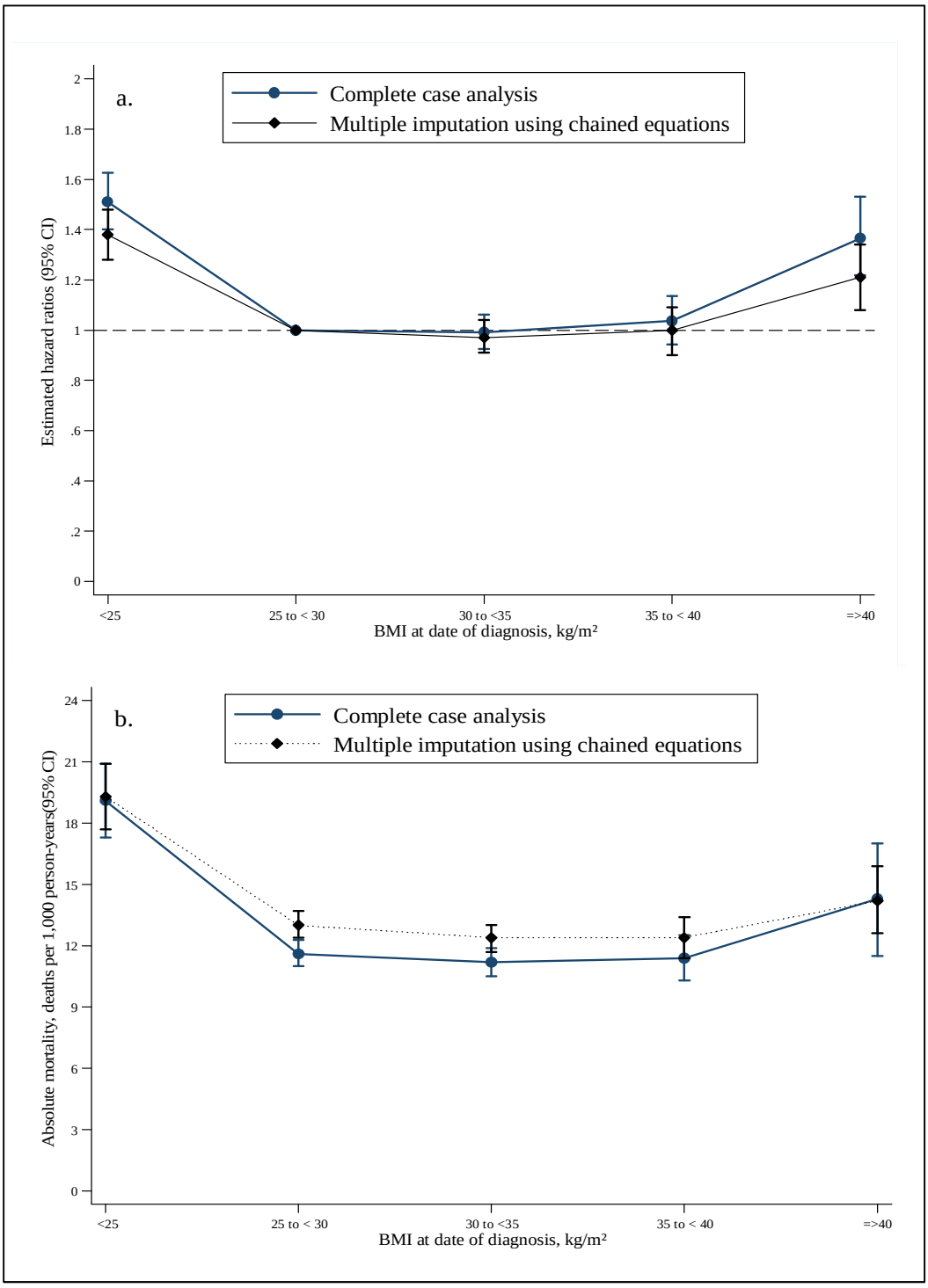

Figure 6.5: Comparison of a) hazard ratio estimates for risk of all-cause mortality and b) deaths per 1,000 person-years by categories of BMI using CCA and MICE. Deaths in first two years after diagnosis of type 2 DM are excluded. Reference category is 25 to less than 30kg/m² group in a).

Like stochastic imputation, MICE acted to flatten out the observed U-shaped relationship in HRs for risk of mortality when CCA was applied. Relative risk of mortality in the highest and lowest BMI categories was lowered when MICE was

applied. The differences in mortality in the lowest and highest BMI categories compared to the reference category were statistically significant.

In absolute terms, mortality rates increased across all BMI groups apart from in the highest BMI category when MICE was applied. Figure 6.6 presents absolute mortality by BMI categories for each imputed dataset. These estimates consistently exhibit a U-shaped relationship. Patients in the ≥40kg/m² BMI category experienced a markedly reduced mortality using this method than when CCA was used (MICE: 7.2 deaths per 1,000 person-years vs. CCA: 8.8).

Despite estimates from stochastic imputation and MICE being largely similar, comparison of the respective assigned BMI categories for each patient and the mode assigned category from MICE, the BMI categories were in agreement in only 4,586 of 13,925 patients with an imputed BMI.

Interestingly, the width of confidence intervals for both the relative and absolute estimates was smaller using multiple imputation than CCA.

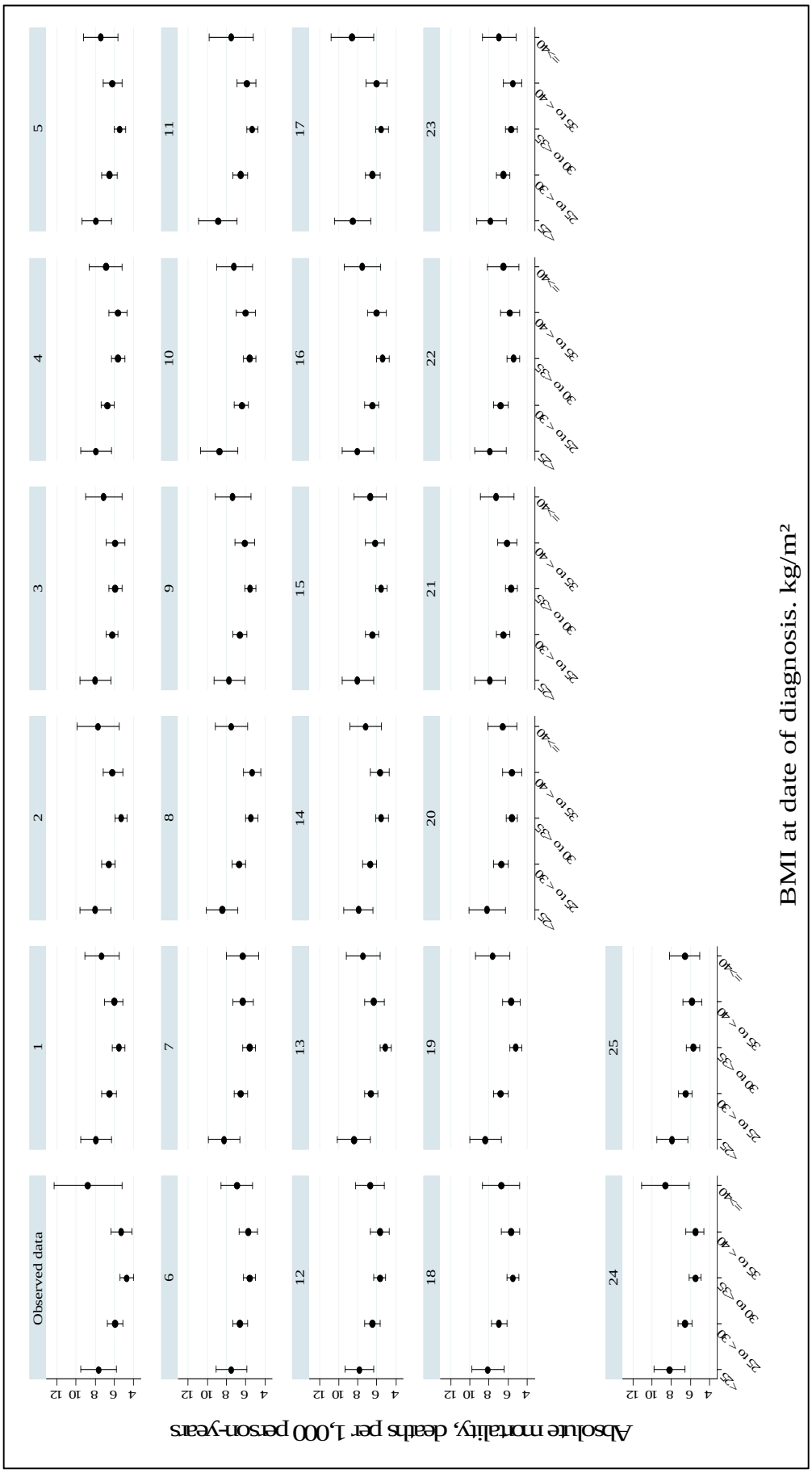

Figure 6.6: Absolute mortality by categories of BMI in each imputed dataset

6.3.6 MAR: MULTIPLE IMPUTATION USING PREDICTIVE MEAN MATCHING (PMM)

Multiple imputation using PMM was used to create 25 datasets. The distribution of BMI categories for each imputed dataset were observed against the observed data and deemed reasonable. Gross discrepancies may have indicated problems with the imputation model.

Results from MICE and multiple imputation using PMM were similar. Compared to results from CCA, multiple imputation using PMM flattened out the U-shaped relationship between BMI and all-cause mortality (Table 6.8, Figure 6.7). Patients in the highest and lowest BMI categories had lowered HRs for risk of mortality than were observed when CCA was used. Compared to results from CCA, multiple imputation using PMM slightly raised the HR for mortality in the 30 to <35kg/m^2 category, this BMI category was associated with lower mortality compared to the reference category (30 to <35kg/m^2 vs. 25 to <30kg/m^2 mv HR: 0.96 [95% CI: 0.83, 1.10]). Table 6.8 also presents the age-standardised mortality results for each BMI category. The absolute risk of mortality was higher in the mid-BMI categories compared to CCA when multiple imputation using PPM was applied. The lowest rates of mortality were observed in the 30 to <35kg/m^2 group. The absolute measure of mortality did not change in the ≤25kg/m^2 group between results from CCA and multiple imputation using PMM.

Table 6.8: Estimates of relative hazard of all-cause mortality and age-standardised mortality by categories of BMI using PMM.

	BMI category (kg/m²)				
	< 25	25 to < 30	30 to < 35	35 to < 40	≥ 40
Average N	7,149 (10.77)	20,568 (30.99)	20,608 (31.05)	10,712 (16.14)	7,335 (11.05)
Average increase in N from CCA (%) [a]	1,496 (10.75)	4,280 (31.10)	4,240 (31.13)	2,209 (16.16)	1,550 (10.87)
Multivariable adjusted HR[b c]	1.19 (1.02, 1.40)	1.00 (REF)	0.96 (0.83, 1.10)	1.08 (0.89, 1.31)	1.30 (1.02, 1.64)
Standard Error	0.0878		0.0713	0.1075	0.1842
No. deaths	1,428	2,732	2,093	848	491
Person years	40,073	117,753	118,289	61,844	42,654
Crude mortality rate [d]	35.6	23.2	17.7	13.7	11.5
Age-standardised mortality rate [c d]	18.8 (16.4., 21.2)	12.9 (12.3, 13.5)	12.5 (11.8, 13.4)	12.5 (11.8, 13.4)	14.5 (12.0, 17.0)

[a] Percentage of missing data, [b] Adjusted for age, sex, smoking status, deprivation status and Charlson comorbidity index, [c] with 95% CI . CCA: CCA

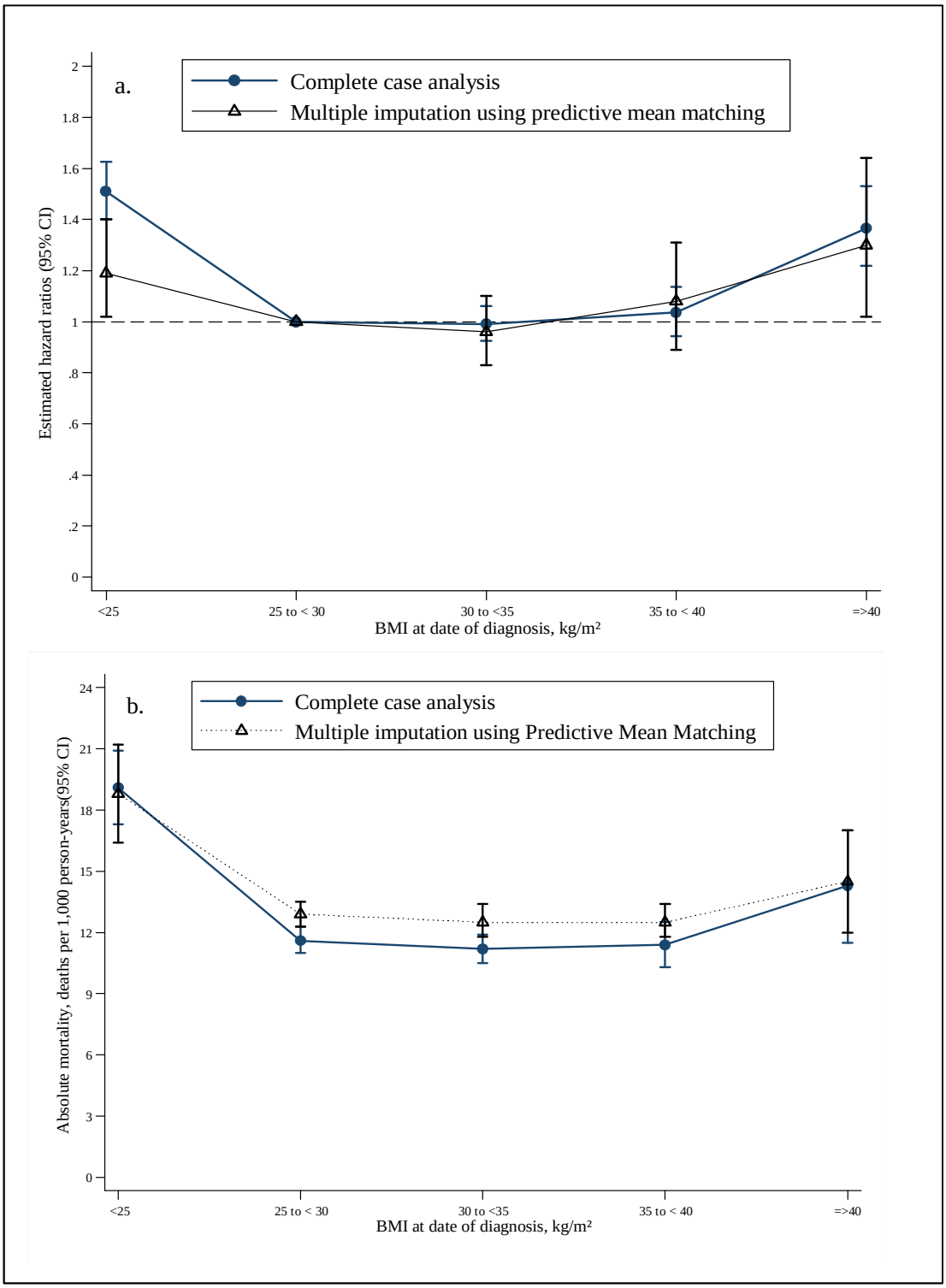

Figure 6.7: Comparison of a) hazard ratio estimates for risk of all-cause mortality and b) deaths per 1,000 person-years by categories of BMI using CCA and multiple imputation using PMM. Deaths in first two years of after diagnosis of type 2 DM are excluded. Reference category is 25 to less than 30kg/m² group in a.

6.3.7 MNAR: SENSITIVITY ANALYSIS USING A PATTERN-MIXTURE BASED APPROACH

A pattern-mixture based sensitivity analysis was conducted on the 25 imputed datasets. Table 6.9 displays the average number of patients in each BMI category, along with HR estimates for all-cause mortality and age-standardised mortality rates. Unsurprisingly, there is a clear left shift in numbers of persons in each category with higher percentages now in the lowest BMI categories. Over 26% of patients with missing data were categorised into the under 25kg/m² group using this pattern-mixture based approach.

Table 6.9: Estimates of relative hazard of all-cause mortality and age-standardised mortality by categories of BMI using a pattern-mixture based sensitivity analysis.

	BMI category (kg/m²)				
	< 25	25 to < 30	30 to < 35	35 to < 40	≥ 40
Average N	9,280 (13.98)	20,509 (30.90)	19,639 (29.59)	10,392 (15.66)	6,551 (9.86)
Average increase in N from CCA (%) [a]	3,628 (26.34)	4,221 (30.64)	3,271 (23.74)	1,927 (13.84)	786 (5.64)
Multivariable adjusted HR[b c]	1.41 (1.31, 1.52)	1.00 (REF)	0.97 (0.90, 1.03)	1.00 (0.91, 1.09)	1.23 (1.09, 1.37)
Standard Error	0.054		0.033	0.045	0.072
No. deaths	1,885	2,598	1,880	789	440
Person years	51,791	116,846	111,383	58,797	36,834
Crude mortality rate [d]	36.3	22.2	16.9	13.4	11.9
Age-standardised mortality rate [c d]	18.9 (17.7, 20.1)	12.8 (12.1, 13.5)	12.1 (11.3, 12.9)	12.2 (11.1, 13.3)	14.1 (12.0, 16.3)

[a] Percentage of missing data, [b] Adjusted for age, sex, smoking status, deprivation status and Charlson comorbidity index, [c] with 95% CI . CCA: CCA

Figure 6.8a displays a comparison of both HR estimates for all-cause mortality by BMI categories between CCA and this pattern-mixture based sensitivity analysis. The two curves are largely similar, with a slightly flattening of the U-shaped relationship. In particular, the risk of mortality in the highest BMI group (≥40kg/m²) was considerably attenuated where this pattern-mixture based sensitivity analysis was applied.

In absolute terms, risk of mortality was slightly raised in BMI categories between 25kg/m² and below 40kg/m². Despite an increase in numbers of patients in the lowest

BMI category, absolute risk of mortality was lowered using this approach compared to CCA.

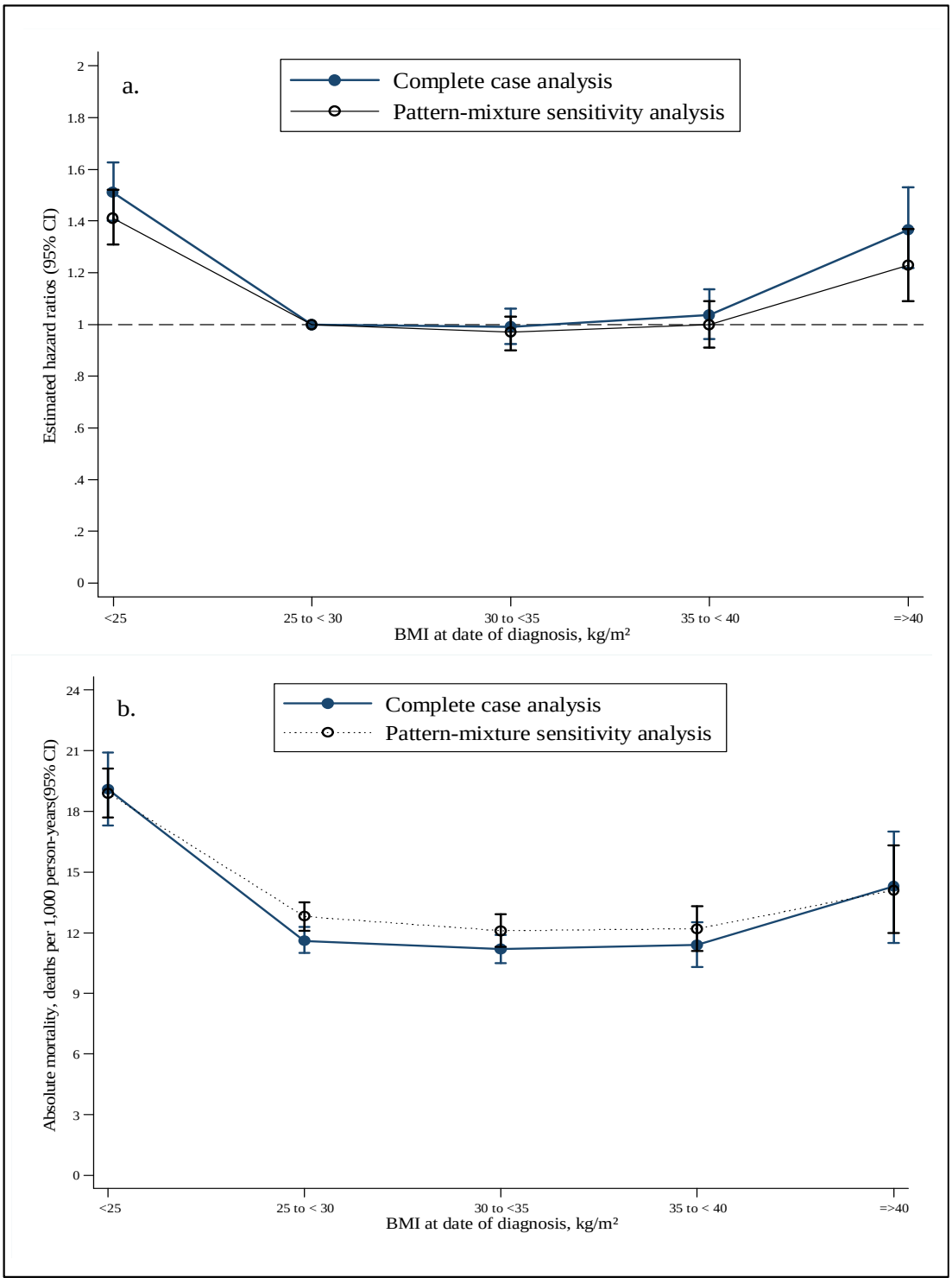

Figure 6.8: Comparison of a) hazard ratio estimates for risk of all-cause mortality and b) deaths per 1,000 person-years by categories of BMI using CCA and a pattern-mixture based sensitivity analysis. Deaths in first two years of after diagnosis of type 2 DM are excluded. Reference category is 25 to less than 30kg/m² group in a.

6.4 DISCUSSION

Dynamic clinical registers offer huge potential for epidemiological research. Clearly, as more large population-based registers become available for research purposes, a key component of all analyses will be the chosen method for handling partially observed data. Unfortunately, the nature of these datasets mean that researchers are typically far removed from the data collection process. Options for improving data quality and minimising the occurrence of missing data are therefore limited. One proposed method which has received considerable attention recently is the use of record linkage to enhance patient records using equivalent data in alternative available datasets (Norris et al., 2000; Faris et al., 2002). Unfortunately no other alternative sources of BMI data were available and therefore this approach was not feasible for this dataset. This problem, whereby data cannot be enhanced with information from other data sources is likely to be common across epidemiology, particularly in health indicators that are not regularly recorded in routine clinical care. Researchers therefore need to apply appropriate statistical techniques to handle the unobserved data. To do so requires knowledge of the likely mechanism through which missing data came about; a complicated process when researchers are not directly involved in the data collection process and data are population-based. Investigators cannot speak to everyone responsible for data collection, and even if they did, those responsible for data collection might not know why data were missing. Furthermore, motivations of clinicians to record risk factors such as BMI are likely to change over time, a feature that was true of our data. It is therefore often necessary to depend instead upon the data itself to explain the probable reasons for missing data. This cannot be achieved with any certainty and therefore the robustness of findings should be assessed by carrying out sensitivity analyses.

6.4.1 RELATIONSHIP BETWEEN BMI AND ALL-CAUSE MORTALITY

Using this large population-based register of persons with type 2 DM in Scotland, the impact of various techniques for handling partially observed data on parameter estimates was investigated.

Six separate methods for handling missing data in the BMI at diagnosis variable were applied. Despite having a wide time-frame during which the recorded measurement

would have been acceptable (10 months prior or 2 months after) in relation to diagnosis date, the proportion of missing data remained high (>20%).

Regardless of the method used for handling missing data, an approximately U-shaped relationship between all-cause mortality and BMI at date of diagnosis was demonstrated. Patients with BMIs in the ≤25kg/m² or ≥40kg/m² category consistently had the highest all-cause mortality. Interestingly, risk of mortality was similar in BMI groups between 25 to <40kg/m² though lowest risk of mortality was consistently observed in the moderately obese (30 to<35kg/m²) category.

These findings correspond well with other studies which have investigated this relationship amongst type 2 DM patients (Church et al., 2005; Hu et al., 2005; Mulnier et al., 2006; Carnethon et al., 2012; Thomas et al., 2014). The problems surrounding missing data were rarely discussed in this body of literature and often approaches chosen to handle missing data were not stated. In fact, rarely did the studies describe the extent of missing data therefore making it difficult to ascertain their potential impact on the respective study's conclusions. Only one study investigating the relationship between BMI and mortality in the general population applied multiple imputation to handle partially observed data (Hotchkiss & Leyland, 2011).

Nonetheless, these studies have identified an apparent protective effect of being overweight or obese, an observation that has also been reported amongst patients with other chronic conditions such as heart disease, COPD and stroke (Oreopoulos et al., 2008; Ovbiagele et al., 2011; Cao et al., 2012). This finding has been restricted to moderately obese patients only, not severely obese.

The finding that obese or overweight patients have a survival advantage over normal weight patients has been termed the obesity paradox and this topic has received considerable attention in recent years as epidemiologists have sought an explanation (Hainer & Aldhoon-Hainerová, 2013; Standl et al., 2013). Typically, it is postulated that this paradox is a consequence of poor study design and methodological flaws (Ferreira & Stehouwer, 2012; Banack & Kaufman, 2013; Hughes, 2013). Residual confounding and selection bias are often cited as potential explanations for the obesity paradox, along with insufficient handling of potential sickness-related weight loss. Smoking history is notoriously prone to misclassification bias and this factor is likely to influence the

validity of estimates for the relationship between BMI and all-cause mortality. In our study, former and current smokers are grouped together which may have diluted the effect of smoking status on estimates.

The largest problem for analyses addressing this topic is potential reverse causation by underlying conditions resulting in high mortality in leaner individuals. In our study, it is possible that this could explain the observation that patients with a BMI in the normal BMI or below category ($\leq 25 kg/m^2$) experienced higher mortality than those in the overweight or obese category. Unfortunately, the short duration of follow-up in our cohort means that this study is particularly prone to reverse causation. To control for this possibility, patients who died within 2 years of diagnosis were excluded and analyses were adjusted for patient Charlson comorbidity index. However, these approaches had a number of problems which may have limited their ability to control against reverse causation. Firstly, patients with some chronic conditions often live more than two years after diagnosis of DM. Secondly, the Charlson comorbidity index only accounts for diagnosed serious comorbid conditions that resulted in hospital admission. Undiagnosed conditions could therefore confuse the observed relationship.

One large study using two prospective cohorts investigating this relationship in DM patients did not identify this effect in their data (Tobias et al., 2014). This study observed a J-shaped association with the lowest mortality occurring in the normal weight category. In a response to this paper, Logue et al, (2014) suggested that this discrepancy may be a consequence of an older cohort. Analyses conducted by Tobias and colleagues included incident DM cases dating back to 1976 and since this time there have been large rises in BMIs whilst improvements in DM management have led to increased life expectancies. These changes may have occurred disproportionately across BMI ranges. Their study also relied on self-reported weight which are typically under-estimated in questionnaires (Gorber et al., 2007). This could therefore lead to attenuated mortality in the normal BMI category.

Given the uncertainty of the true relationship due to conflicting studies and the methodological difficulties in measuring this association, it is difficult to assess whether results obtained from these analyses accurately reflect the real association between BMI and mortality. To date, only observational studies have been carried out. Intervention

studies capable of discerning intentional weight loss from non-intentional weight loss are required.

6.4.2 PERFORMANCE OF MISSING DATA METHODS

6.4.2.1 CCA

Given the findings from earlier studies and information from the data itself, it appears that some methods for handling missing data were more appropriate for this specific analysis than others, and thus gave more valid estimates. For instance, initial investigations into the probable mechanism of missingness indicated that a substantial quantity of missing BMI data were MAR conditional primarily on survival. For this reason, methods which assumed MCAR were unlikely to produce appropriate estimates since the complete cases were no longer representative of the full sample. Consequently, whilst this method was quick and easy to implement, absolute estimates from CCA were likely to have been over-estimated in this cohort.

The exclusion of 13,775 patients with missing data was clearly inefficient and reduced the number of events in the sample from 1,593 deaths to 1,096, reflecting a significant loss of information.

6.4.2.2 POPULATION MEAN IMPUTATION

The second method employed to handle the missing data was mean imputation. Here, the risk of mortality in the BMI category (30 to <35kg/m^2) where all patients with missing data were placed was considerably attenuated relative to the reference group. Clearly, this method's implicit assumption is hugely restrictive, namely that the unseen data all fall in the same category. The method therefore reduced the variability of the observations, a problem which is considerably more problematic when proportions of missing data are high. Despite this method being wholly inappropriate for the problem at hand, it was able to illustrate our finding that patients with missing data had considerably worse survival than those with unobserved data. In some circumstances, this method may therefore be useful as an exploratory analysis, but certainly cannot be recommended as the primary or secondary analysis tool for handling partially observed data.

6.4.2.3 *IMPUTATION APPROACHES FOR MAR DATA*

Estimates from stochastic and multiple imputation approaches were largely similar. Compared to CCA, these techniques acted to increase absolute mortality in all but the highest BMI category. In relative terms this resulted in the flattening of the U-shaped relationship. The finding that the highest BMI category experienced lower absolute mortality when these imputation approaches were implemented indicates that these methods acted to place the healthiest patients with missing data into this category.

Changes in relative estimates from those obtained from CCA would not affect overall inferences. This was in agreement with findings from a simulation study which compared estimates from CCA and multiple imputation (Lee & Carlin, 2012). This study reported that when missing data occurred in the exposure of interest, the benefits of using multiple imputation was minimal. The authors found that multiple imputation was most useful when missing data occurred in confounding variables as then multiple imputation is able to recover information about the relationship between the predictor of interest and the outcome. However, these analyses were restricted to logistic regression analyses. Another simulation study conducted in the context of a survival analysis with unobserved covariate data found that multiple imputation approaches yielded estimates with large bias when the censoring time depended on missing covariates (Qi et al., 2010). The authors therefore suggested that in order to get consistent estimates from multiple imputation, missing data in covariates should be independent of censoring times. Clearly this was not the case in our study and therefore the applied multiple imputation approaches may have produced biased estimates. Qi and colleagues suggest that weighted analyses may perform better in these scenarios. Unfortunately, this approach is less straightforward to apply in standard statistical software.

White and Carlin (2010) also identified that multiple imputation was not always superior to CCA for missing covariate problems. For example when missingness in covariates was independent of the outcome given other covariates, CCA had negligible bias whilst estimates from multiple imputation were biased away from the null. Nonetheless, the authors found that multiple imputation performed better than CCA in a wider range of settings and ought to be preferred due to its greater efficiency. However, it was highlighted that multiple imputation is more prone to human error and poorly specified

models are likely to introduce additional bias that is absent in CCA. This potential limitation of multiple imputation has been observed elsewhere whereby incompatible imputation and analysis models have led to biased parameter estimates in the latter (White et al., 2011; Bartlett et al., 2014).

In accordance with the approach proposed by Royston and White (2009), the Nelson Aalen estimate of cumulative hazard instead of the survival time was included in a separate imputation model. It has been suggested that the inclusion of this variable in the imputation model will reduce the amount of bias in the later Cox proportional hazards model. Despite this being tested in a single simulation study, it has now become widely accepted as the best approach for including the outcome in missing data analyses for survival analyses. When used here, the model produced largely similar results to those presented in 6.3.5 (See Appendix Table A6.4).

Meanwhile, confidence intervals from multiple imputation were narrower than those from CCA. This observation therefore suggests that even when taking into account between-imputation variance, multiple imputation produced smaller standard errors.

Most critically perhaps, was the poor prediction of the BMI at diagnosis variable (r^2=0.15). Unfortunately, few of the available variables were strongly correlated with the BMI at diagnosis variable and therefore the predictive power of the imputation model was poor, leading to uncertain imputed values. This process may have led to patients being placed into the incorrect BMI category, resulting in random misclassification bias. It is uncertain what effect this random misclassification may have had on estimates but it is possible that findings were biased towards the null as demonstrated by the flattening of the U-shaped curve.

The problem of poor predictive power of the imputation model was possibly attenuated through the categorisation of BMI in the imputation model. Categorisation of BMI was carried out to reduce the risk of dilution of the association between BMI and all-cause mortality since the imputation model would assume this relationship was linear. However, categorisation of BMI was likely to have led to significant loss of information. Interestingly, upon the imputation of BMI as a continuous variable, the resulting HRs exhibited a greater curve than was observed when BMI was imputed as a categorical variable (See Appendix Table A6.5). There is currently no general consensus as to how

to handle non-linear relationships and interactions in multiple imputation models (Seaman et al., 2012; Vink & van Buuren, 2013). In order to produce valid estimates, the imputation model should preserve all relationships of interest in the final model. This becomes difficult when the analysis model contains squares and interaction terms and these variables need to be accounted for in the imputation model. At present a number of methods to include variables with more than one functional form have been investigated including passive imputation and the just another variable approach (Von Hippel, 2009). However, these approaches have been shown to produce biased estimates when data are MAR and have not yet been tested in a survival analyses setting (Seaman et al., 2012). Due to this lack of agreement regarding the best approach, BMI at diagnosis was categorised in these analyses. This had the additional benefit of easing comparisons with other studies investigating the association between BMI and all-cause mortality.

Without using BMI data obtained prior to or after diagnosis, there are limited options for improving the imputations. Unfortunately, some potentially reasonable predictors of BMI at diagnosis could not be included in the imputation model due to concerns surrounding collinearity issues and high levels of unobserved data. Nonetheless, in their recent study exploring the inclusion of auxiliary variables, Hardt et al. (2012) concluded that the use of too many auxiliary variables could be disadvantageous unless correlations were high. In order to achieve better predictions using this dataset, it was necessary to utilise the available repeated measurements (Chapter 7). This may not always be possible in other datasets.

Secondly, even with a modest number of predictors and only a single variable to impute, multiple imputation was computationally intensive and time-consuming. In real-life analyses involving routine data, it is unlikely that the pattern of missingness is univariate, and therefore multiple variables will need to be imputed. Consequently, real-life imputation processes are likely to be even more intensive due to the need for multiple iterations. Furthermore, compared to the conventional methods, its application required considerable thought and prior exploratory analyses to correctly specify the imputation model. A thorough understanding of the missing data problem and the subsequent analysis model is required to produce congenial models capable of producing proper imputations. This feature in particular may prove problematic where investigators

involved with handling the missing data are separate from those who subsequently analyse the data, as Rubin (1987) initially proposed when he developed multiple imputation.

Therefore, whilst multiple imputation has become more popular, with all mainstream statistical packages now offering multiple imputation software and numerous review articles providing considerable guidance on its implementation, the aforementioned difficulties may contribute to its low uptake across epidemiological research (Klebanoff & Cole, 2008).

6.4.2.4 *Imputation Approaches for MNAR data*

Regularly, researchers are encouraged to display sensitivity analyses of their results when it is likely that data could be MNAR. However, not a great deal of guidance is available on how to go about doing this. Most literature relating to this problem concerns MNAR data in clinical trials where differential attrition by treatment arm has occurred. Unfortunately, there are few instances whereby MNAR methods have been applied in the non-clinical trial setting. Indeed, software options for these methods are currently limited in standard statistical packages.

Figure 4.2 highlighted that missingness was highest amongst patients whose median ever-recorded BMI were in the under- or normal-weight categories. Since repeat measurements were not included in this chapter's analyses, the risk of MNAR data was high. It was therefore necessary to explore the sensitivity of estimates to departures from MAR. Upon the application of a basic sensitivity analysis, inferences did not change dramatically, with relative risk of mortality increasing in only the <25kg/m² BMI category compared to the 25 to 30kg/m² group. This finding suggests that multiple imputation estimates were robust to possible departures to the MAR assumption. The present analysis used arbitrarily defined assumptions, whereby 50% of patients were considered to have missing data that were MNAR. It was assumed that these patients' BMI were not recorded as it was considered clinically irrelevant. Accordingly we assumed that multiple imputation over-estimated imputed BMI values. Of course, these assumptions are completely unverifiable in these data. It would therefore have been beneficial to explore the sensitivity of MAR-based estimates through further analyses which made both more and less conservative assumptions regarding the mechanism of

missingness. Unfortunately, time constraints prevented further exploration of this issue. However, Chapter 7 investigates the inclusion of longitudinal BMI data which should reduce the risk of data being MNAR.

6.4.3 PROBLEMS & LIMITATIONS

In addition to some of the problems already mentioned, there were a few other potential limitations to these analyses. One complicating factor of these analyses was the fact that patients with missing BMI at diagnosis were also missing a date of measurement. Accordingly, it was necessary to use date of diagnosis as the starting point to calculate follow-up time, rather than measurement date. It is difficult to ascertain the likely effect, if any, this would have on analyses. It is possible that there were differences in survival time by time of measurement from date of diagnosis. For example, if patients were unwell at date of diagnosis, the only available BMI measurement may have been taken long in advance of this date. These un-well patients may therefore have a shorter survival time if they subsequently died during follow-up. Nonetheless, a CCA which used measurement date as the starting point for survival time produced very similar results to those presented earlier which used date of diagnosis as the starting point (not shown).

The method of categorisation of BMI may be a further limitation to this work. More specifically, defining the lowest BMI group as below 25kg/m² may have obscured the dose-response relationship. As a result, it was not possible to obtain an accurate understanding of the effect of low BMI on all-cause mortality. It is possible that the inclusion of patients with a BMI below 20kg/m² in this category considerably inflated the risk of mortality in the below 25kg/m² group. Patients with a BMI below 20kg/m² are likely to be significantly different to patients with a BMI of 25kg/m² in respect to their health status. Similarly, using a BMI of above 40kg/m² as the definition for the highest BMI category may have obscured the association between high BMI and all-cause mortality. Indeed, upon the re-categorisation of BMI with the lowest BMI category becoming below 20kg/m² and the highest BMI group becoming 45kg/m² or above, an analysis of the association between BMI and all-cause mortality using CCA produced a more pronounced U-shaped relationship (see Appendix Figure A6.1). This figure shows that the risk of mortality in the 20 to less than 25kg/m² group compared

to the reference category was lowered from 1.51 using the original classification system to 1.44. Excluding patients with a BMI below 20kg/m² from this category therefore attenuated the risk of mortality in this group. The creation of an additional category for patients with a BMI above or equal to 45kg/m² also led to the attenuation of the risk of mortality in the 40 to less than 45kg/m² group (from 1.37 to 1.23). However, this categorisation scheme was defined prior to conducting the analysis and was used due to concerns surrounding small numbers of patients with a BMI below 20kg/m² and subsequent risk of low power. The number of patients with a BMI below 20kg/m² was 464, of which there were 133 deaths. Of these, 193 patients had a BMI below 18.5kg/m², the "healthy weight" BMI definition. Furthermore, this definition assisted in the comparison of estimates with those obtained from earlier analyses using this dataset. Future work using this dataset will have larger sample sizes and therefore can describe the relationship between low BMIs and all-cause mortality in greater detail.

A further potential problem with these analyses may be missing outcomes. This could arise where patients have emigrated and possibly died elsewhere creating an immortal time bias. This may be of particular concern if patients without a BMI at date of diagnosis may be more likely to be lost to follow-up, thus leading to the belief that they are alive when they are not. Nonetheless, it seems unlikely that this would account for such large differences in survival by missingness.

6.4.4 CONCLUSIONS

Though recording practices in clinical care are improving through initiatives such as QOF, missing data are always going to be a concern where population-registers are used. Despite data being recorded more regularly, the timing of measurements may not occur within the time-frame of interest in any given analysis. This is a unique feature of data analyses using routine data and means that appropriate methods for overcoming missing data need to be used to avoid bias and inefficiency.

In our study we have illustrated the difficulties in assessing the performance of each missing data method without knowing the true associations. However, careful consideration of the plausibility of the assumptions made by each respective method and the use of model diagnostics was able to provide insight into the likely performance

of each method. In these analyses, given the aforementioned problems associated with CCA and the poor prediction of BMI at diagnosis for multiple imputation approaches, it seems probable that the true relationship lies somewhere between estimates observed for CCA and either multiple imputation approach, providing data are MAR. The true relationship in our data is therefore likely to exhibit an approximate U-shaped curve.

These analyses have emphasised the importance of presenting a primary analysis whose assumptions regarding the missing data mechanism are most plausible, followed by a second analysis which investigates the robustness of estimates to departures from the proposed missing data mechanism. Results from these analyses should be compared and potential reasons for disparities should be discussed.

7. USING REPEAT MEASUREMENTS OF BMI TO IMPUTE MISSING DATA

7.1 INTRODUCTION

In the previous chapter, missing data methods which utilised cross-sectional data to predict missing values were applied. It was shown that any additional efficiency benefits obtained by using multiple imputation could be hindered by the limited availability of strong predictors of the missing BMI variable. To overcome this problem, it may be possible to utilise longitudinal patient data to reduce uncertainty in imputed values. Within-patient measurements are likely to be correlated and therefore should provide a powerful guide to plausible values for the missing data. Additionally, conditioning imputation models on longitudinal data may also strengthen the plausibility of the MAR assumption (Welch et al., 2014).

The incorporation of longitudinal data from routinely collected data sources is not straightforward. In these databases, important health indicators are typically recorded as part of a patient's routine clinical care and therefore tend to be recorded on the basis of their clinical relevance. Consequently, recordings of important health indicators occur at irregular time-intervals resulting in wide differences in the number of available measurements per patient. This form of data are termed unbalanced and introduce several additional complexities into statistical analyses with missing data, such as the necessity to arbitrarily define missing data.

This chapter will present and document the extent of unbalanced data in the SCI-DC register. Later, a series of imputation models capable of incorporating longitudinal data will be applied to an analysis of the association between BMI at diagnosis and all-cause mortality.

7.2 DESCRIPTION OF DATA PROBLEM

7.2.1 INTRODUCTION

The key problem with incorporating longitudinal data from the SCI-DC register into missing data analyses is the variation in both the number of available measurements and

in the timing of measurements. This problem is illustrated in Figure 7.1 which presents the pattern of BMI measurements of a typical sample of ten patients without a BMI at date of diagnosis. Within this sample of ten patients, three patients had BMI measurements recorded both before and after the BMI timeframe, two patients only had BMI values taken before the diagnosis timeframe whilst all other patients only had values taken after the diagnosis timeframe. The timing of available measurements was markedly different too. Patient 1 did not have a BMI recording until two years after their diagnosis with DM whilst Patient 9's closest BMI measurement was only a few days outside of the diagnosis time-frame. The number of available measurements by patient fluctuated considerably. For example, patient 2 only had one available BMI measurement, whilst patient 7 had 14 measurements.

Figure 7.1 also highlights large differences in BMI trends over time in patients. Some patient's BMI remained similar across large periods of time and others varied widely in relatively short periods of time. In the context of this study this is important since it is easier to predict patient BMI when individual-level measurements exhibit a stable pattern. The extent of these problems in the overall sample is discussed in more detail in section 7.2.2.

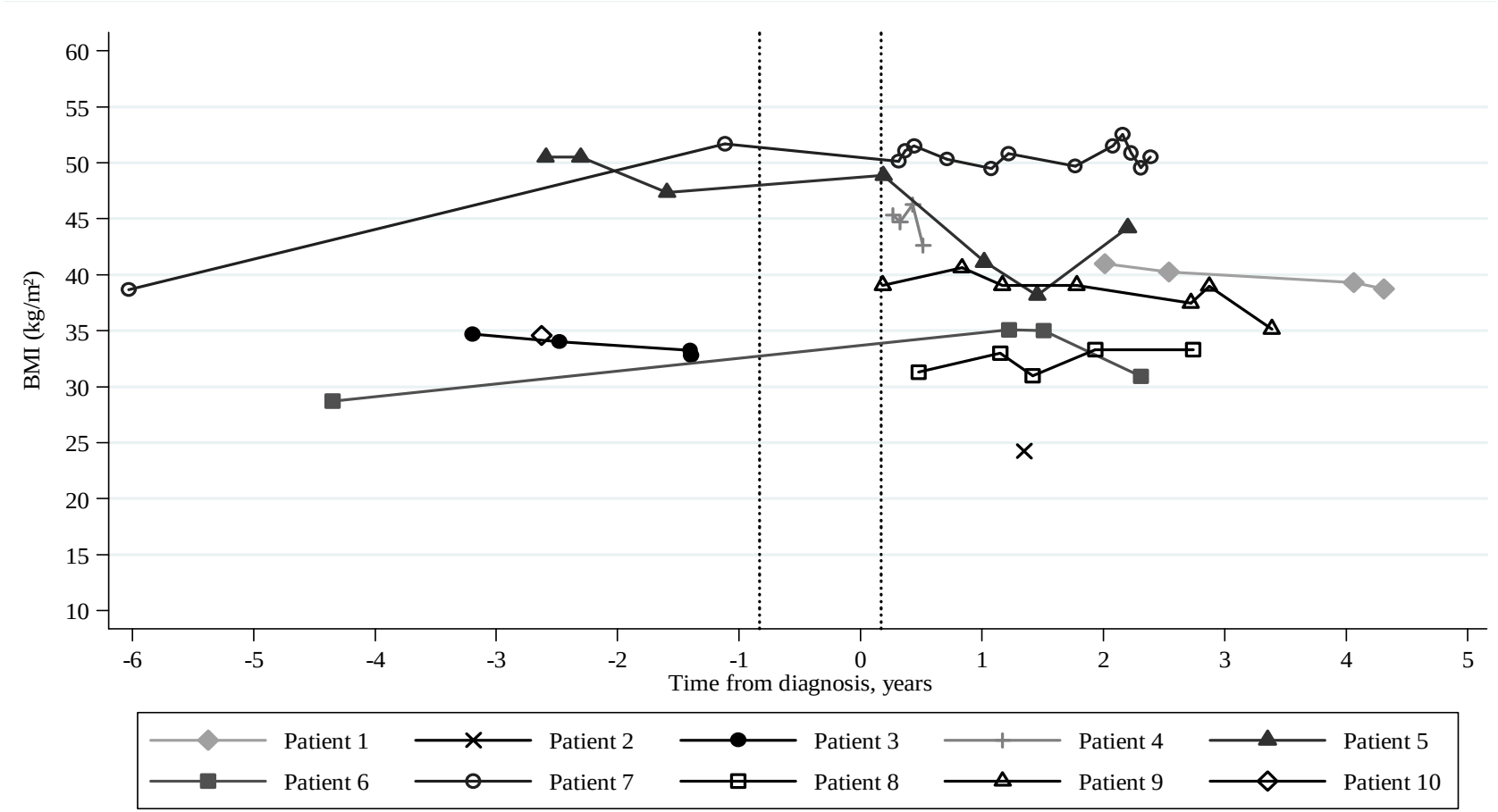

Figure 7.1: Pattern of BMI measurements in relation to date of diagnosis in a random selection of 10 patients. Black dotted reference lines reflect date of diagnosis timeframe.

7.2.2 NUMBER OF OBSERVATIONS PER PERSON

With the widespread use of annual DM review appointments during which key health indicators are recorded, it might be expected that patients would have multiple measurements of BMI available over several years, provided their diagnosis occurred sufficiently long ago. Table 7.1 presents both the number of measurements per patient in the overall sample of patients diagnosed between 2004 and 2008 and in those patients without a BMI at diagnosis. Number of measurements per person were categorised into 6 distinct groups; no measurements, 1, 2, 3 to 10, 11 to 20 and more than 20. The choice of these groups was arbitrary, but allowed the inspection and comparison of patients with vastly different recording patterns. The number of patient BMI measurements ranged between 0 and 79, with a median of 6 measurements per person.

Table 7.1: Number of BMI observations per patient

Number of measurements	N in whole sample (% of 66,372)	N in patients without a BMI at diagnosis value (% of 13,775)
0	1,401 (2.11)	1,401 (10.13)
1	2,241 (3.38)	1,302 (9.44)
2	4,244 (6.39)	1,712 (12.42)
3-10	47,713 (71.89)	8,569 (62.20)
11-20	9,753 (14.69)	763 (5.52)
>20	1,020 (1.54)	40 (0.28)

The table above shows a large amount of variability in the number of BMI measurements per patient in the SCI-DC register for those patients meeting the eligibility criteria described in Section 6.2. As a reminder, this included patients who were diagnosed with type 2 DM between 2004 and 2008, who had not died within two years of diagnosis and who had complete data for variables to be included in either the analysis or missing data model. In the whole sample, 1,401 did not have any BMI measurements and 2,241 patients had only one BMI recording. Of the patients with only a single measurement, 1,302 of these measurements were outside of the date of diagnosis timeframe, defined as a measurement recorded 10 months prior to or 2

months after diagnosis. Identifying a BMI at date of diagnosis for these individuals would therefore mean extrapolating information from a single BMI measurement recorded some time from diagnosis.

In patients without a recording of BMI at diagnosis, 67% had three or more BMI measurements. A substantial quantity of information from which a BMI at diagnosis could be estimated was therefore available.

7.2.3 CHARACTERISTICS OF PATIENTS BY NUMBER OF OBSERVATIONS

To acquire a greater understanding of attendance rates and potential reasons for differences in patterns of missing data, baseline characteristics of patients by their respective number of available BMI recordings were examined (Table 7.2). The number of measurements per person was categorised as in Section 7.2.2. To display patterns of measurements for year of diagnosis and deprivation status, box and whisker plots are presented in Figure 7.2. According to Figure 7.2, patient year of diagnosis was a major determinant of number of BMI measurements per person. Patients diagnosed in 2008 had many fewer measurements than patients diagnosed in 2004. Since BMI data were extracted in 2008, the amount of data collected after diagnosis amongst patients diagnosed with DM in 2008 was severely limited. This may also partly explain the observation that the median years of follow-up increased with increasing available measurements per patient (Table 7.2).

Patients with fewer measurements were older (0 measurements: 64.36 years vs. >20 measurements: 55.64 years) and were less likely to be alive at the end of the study period (0 measurements: 81% vs. >20 measurements 88%). The older age of patients without a single measurement of BMI could be a reflection of the perceived lack of need to measure BMI in elderly patients as weight control may not be prioritised in these patients. Meanwhile, since patients diagnosed at an early age tend to be considerably more overweight than patients diagnosed at a later stage in life, these patients may be monitored more closely for a longer period.

A strong association between number of BMI measurements and median BMI values was observed (1 measurement: Median overall BMI 31.0kg/m² vs. 35.3kg/m² for people with >20 measurements:), indicating patients with higher, more clinically-concerning BMIs were monitored more closely. This relationship may be confounded by age. As

previously stated, patients diagnosed at an earlier age tend to be heavier whilst also tending to have more available BMI recordings. The prevalence of never smokers decreased with increasing number of available measurements indicating more intensive monitoring of smokers. A U-shaped relationship was observed between number of measurements of BMI and the presence of comorbid conditions. BMI measurement may not be a priority for very unwell patients, thus accounting for high levels of comorbidity amongst patients without an available BMI. Conversely, patients with additional comorbidities may have higher GP attendance levels, thus resulting in more opportunities to record BMI measurements.

Mean glycated haemoglobin levels exhibited a general positive association with number of BMI measurements, suggesting patients with poorly controlled diabetes had higher attendance rates, a finding supported by discussions with clinicians (Section 5.3). Upon restricting these comparisons to patients without a BMI at diagnosis only, similar patterns were observed (See appendix Table A7.1 and Figure A7.1). To conclude, there are large differences between patients by number of available BMI measurements, though it is uncertain whether these differences are reflective of shorter periods of follow-up for those patients diagnosed in 2008. Nonetheless, observed differences in median patient BMI by number of measurements were noteworthy and could reflect differences in recording practices by patient weight.

Table 7.2: Characteristics of patients by number of BMI measurement per person

	Number of BMI measurements per patient						TOTAL
	0	1	2	3-10	11-20	>20	
N (%)	1,401 (2.11)	2,241 (3.38)	4,244 (6.39)	47,713 (71.89)	9,753 (14.69)	1,020 (1.54)	66,372
Mean age at diagnosis, years (SD)	64.36 (14.18)	60.99 (13.82)	60.44 (13.33)	61.65 (12.22)	61.79 (11.72)	60.31 (11.72)	61.61 (12.34)
% Male	51.96	56.49	57.49	55.38	50.50	43.24	54.58
% Alive at end of follow-up	80.94	87.10	88.97	88.89	88.24	88.24	88.56
Median no. of years follow-up (IQR)	4.61 (4.07, 5.39)	4.59 (4.05, 5.50)	4.72 (4.15, 5.62)	5.75 (4.74, 6.81)	6.72 (5.62, 7.50)	6.62 (5.31, 7.48)	6.04 (4.89, 7.09)
% With comorbidities	37.26	29.54	28.06	31.18	37.05	41.96	32.08
% Never smoker	47.11	45.65	44.34	42.08	40.27	40.39	42.16
Median ever-recorded BMI, kg/m^2 (SD)	-	30.99 (6.48)	31.16 (6.27)	31.44 (6.08)	32.68 (6.45)	35.27 (7.12)	31.65 (6.22)
Mean Systolic BP, mmHg (SD) [a]	140 (20.25)	139 (19.51)	139 (19.13)	140 (18.77)	139 (18.54)	138 (18.62)	140 (18.82)
Mean Diastolic BP, mmHg (SD) [a]	80 (11.51)	81 (11.07)	81 (10.96)	81 (10.92)	80 (10.77)	80 (10.56)	81 (10.91)
Mean Glycated haemoglobin, % (SD) [a]	8.08 (2.14)	8.10 (2.20)	8.16 (2.16)	8.00 (2.10)	7.97 (2.06)	7.87 (1.99)	8.01 (2.10)
Mean Serum creatinine, umol/L (SD) [a]	90.16 (25.78)	88.94 (29.79)	87.18 (23.71)	90.49 (26.81)	92.34 (26.87)	93.06 (29.11)	90.53 (26.
Mean Total cholesterol, mmol/L (SD) [a]	5.08 (1.29)	5.15 (1.38)	5.23 (1.33)	5.19 (1.30)	5.10 (1.27)	4.94 (1.27)	5.17 (1.30)
Median Triglycerides, mmol/L (IQR) [a]	2.01 (1.34, 2.96)	2.00 (1.43, 2.90)	2.05 (1.40, 2.86)	2.00 (1.40, 2.91)	2.00 (1.42, 2.90)	2.08 (1.45, 2.93)	2.00 (1.41, 2.93)
Mean HDL-cholesterol, mmol/L (SD) [a]	1.23 (0.37)	1.23 (0.39)	1.23 (0.40)	1.25 (0.40)	1.24 (0.40)	1.22 (0.41)	1.24 (0.40)

[a] Closest recording to date of diagnosis with DM

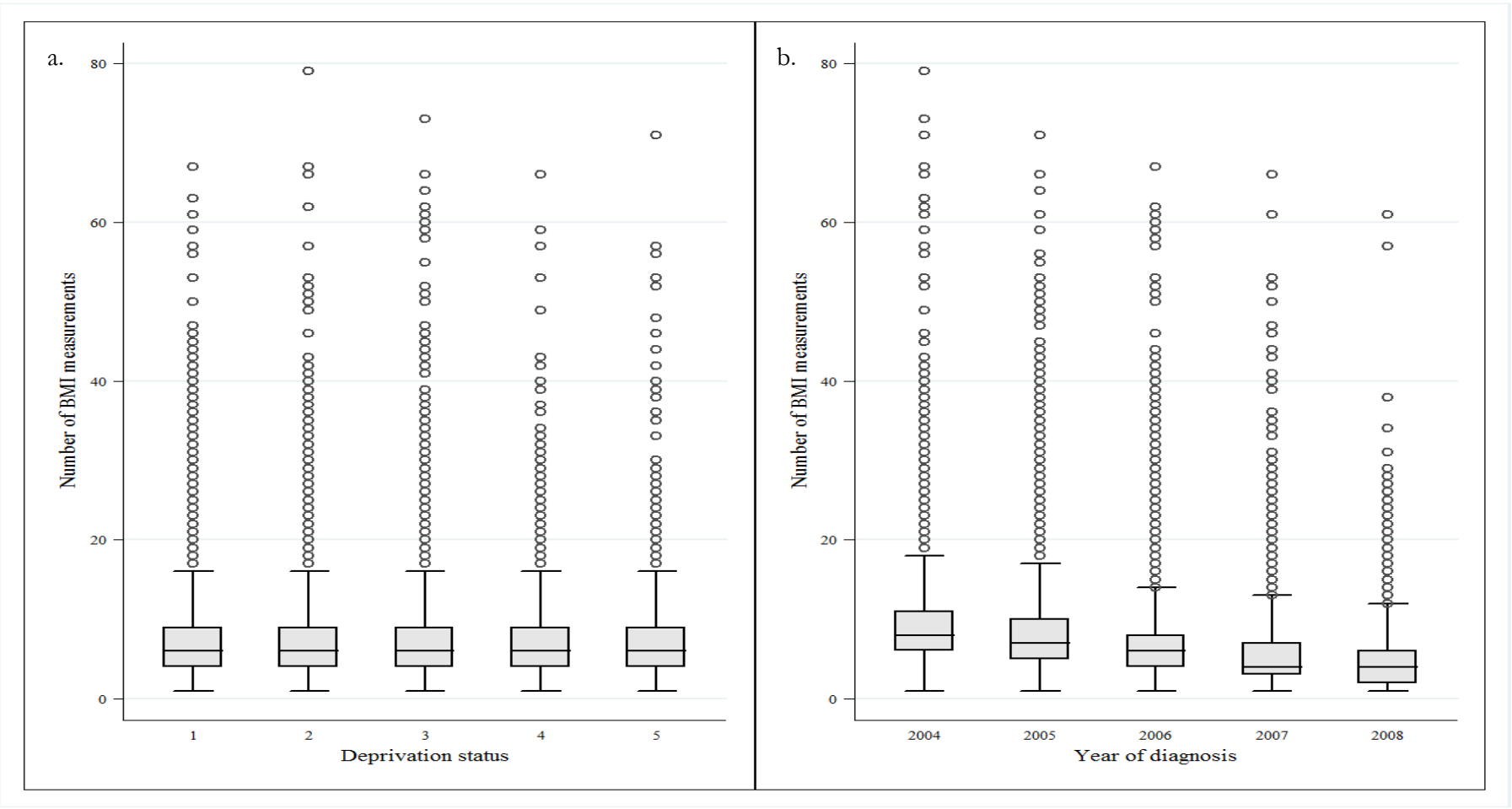

Figure 7.2: Box and whiskers plots of number of measurements by a. Deprivation status (1=Most deprived) and b. Year of diagnosis

7.2.4 TIMING OF MEASUREMENTS

The timing of available measurements relative to the diagnosis time-frame was examined (Table 7.3). For patients without a BMI at diagnosis (n=13,775) who could potentially benefit from data imputation, only 7,557 patients had a valid BMI recorded prior to the date of diagnosis timeframe (Patterns A + C). This would therefore limit the feasibility of applying a Last Observation Carried Forward (LOCF) method to handle the missing data. There were 6,684 patients without an observed BMI who had BMI records either side of diagnosis with DM (Pattern C). Accordingly, interpolation of BMI could not be applied to estimate an approximate BMI value at diagnosis for all patients.

Table 7.3: Patterns of observations in relation to date of diagnosis timeframe.

	Pattern	Pattern of observations			Frequency (% of 66,372)
		Before date of diagnosis	Date of diagnosis timeframe	After date of diagnosis	
Patients without a BMI at date of diagnosis	A	✓	.	.	873 (1.32)
	B	.	.	✓	4,817 (7.26)
	C	✓	.	✓	6,684 (10.08)
	D	.	.	.	1,401 (2.11)
Patients with a BMI at date of diagnosis	E	✓	✓	✓	34,445 (51.90)
	F	.	✓	✓	12,298 (18.53)
	G	✓	✓	.	4,377 (6.59)
	H	.	✓	.	1,477 (2.22)

In addition, the median number of years that the last measurement was taken prior to date of diagnosis was 2.94 years, with an IQR of 1.45 to 6.97 years (or 2.12 years [IQR:0.63, 6.15] from diagnosis timeframe), reflecting a considerable time before time of diagnosis. Indeed, only 7% of patients had a valid BMI reading in the preceding year before the date of diagnosis timeframe.

Following a diagnosis with type 2 DM, BMI appears to be monitored more closely with 87% of patients having an available measurement taken after the defined date of diagnosis timeframe. The median time until the closest available BMI measurement in

patients without a BMI at diagnosis value was equal to 0.37 years with 50% of measurements occurring between 0.24 and 0.73 years (or 0.21 years after defined diagnosis timeframe [IQR: 0.08, 0.57]).

The number of available measurements before and after the diagnosis time-frame highlights the intensification of risk factor monitoring upon diagnosis with DM; the median number of within-patient measurements before diagnosis: 1 (IQR: 0, 2) vs. after diagnosis: 3 (IQR: 1, 5).

7.3 INCORPORATING LONGITUDINAL DATA TO IMPROVE IMPUTATIONS

7.3.1 INTRODUCTION

There are various options for incorporating longitudinal data into missing data analyses, each of which have features that make them more or less appropriate to apply to this routinely collected dataset. This section illustrates the use of these methods in an analysis of the association between BMI at diagnosis with DM and all-cause mortality.

7.3.2 METHODS

This subsection will describe the methods applied to handle missing data in the SCI-DC register and is divided into single imputation and multiple imputation approaches. Results from each approach will be compared to results from CCA.

7.3.2.1 EXCLUSION CRITERIA

The exclusion criteria are displayed in Figure 7.3. Unfortunately, some of the selected methods for handling missing data could not identify a suitable value for the missing data point when patients did not have any BMI measurements at all. Therefore to ease the comparison of estimates from each method, a further 1,401 patients with no BMI readings were excluded from the analyses.

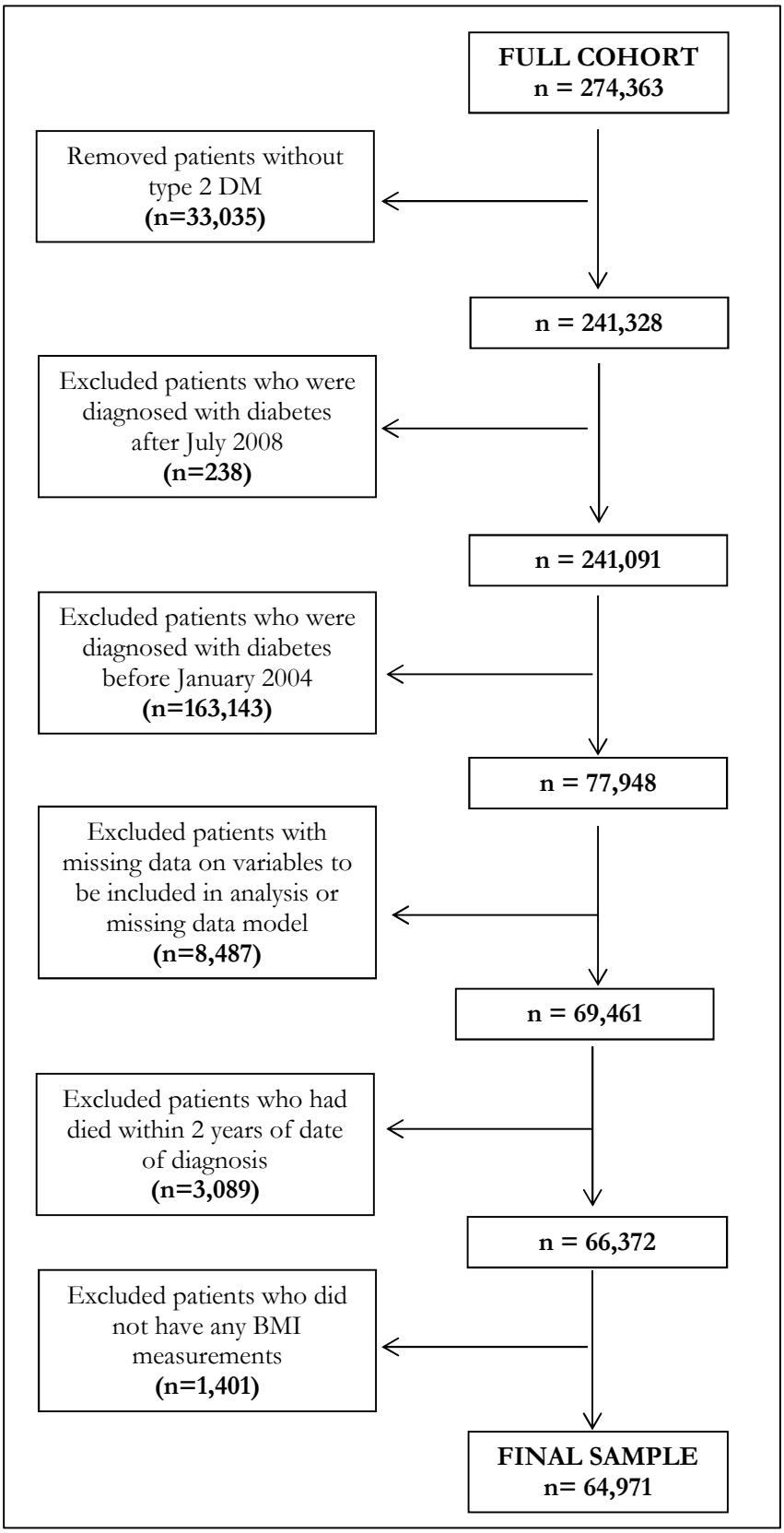

Figure 7.3: Exclusion criteria for analysis involving use of longitudinal data to impute missing data

7.3.2.2 DELETION APPROACH

COMPLETE CASE ANALYSIS

Analyses were conducted only on patients with complete data. This analysis will serve as a comparison to the other analyses using alternative methods of handling unobserved data.

7.3.2.3 SINGLE IMPUTATION APPROACHES

CLOSEST OBSERVATION IMPUTATION

Missing values were replaced with those observations that lay closest to the date of diagnosis timeframe. In essence, this method simply extends the diagnosis time-frame.

EXTRAPOLATED BMI IMPUTATION

Missing BMI values at diagnosis were replaced with predicted values from individual-level linear regressions using all available BMI measurements. Where patients only had a single BMI measurement, this single BMI measurement was used.

7.3.2.4 MULTIPLE IMPUTATION APPROACHES

INCLUDING EXTRAPOLATED BMI IN IMPUTATION MODELS

uvis for STATA was used to create 25 multiply imputed datasets to coincide with the proportion of missing data, as recommended by Bodner (2008). The imputation model included the following variables BMI at diagnosis (categorical), age at diagnosis (continuous), sex (binary), year of diagnosis (categorical), deprivation status (categorical), vital status (binary), smoking status (binary), follow-up time (continuous), Charlson comorbidity index (binary), diastolic BP (continuous), total cholesterol (continuous), HBA (categorical) and extrapolated BMI (categorical). Extrapolated BMI was divided into 5kg/m^2 intervals (<25kg/m^2, 25 to < 30kg/m^2, 30 to <35kg/m^2, 35 to 40kg/m^2 and ≥40kg/m^2).

INCLUDING MEASUREMENTS RECORDED BEFORE AND AFTER DIAGNOSIS

The *ice* command in STATA was used to include measurements taken one (t-1), two (t-2) and three years (t-3) before and one (t+1), two (t+2) and three (t+3) years after the diagnosis time-block. Where more than one measurement was available in the year block, the nearest measurement to diagnosis chronologically was used. The inclusion of these measurements introduced an additional six variables with missing data and

therefore necessitated the use of the *ice* command rather than *uvis* (Royston, 2005). *uvis* is primarily designed for imputing missing values in a single variable based on multiple regression on fully observed variables in the imputation model.

Briefly, *ice* imputes missing values in a number of variables by using regression switching. This involves the generation of conditional distributions for each variable with missing data using observed data on all variables included in the imputation model. Imputed values are generated by taking simulated draws from the respective posterior predictive distribution (more detail may be found in Section 2.3.1.1). This process is repeated for several cycles or iterations in order to stabilise results.

In its application here, *ice* was used to impute 60 datasets. This number of imputed datasets was selected following the inclusion of further incomplete variables. The model included the follow variables BMI at diagnosis, age at diagnosis, sex, year of diagnosis, deprivation status, vital status, smoking status, Charlson comorbidity index, follow-up time, diastolic BP, total cholesterol, HBA, BMI (t-1, t-2, t-3, t+1, t+2, t+3). These BMI measurements were included as categorical variables and were imputed using multinomial logistic regression. The imputation process was iterated through 10 cycles, the default in STATA.

TWO FOLD FULLY CONDITIONAL SPECIFICATION (FCS)

Two-fold FCS is an extension of the MICE model and was proposed by Nevalainen et al, (2009) to handle intermittent patterns of missing longitudinal data. To achieve this, two fold FCS only conditions on variables at the time-point of interest and measurements in immediately adjacent time-blocks only. It has been proposed that this approach has two chief benefits over the inclusion of all time-blocks in the same model. Firstly, this approach reduces the risk of over-fitting and therefore collinearity problems by reducing the number of variables conditioned on in each regression. Secondly, this approach allows for the temporal ordering of the measurements to be taken into account.

The STATA command *twofold* was used to implement this technique (Welch et al., In print 2014). Available BMI measurements were again divided into yearly time-blocks outside of the BMI at diagnosis time-frame up to 3 years before and after. BMI was therefore included at time *t* (diagnosis time frame), t-3, t-2, t-1, t+1, t+2 and t+3.

As per the default settings, the number of among-time iterations was set to 10 and the number of within-time iterations was set to 5. Twenty-five imputed datasets were created. The following fully observed time-independent variables were included; age at diagnosis, sex, year of diagnosis, deprivation status, Charlson comorbidity index, HBA, systolic blood pressure and total cholesterol. Since smoking status was the most correlated variable with BMI and multiple measurements were available, this was included as a binary time-dependent variable. The outcomes for the substantive model, vital status and follow-up time were also included.

7.3.3 RESULTS

After exclusion of patients without any BMI measurement, there were a total of 64,971 patients in the cohort, of whom 12,380 patients were without a recorded BMI at diagnosis. Patients without missing data contributed 300,334 person years of follow up during which there were 5,491 deaths at a rate of 18.3 deaths/1,000 person-years. Patients with missing BMI data contributed 73,559 person-years during which there were 1,834 deaths at a rate of 24.9 deaths per 1,000 person-years.

7.3.3.1 CCA

CCA resulted in the exclusion of 12,380 patients without a BMI at date of diagnosis. Obtained results were the same as presented in Section 6.3.2.

7.3.3.2 CLOSEST OBSERVATION IMPUTATION

Closest observation imputation was applied to impute values for the 12,380 patients who were without a BMI at diagnosis. Table 7.4 presents the multivariate HRs, adjusted for age at diagnosis, sex, smoking status, deprivation status and Charlson comorbidity index, along with absolute age-standardised mortality rates. Figure 7.4 presents a comparison of relative and absolute estimates from CCA and closest observation imputation.

Compared with estimates from CCA, replacing missing values with the closest available measurement had very little impact on the observed relationship between BMI at diagnosis and mortality. Relative risk of mortality was slightly lowered in the 30 to 35kg/m² group and raised in the less than 25kg/m² group when compared to CCA estimates, though these changes were negligible. In absolute terms, levels of mortality

increased across all BMI categories upon using this imputation approach. Mortality in the less than 25kg/m² group was raised from 19.1 deaths per 1,000 person-years using CCA to 20.9 deaths per 1,000 person-years when closest observation imputation was applied. These results were almost identical to estimates obtained when replacing individual-level median BMIs to replace missing values (results not shown).

Table 7.4: Estimates of relative hazard of all-cause mortality and age-standardised mortality by categories of BMI using closest observation imputation.

	BMI category (kg/m²)				
	< 25	25 to < 30 (REF)	30 to < 35	35 to < 40	≥ 40
N (%)	7,427 (11.43)	20,343 (31.31)	19,864 (30.57)	10,313 (15.87)	7,024 (10.81)
Increase in N from CCA (%)[a]	1,776 (14.35)	4,056 (32.76)	3,498 (28.26)	1,811 (14.63)	1,239 (10.01)
Multivariable adjusted HR (95% CI)[bc]	1.54 (1.45, 1.64)	1.00	0.98 (0.92, 1.04)	1.05 (0.97, 1.14)	1.36 (1.23, 1.51)
Standard Error	0.050		0.030	0.043	0.070
No. deaths	1,592	2,545	1,891	797	490
Person years	41,181	115,437	113,133	58,567	39,722
Crude mortality rate[c]	38.6	21.9	16.7	13.6	12.3
Age-standardised mortality rate[c]	20.9 (19.3, 22.5)	12.2 (11.5, 12.8)	11.8 (11.1, 12.2)	12.1 (11.1, 13.2)	15.8 (13.5, 18.1)

[a] Adjusted for age, sex, smoking status, deprivation status and Charlson comorbidity index, [b] with 95% CI, [c] deaths per 1,000 person-years. CCA: CCA

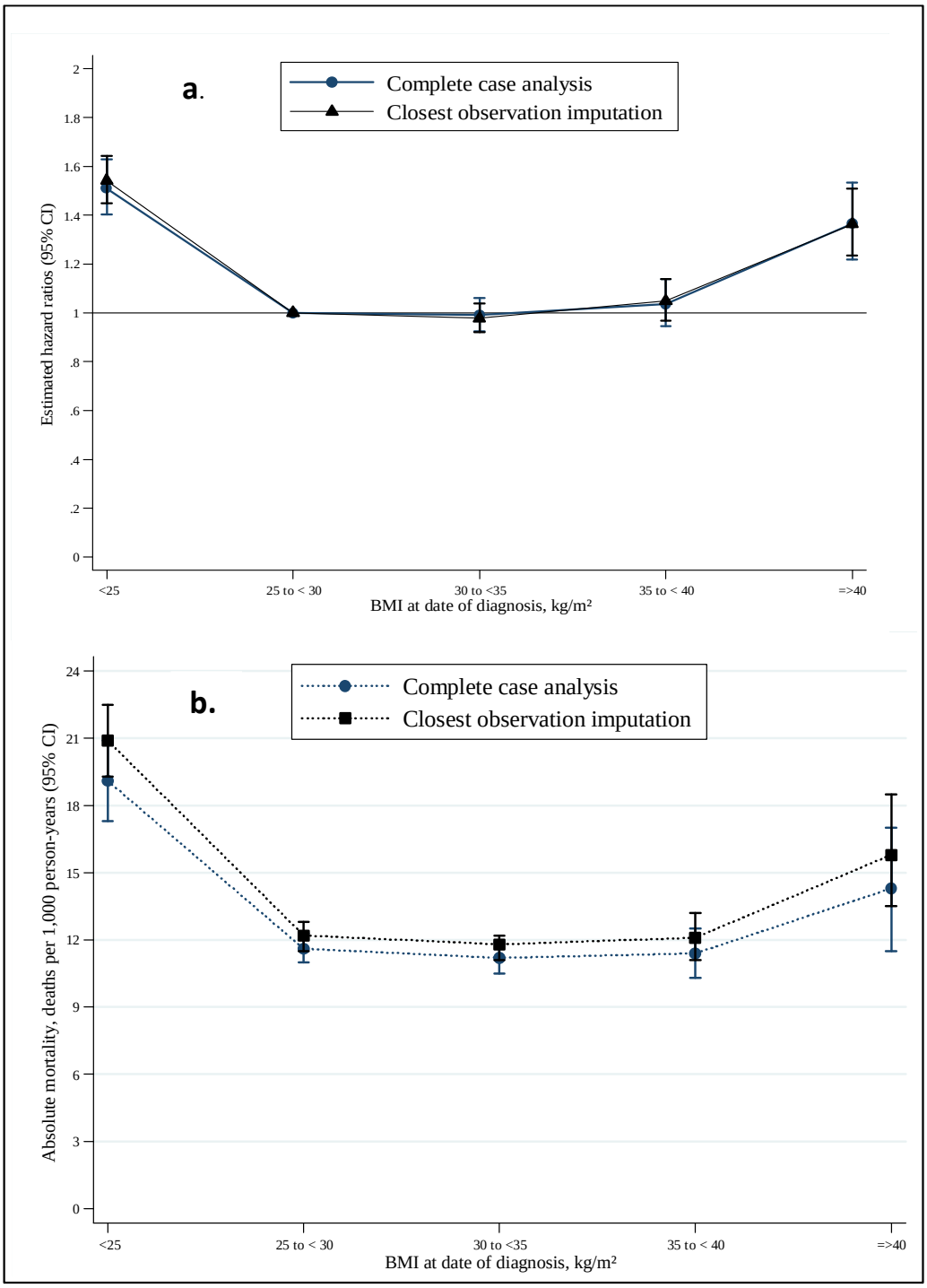

Figure 7.4: Comparison of a) Estimates of relative hazard of all-cause mortality and b) deaths per 1,000 person-years by categories of BMI at diagnosis of diabetes after using closest observation imputation and CCA. Deaths in the first two years after diagnosis of type 2 DM were excluded. Reference category is 25 to less than 30kg/m² group in a).

Overall, the median time from diagnosis at which a BMI was available was 23 days (IQR: 6, 66). Ninety percent of patients had an available BMI measurement within 168

days of diagnosis and 99% had a measurement within 969 days of measurement. Amongst imputed values, the median time from diagnosis at which measurements were recorded was 148 days (IQR: 91, 324). Ninety-one percent of imputed measurements were recorded after diagnosis at a time when DM treatments were likely to have affected measurements.

To ascertain how well measurements recorded outside of the BMI time-frame described BMI at diagnosis measurements, a series of comparisons were made between measurements recorded at diagnosis and closest available measurements. In order to do this, it was necessary to firstly identify patients with a BMI at diagnosis and a BMI measurement recorded before diagnosis (Patterns E & G in Table 7.3) and secondly patients with a BMI at diagnosis and a BMI measurement recorded after diagnosis (Patterns E & F in Table 7.3). The validity of using patients with observed BMI data and additional records prior to or after diagnosis to assess the likely performance of using measurements recorded before and after diagnosis is questionable. Patients without data at diagnosis may have very differing trends in BMI around diagnosis than those with BMI data at diagnosis. Caution is therefore required when interpreting these results.

To compare measurements, Pearson's correlation coefficients and linear regression coefficients were calculated. Mean differences between measurements taken before and at diagnosis with DM were also calculated. Both measurements of BMI were then categorised and the percentage concordance between categories was examined. All results are stratified by time at which measurements recorded before were available.

Measurements taken before diagnosis time-period

Overall, 38,820 patients had an available measurement recorded before the diagnosis timeframe, Table 7.5 displays the correlation and concordance between measurements recorded before and measurements recorded within the diagnosis timeframe. Results are stratified by years before diagnosis. Fifty percent of closest measurements recorded before diagnoses were measured within two years prior to the diagnosis time-frame. A further 25% of closest measurements were recorded five years or more before diagnosis.

Overall, the Pearson's Correlation coefficient for closest measurements recorded before diagnosis was 0.88. The mean difference between measurements (Closest BMI before

diagnosis timeframe – BMI at diagnosis) was equal to 0.51kg/m² (SD: 3.15), indicating a slight weight gain before diagnosis with DM. The overall concordance between measurements recorded before and at diagnosis was reasonable. Sixty-seven percent of patients would not change BMI category if the closest BMI measurement was used, with 33% of patients changing categories. Of the patients whose categories were not in agreement, the vast majority differed only by a single higher or lower level of BMI (See Appendix Table A7.2).

The level of agreement between measurements decreased as the time between measurement and diagnosis increased. Measurements recorded one year prior to the diagnosis timeframe were more strongly correlated with measurements at diagnosis (Pearson's correlation 0.93) than measurements recorded over five years before diagnosis (Pearson's correlation 0.77).

Table 7.5: Agreement between closest measurement recorded before diagnosis and measurement recorded at diagnosis with DM.

Timing of available measurement *before* diagnosis (Years)	N	Mean difference (BMI at diagnosis – Closest BMI before)	Pearson's Correlation coefficient	Linear regression Coefficient	% Concordance in BMI category
0 to <1	4,221 (10.87)	0.05 (2.44)	0.93	0.954 (0.943, 0.965)	76.23
1 to <2	15,096 (38.89)	0.13 (2.47)	0.92	0.955 (0.949, 0.961)	74.67
2 to <3	5,072 (13.07)	0.21 (2.91)	0.90	0.943 (0.930, 0.956)	68.42
3 to <4	2,696 (6.94)	0.40 (3.19)	0.88	0.942 (0.923, 0.961)	65.88
4 to <5	2,006 (5.17)	0.77 (3.38)	0.85	0.929 (0.903, 0.954)	61.31
≥5	9,729 (25.06)	1.44 (4.11)	0.77	0.890 (0.875, 0.904)	53.29
Overall	38,820	0.51 (3.15)	0.88	0.925 (0.920, 0.930)	67.34

Measurements taken after diagnosis time-period

There were 46,741 patients with both a BMI at diagnosis and a BMI recorded after diagnosis. Table 7.6 displays the correlation and concordance between these measurements. Results were stratified into much smaller time intervals than in Table 7.5 due to the high availability of recordings in the first few years following diagnosis.

Overall, the correlation between available measurements was strong (Pearson's correlation: 0.95). The correlation between closest measurements and BMI at diagnosis was strongest amongst measurements recorded 0.25 years after diagnosis (Pearson's correlation: 0.96) and gradually declined with increasing time from diagnosis. Percentage concordance also decreased with increasing time, despite the use of very narrow time-frames. Overall concordance of measurements used after the diagnosis time-frame was 78%. This value is higher than observed in Table 7.5 for measurements recorded before diagnosis but this possibly reflects the shorter time-frame. For those measurements which were not in agreement, 12% of BMI measurements recorded after diagnosis were in higher BMI categories and 10% were lower than those measurements recorded at diagnosis (Appendix Table A7.3).

The mean difference between measurements recorded at diagnosis and the closest measurement recorded after diagnosis was -0.45kg/m², thus reflecting an overall decrease in BMI after diagnosis. Interestingly, the mean difference in measurements decreased with increasing time (Mean difference in BMI measurements at 0 to <0.25years: 0.47kg/m² vs. Mean difference in BMI measurements at ≥1.25 years: 0.28kg/m²). This could reflect weight loss followed by weight gain over the longer period of time or perhaps biased recording of weight for people that lost weight soon after diagnosis. These patients were therefore more likely to have measurements recorded.

Table 7.6: Agreement between closest measurement recorded after diagnosis and measurement recorded at diagnosis with DM.

Timing of available measurement *after* diagnosis (Years)	N	Mean difference (BMI at diagnosis – Closest BMI after)	Pearson's Correlation	Coefficient	% Concordance in BMI category
0 to <0.25	10,412 (22.28)	0.47 (1.85)	0.96	0.984 (0.979, 0.990)	80.70
0.25 to < 0.50	17,310 (37.03)	0.51 (1.94)	0.96	0.976 (0.971, 0.980)	77.82
0.50 to <0.75	8,361 (17.89)	0.47 (2.09)	0.95	0.952 (0.945, 0.959	76.56
0.75 to <1.00	4,257 (9.11)	0.39 (0.26)	0.94	0.952 (0.941, 0.962)	75.47
1.00 to <1.25	3,752 (8.03)	0.32 (2.41)	0.93	0.923 (0.911, 0.935)	74.91
≥ 1.25	2,649 (5.67)	0.28 (6.25)	0.93	0.937 (0.922, 0.951)	73.67
Overall	46,741 (100.00)	0.45 (2.05)	0.95	0.965 (0.962, 0.968)	77.55

7.3.3.3 *EXTRAPOLATED BMI IMPUTATION*

Absolute and relative risks of all-cause mortality by category of BMI are presented in Table 7.9. These estimates are presented along with estimates from CCA in Figure 7.5. Compared with CCA, extrapolated BMI imputation produced a more marked U-shaped relationship with the lowest risk of death occurring in the 30 to <35kg/m² category. The risk of mortality in this group dropped from 0.98 to 0.87 upon using extrapolated BMI imputation. Similarly, compared to when CCA was used, the risk of mortality was considerably lowered in the <25kg/m² category (mv. HR using CCA: 1.56 vs. 1.35 using extrapolated BMI imputation), whilst the HR remained the same in the two highest BMI categories. Absolute risk of mortality increased across all BMI categories using extrapolated BMI category. The smallest changes in absolute mortality were observed in the 30 to <35kg/m² category.

Table 7.7: Estimates of relative hazard of all-cause mortality and age-standardised mortality by categories of BMI using extrapolated BMI imputation

	BMI category (kg/m²)				
	< 25	25 to < 30 (REF)	30 to < 35	35 to < 40	≥ 40
N	7,429 (11.43)	20,396 (31.33)	19,841 (30.53)	10,261 (15.84)	7,044 (10.84)
Increase in N from CCA (%)[a]	1,778 (14,36)	4,109 (33.19)	3,475 (28.07)	1,759 (14.21)	1,259 (10.17)
Multivariable adjusted HR (95% CI)[bc]	1.35 (1.18, 1.54)	1.00	0.87 (0.76, 0.99)	1.04 (0.87, 1.24)	1.39 (1.13, 1.73)
Standard Error	0.092		0.059	0.094	0.154
No. deaths	1,583	2,577	1,859	789	507
Person years	41,197	116,774	113,049	58,231	39,784
Crude mortality rate [d]	38.4	22.1	16.4	13.5	12.2
Age-standardised mortality rate [c d]	20.7 (19.1, 22.4)	12.4 (11.8, 13.0)	11.6 (10.9, 12.2)	12.2 (11.0, 13.2)	16.1 (14.1, 18.1)

[a] Percentage of missing data, [b] Adjusted for age, sex and smoking status, deprivation status and Charlson comorbidity index [c] with 95% CI, [d] deaths per 1,000 person-years. CCA: CCA

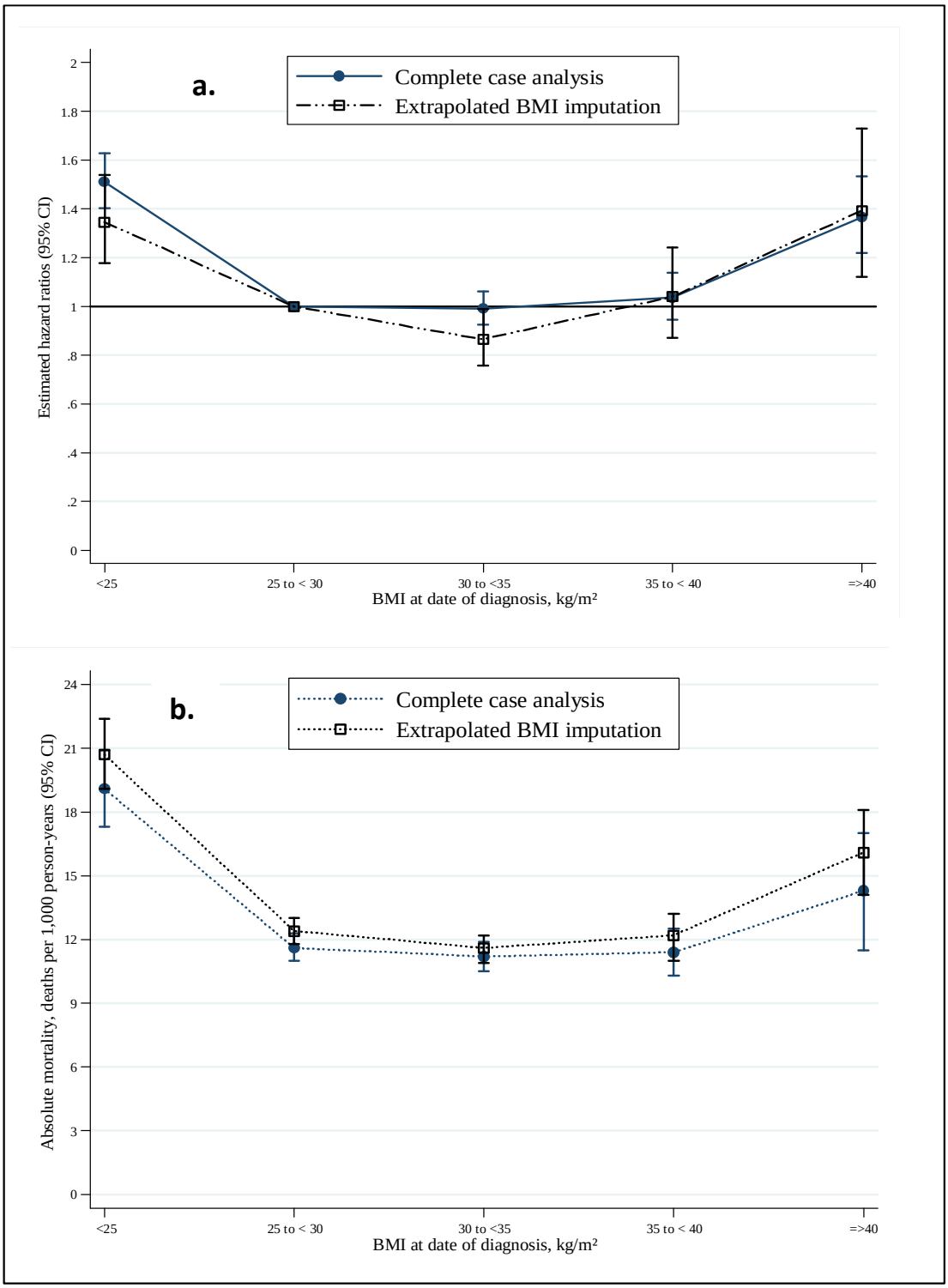

Figure 7.5: Comparison of a) hazard ratio estimates for risk of all-cause mortality and b) deaths per 1,000 person-years by categories of BMI after using CCA and extrapolated BMI imputation. Deaths in first two years of after diagnosis of type 2 DM are excluded. Reference category is 25 to less than 30kg/m² group in a.

Table 7.8: Agreement between categories of extrapolated BMI and BMI at diagnosis

		BMI at date of diagnosis (kg/m²)					Diagonal Total (% of 52,591)
		< 25	25 to < 30	30 to < 35	35 to < 40	≥ 40	
Extrapolated BMI (kg/m²)	< 25	4,856 (85.93)	759 (4.66)	7 (0.04)	2 (0.02)	0 (0.00)	-
	25 to < 30	768 (13.59)	14,401 (88.42)	1,943 (11.87)	16 (0.19)	1 (0.02)	0 (0.00)
	30 to < 35	17 (0.30)	1,116 (6.85)	13,734 (83.92)	1,553 (18.27)	22 (0.38)	3 (0.01)
	35 to < 40	5 (0.09)	7 (0.04)	665 (4.06)	6,606 (77.70)	897 (15.51)	45 (0.09)
	≥ 40	5 (0.09)	4 (0.02)	17 (0.10)	325 (3.82)	4,865 (84.10)	5,152 (9.80)
Diagonal Total (% of 52,591)		-	5 (0.01)	9 (0.02)	41 (0.08)	2,874 (5.46)	44,462 (84.64)

Upon categorisation of extrapolated BMI values and BMI at date of diagnosis levels, BMI groups were concordant for 85% of cases (Table 7.8). Where concordance was not achieved, the vast majority of patient's BMI categories differed by only a single BMI category, with 10% of patients having an extrapolated BMI that was lower than the observed BMI at diagnosis and 5% having a value that was above the observed value. On average, extrapolated BMI values were 0.26kg/m² lower than observed BMI values at diagnosis.

7.3.3.4 MULTIPLE IMPUTATION – INCLUDING EXTRAPOLATED BMI

Multiple imputation was used to create 60 imputed datasets. The relative and absolute mortality risks are presented in Table 7.9, and a comparison of estimates from CCA and multiple imputation are displayed in Figure 7.6. Estimates for relative and absolute mortality have been pooled across the 25 datasets. The observed patterns of mortality risk by BMI category are largely similar to those observed when closest imputation was applied. Estimates from multiple imputation including BMI measurements taken before and after diagnosis were identical to estimates obtained including extrapolated BMI and are displayed in Appendix Table A7.4. Unfortunately, only measurements taken at time-points t, t+1 t+2, t-1 and t-2 could be included in this imputation model due to issues with collinearity when further measurements outside this timeframe were incorporated into the multiple imputation model. The extent of the missing data in these variables is also presented in the Appendix Table A7.5.

Table 7.9: Estimates of relative hazard of all-cause mortality and age-standardised mortality by categories of BMI using multiple imputation (+extrapolated BMI).

	BMI category (kg/m²)				
	< 25	25 to < 30 (REF)	30 to < 35	35 to < 40	≥ 40
N	7,255 (11.17)	20,265 (31.19)	19,971 (30.74)	10,353 (15.93)	7,127 (10.97)
Increase in N from CCA (%)[a]	1,604 (12.96)	3,978 (32.13)	3,605 (29.12)	1,851 (14.95)	1,342 (10.84)
Multivariable adjusted HR (95% CI)[bc]	1.51 (1.41, 1.60)	1.00	0.96 (0.90, 1.02)	1.01 (0.93, 1.09)	1.33 (1.21, 1.47)
Standard Error	0.050		0.030	0.043	0.068
No. deaths	1,543	2,583	1,900	789	500
Person years	40,183	115,849	113,816	58,857	40,332
Crude mortality rate [d]	38.4	22.3	16.7	13.4	12.4
Age-standardised mortality rate[c d]	20.6 (18.5, 22.7)	12.5 (11.4, 13.6)	11.7 (10.6, 13.0)	12.0 (10.5, 13.5)	15.7 (12.9, 18.6)

[a] Percentage of missing data, [b] Adjusted for age, sex and smoking status, deprivation status and Charlson comorbidity index [c] with 95% CI, [d] deaths per 1,000 person-years. CCA: CCA

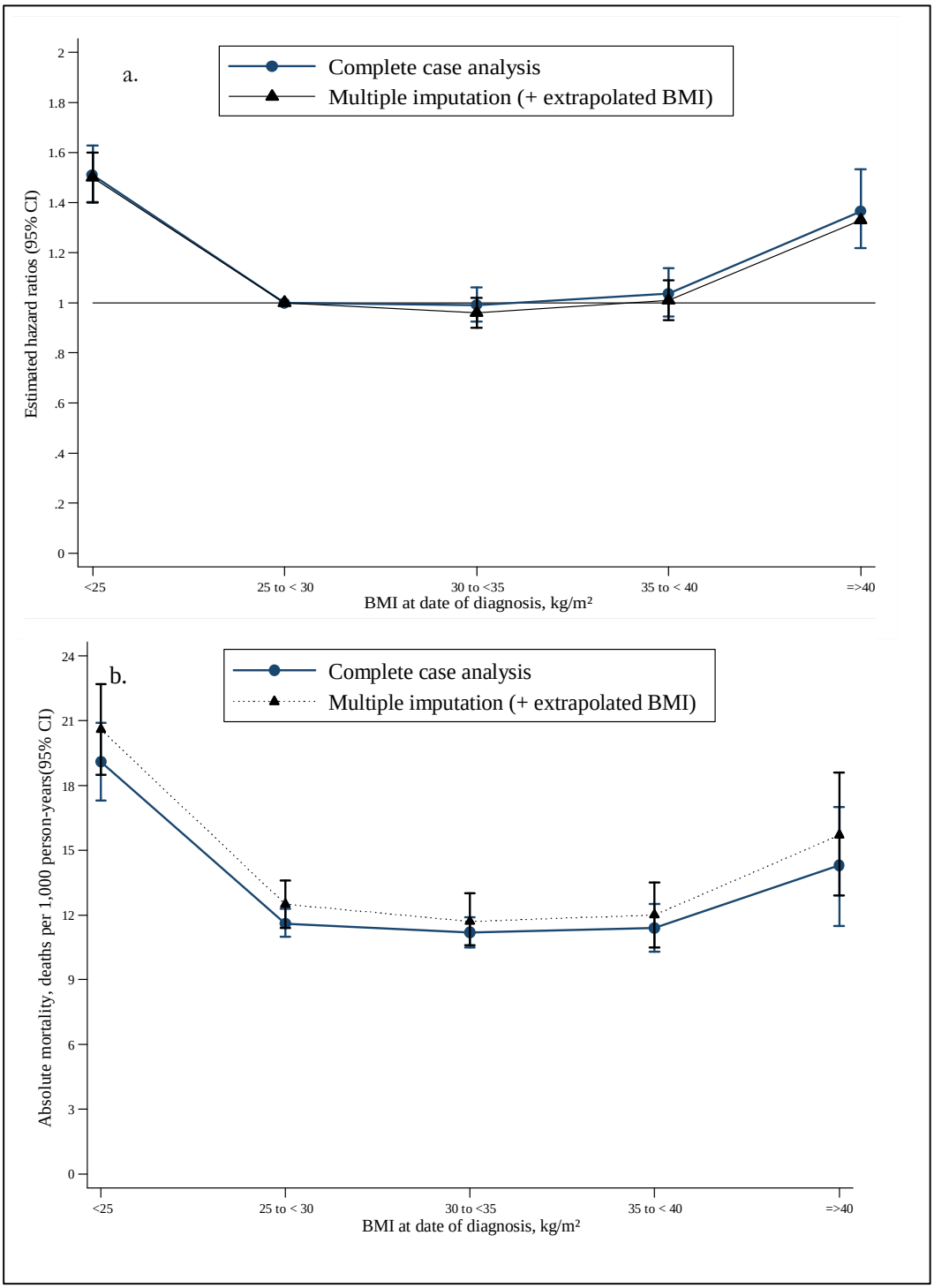

Figure 7.6: Comparison of a) hazard ratio estimates for risk of all-cause mortality and b) deaths per 1,000 person-years by categories of BMI after using CCA and multiple imputation. Deaths in first two years after diagnosis of type 2 DM are excluded. Reference category is 25 to less than 30kg/m² group in a).

7.3.3.5 TWO FOLD FULLY CONDITIONAL SPECIFICATION (FCS)

Two fold FCS was conducted to produce 25 imputed datasets. Minor problems with using this technique were quickly resolved upon corresponding with Catherine Welch, a member of the team which developed the two fold FCS tool in STATA. For example, unlike in the majority of other STATA commands the whole name of variables need to be specified to allow the model to run. Running the imputation model was time-consuming and complicated. It was necessary to find a balance between incorporating more predictors, whilst attempting to limit the introduction of additional missing data. Using two-fold FCS did not affect the overall observed relationship between BMI at diagnosis and all-cause mortality (Table 7.10 & Figure 7.7). A U-shaped relationship between patient BMI was observed with the lowest risk of mortality occurring in the 25 to <30kg/m² and 30 to <35kg/m² group. The relative risk of mortality in the highest and lowest BMI categories was again lowered using this multiple imputation approach.

Table 7.10: Estimates of relative hazard of all-cause mortality and age-standardised mortality by categories of BMI using two fold fully conditional specification

	BMI category (kg/m²)				
	< 25	25 to < 30 (REF)	30 to < 35	35 to < 40	≥ 40
N	7,199 (11.08)	20,402 (31.40)	20,142 (31.00)	10,248 (15.77)	6,980 (10.74)
Increase in N from CCA (%)[a]	1,548 (12.50)	4,115 (33.24)	3,776 (30.50)	1,746 (14.10)	1,195 (9.65)
Multivariable adjusted HR (95% CI)[bc]	1.37 (1.27, 1.47)	1.00	1.00 (0.93, 1.07)	1.06 (0.97, 1.16)	1.30 (1.16, 1.45)
Standard Error	0.052		0.035	0.050	0.076
No. deaths	1,520	2,597	1,943	748	507
Person years	40,943	116,201	114,008	58,351	39,510
Crude mortality rate[d]	37.1	22.3	17.0	12.9	12.8
Age-standardised mortality rate[c d]	18.7 (17.1, 20.3)	11.9 (11.3, 12.5)	11.8 (11.1, 12.5)	12.8 (11.6, 14.0)	15.5 (13.1, 17.9)

[a] Percentage of missing data, [b] Adjusted for age, sex and smoking status, deprivation status and Charlson comorbidity index [c] with 95% CI, [d] deaths per 1,000 person-years. CCA: CCA

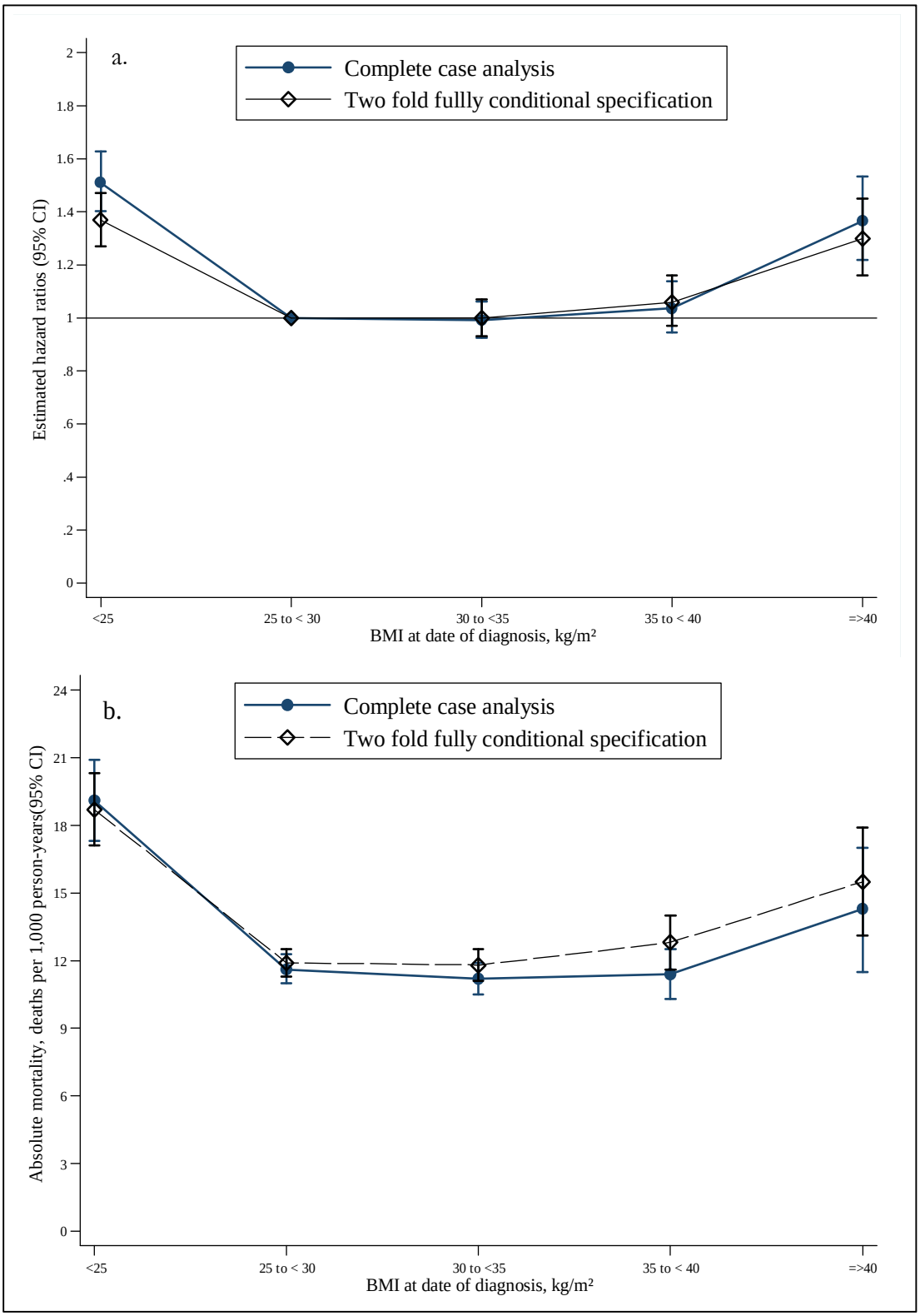

Figure 7.7: Comparison of a) hazard ratio estimates for risk of all-cause mortality and b) deaths per 1,000 person-years by categories of BMI after using CCA and two-fold fully-conditional specification.Deaths in first two years after diagnosis of type 2 DM are excluded. Reference category is 25 to less than 30kg/m² group in a.

7.4 DISCUSSION

7.4.1 INTRODUCTION

In the previous chapter, the poor predictive value of available characteristics at diagnosis was a key limitation of the performance of multiple imputation. Available characteristics were only able to account for around 15% of the variation in the BMI at diagnosis measurements. This problem is not unusual in epidemiological studies (van Buuren et al., 1999). Fortunately, routine data sources commonly have longitudinal records of important health indicators which may be used to improve the prediction of missing values. Nevertheless, irregular patterns of risk factor recording make the incorporation of these measurements into statistical analyses difficult. This chapter attempted to summarise the use of a number of approaches for utilising repeatedly measured data to improve imputations and to highlight the strengths and weaknesses of respective methods.

7.4.2 RELATIONSHIP BETWEEN BMI AND ALL-CAUSE MORTALITY

Across all methods for handling missing data, risk estimates for mortality across BMI groups remained fairly stable. Each approach produced a U-shaped relationship in mortality with the highest risk of mortality occurring in the lowest BMI category ($<25kg/m^2$). It is difficult to be certain about explanations for the stability in estimates. One reason may have been the relatively small amount of missing data occurring in the BMI at diagnosis variable (20%). Alternatively, there may well have been a lack of bias in the CCA approach.

Despite age-standardised mortality rates between patients with and without (17.8 deaths per 1,000 person years vs. 26.3 deaths per 1,000 person years) BMI at diagnosis data appearing considerably different, these differences did not affect the relative relationships between BMI and all-cause mortality. Each method indicated that patients with missing data exhibited a similar distribution across the categories of BMI to those patients with observed data. Of the patients with imputed data, there did not appear to be a clustering of the patients at the highest risk of mortality into a single BMI group. Indeed, relative estimates did not change and absolute estimates appeared to increase

fairly uniformly across all BMI categories after imputation regardless of the method for handling missing data.

7.4.3 PERFORMANCE OF MISSING DATA METHODS

7.4.3.1 CLOSEST OBSERVATION IMPUTATION

Closest observation imputation was the first method utilised to handle missing data in BMI. Relative estimates from this method were virtually identical to those observed using CCA, whilst absolute estimates were raised fairly uniformly across all BMI categories. Marginally larger increases in absolute mortality were observed in the highest and lowest than in other BMI categories. The relationship remained U-shaped and therefore it is unlikely that these small changes would alter overall inferences.

Of the imputed data, 90% of imputed values were derived from measurements taken after diagnosis with type 2 DM and 50% of these measurements were recorded 3 months after the defined diagnosis timeframe.

Closest observation imputation makes the assumption that closest observations accurately describe true BMI at diagnosis and therefore that BMI remains constant for considerable periods of time in some cases. This may not be a valid assumption, particularly where imputed data were derived from measurements recorded after diagnosis with DM. The diagnosis of type 2 DM is likely to trigger weight changes for two main reasons. Firstly, upon diagnosis with type 2 DM, current guidelines advise practitioners to offer individualised interventions, such as lifestyle changes and pharmacological therapies to encourage patient weight loss (Scottish Intercollegiate Guidelines Network, 2010). Secondly, the initiation of treatment with anti-hyperglycaemic agents is well known to affect body weight. For example, the association between insulin therapy and weight gain is well-established (UK Prospective Diabetes Study Group, 1998). Nonetheless, metformin treatment is more usually the first line pharmacological treatment, whilst insulin treatment is usually only required in the longer term (UK Prospective Diabetes Study Group (UKPDS), 1998). Metformin has been shown, in some studies to cause small, yet long term weight losses (Stumvoll et al., 1995; The Diabetes Prevention Program Research Group, 2012).

Observational studies that have examined weight changes in newly diagnosed patients with type 2 DM have had mixed findings. De Fine Olivarius et al. (2006) reported sustained weight loss in newly diagnosed patients in Denmark, up to 5 years after diagnosis date, though this was largely dependent on treatment type. Weight loss was highest in patients whose treatment was advice on diet, whilst patient weight remained stable upon initiation of treatment with metformin or sulfonylureas. Across all treatment groups, weight changes were greatest in the first five months following diagnosis. In another study, Tuthill et al. (2008) observed that 60% of patients with type 2 DM identified in a hospital database in Ireland exhibited either a decrease or no change in weight in six months following treatment initiation. However, both of these studies were conducted on relatively small samples (<1,500 participants) and while the Danish study used CCA to handle unobserved data for 17 participants, there was no description of the chosen method for handling missing data in Tuthill et al.'s study.

Another study observed that 53% of newly diagnosed type 2 DM patients in Sweden (n=8,486) had no change in their BMI (i.e. within 1 units) during the first 18 months after diagnosis (Bodegard et al., 2013). Thirty-two percent of patients had decreased BMI (defined as a fall of 1 or more units) and 14% had increased BMIs (gain of 1 or more units). Again, this study used CCA to handle unobserved data. Using trajectories to observe trends in BMI following diagnosis, Chiu et al. (2013) reported that 75% of participants in the Health and Retirement Study in the United States exhibited stable weight trajectories, whilst 18% of participants lost weight before regaining and 7% gained weight during 10 years of follow-up. The analyses in this study were conducted using maximum likelihood estimation and therefore were able to incorporate patients with missing data. Additionally, the authors reported results from those patients with CCA and the findings were similar. No studies were identified which provided data on weight change following diagnosis of DM from patients in the United Kingdom.

Amongst patients with BMI recordings at diagnosis and after diagnosis in our study, closest measurements recorded after diagnosis were generally consistent with measurements within the diagnosis time-frame. Generally, patients exhibited a slight decline in BMI (-0.37kg/m^2) after diagnosis, though these changes resulted in a change of BMI category in 23% of people.

The median time between the diagnosis time-frame and the next closest measurement was 91 days amongst patients with a BMI recording at diagnosis. Amongst patients with imputed BMI values, the time between diagnosis time-frame and closest BMI measurement was 88 days. Providing the trends in BMI were similar between patients with and without a BMI at diagnosis, the imputation of closest measurements taken after diagnosis (91% of imputed values) to reflect BMI at diagnosis could be expected to provide a reliable estimate of BMI around diagnosis of DM. However, this is an unverifiable assumption.

Similarly the accuracy of the nine percent of imputed values recorded prior to the diagnosis time-frame cannot be verified. Previous studies have indicated a steady rise in weight up until diagnosis of DM. One observational study has presented findings from incident type 2 patients indicating an average weight gain of 14.7kg from young adulthood until time of diagnosis (De Fine Olivarius et al., 2008). Overall, weight gain from 10 years prior to diagnosis was considerably smaller (+1kg), however significant weight loss was noted in the year immediately preceding diagnosis with DM. De Fine Olivarius et al. (2008) suggested that this may reflect the effect of hyperglycaemia and the resultant energy losses through glycosuria. Another study analysing data from 1965 to 2000 in Central Arizona reported that patient BMI steadily climbed between 0.43 and 0.71kg/m^2 per year before diagnosis but declined after diagnosis (Looker et al., 2001). However, using latent class trajectory analysis in a population of newly-diagnosed patients with diabetes in the Whitehall II study, Vistisen et al. (2014) found that 94% of patients could be categorised as 'stable overweight' prior to diagnosis with diabetes. These patients were characterised as having an average BMI in the overweight range for the period prior to diagnosis of diabetes (18 years), though exhibiting an average weight gain of 2.3 BMI units between baseline and end of follow-up. In a general population study, Bhaskaran et al. (2013) reported that previous BMI measurements (recorded within 3 years) were comparable to 'most recent' BMI values whilst measurements recorded more than 3 years prior to time of interest tended to underestimate later BMI values. This error was reduced by using a model-based imputation approach whereby a linear model for change in BMI over time was used to predict the most recent BMI. However, it is uncertain if these findings would be applicable to a population of type 2 DM patients.

In our data, the correlation between patient measurements taken before and at diagnosis (Pearson's Rank Correlation: 0.88) was weaker than the correlation between measurements taken after and at diagnosis (Pearson's Rank Correlation: 0.95). This weaker correlation may reflect the longer period of time at which measurements taken before diagnosis are available, rather than true biological differences.

Overall, in patients with a measurement taken at diagnosis, measurements recorded before and after diagnosis appear to be a reasonable estimate of BMI category at diagnosis. Unfortunately, whether these trends apply to patients without a BMI at diagnosis cannot be ascertained. It seems plausible that some practitioners may not have recorded patients whose weight was in agreement with previous recordings. In this scenario, closest observation imputation may produce unbiased estimates. Again though, this is an untestable theory using the data at hand.

In previous analyses, BMI at diagnosis was defined as a measurement recorded within 10 months prior to or 2 months after diagnosis. Here, this definition was expanded to include the closest BMI measurements and produced very similar results to those obtained from the CCA using the original definition of BMI at diagnosis. This work indicates that instead of using complicated modelling methods for handling unobserved repeatedly measured variables in routinely collected data sources, the expansion of the time-interval of interest may be sufficient to overcome problems associated with unobserved data. This approach to handling unobserved values is considerably more straight-forward and quicker to execute than more advanced approaches. Future work using routine data sources could investigate whether this approach produces valid estimates that are similar to those obtained using complicated methods.

Closest observation imputation is largely similar to the traditionally used last observation carried forward (LOCF) approach. LOCF itself was not feasible in this dataset where some patients did not have BMI measurements prior to the diagnosis timeframe. However, LOCF and its derivatives have been widely criticised, especially in the clinical trial literature where dropout is likely to be related to the outcome or the assigned treatment group. In this dataset, missingness in BMI was strongly associated with the outcome of mortality and unintentional weight loss often precedes death. Depending on the presence of chronic diseases, weight loss prior to death can occur a considerable

time before death. For example, weight loss in patients with cancer has been shown to accelerate significantly three years before death, regardless of age (Alley et al., 2010). This creates two problems. Firstly, using closest observation imputation may not have been valid if these patients had other chronic conditions as these conditions may have contributed to their observed weight taken outside the diagnosis timeframe. Comorbidity status was assessed at diagnosis and therefore may not reflect comorbidity status at BMI measurement date. Secondly, the exclusion of patients who had died within two years of diagnosis of type 2 DM may not have adequately controlled against reverse causation.

A further criticism of this single imputation approach is its inability to take into account the uncertainty of the imputed data (Little & Rubin, 1987). Accordingly, closest observation imputation is likely to have produced standard errors which were too small and confidence intervals that were too narrow.

7.4.3.2 *EXTRAPOLATED BMI IMPUTATION*

Absolute estimates of the association between BMI at diagnosis and all-cause mortality obtained using extrapolated BMI imputation were almost identical to those obtained using CCA and closest observation imputation. The similarity between estimates using the two imputation methods is unsurprising given that assigned BMI categories from the two methods were in agreement for 10,423 (84% of 12,380) individuals. However 1,302 of these were the same since only a single BMI measurement was available. Relative estimates however indicated that a higher proportion of relatively healthier patients were placed in the 30 to <35kg/m^2 group using extrapolated BMI imputation.

Estimates from extrapolated BMI imputation are likely to be more reliable than those obtained from closest observation imputation. Imputes using closest observations were generally available after diagnosis. Since patients with missing data had worse survival, the date of determination of their closest BMIs was likely to be closer to their date of death. This therefore increased the likely influence of reverse causation and therefore may have accounted for the increased risk of death in the <25kg/m^2 category when using closest observation imputation.

Nonetheless, the validity of using extrapolated BMI imputation to handle missing BMI data is questionable for a number of reasons. Firstly, as previously mentioned, some

individuals only had a single BMI measurement from which to calculate a BMI at diagnosis. This meant that the predicted BMI at diagnosis was equal to the available measurement, regardless of time from diagnosis. Secondly, the use of this technique did not allow for within subject variability. Thirdly, this method led to the production of implausible BMI measurements (above 75kg/m^2 or below 15kg/m^2) where recordings were available a long time prior to or after diagnosis (See section 4.3.3.3 for more detail). These individuals were included in the highest and lowest BMI categories to avoid the exclusion of further patients. Interpolation of missing values, whereby measurements taken before and after are used to predict the unobserved measurement may have helped to overcome this problem. Unfortunately, this approach was not feasible in this dataset where not every patient had measurements available before and after the diagnosis timeframe. Finally, amongst individuals with few measurements, extrapolated BMI values were likely to have been heavily affected by BMI outliers.

Nonetheless, according to findings where extrapolated BMI and observed BMI at diagnosis measurements were compared, the performance of extrapolated BMI was good. Categories of extrapolated BMIs and observed BMIs at diagnosis were in agreement 87% of the time, and the mean difference was equal to only 0.24kg/m^2. This difference indicates that extrapolated BMI underestimated BMI at diagnosis, though this difference was small enough that it would rarely result in a change of BMI category. Unfortunately, these findings may not be representative of the performance of this measure amongst patients with missing BMI data.

Furthermore, this is another single imputation approach and therefore the uncertainty in imputations was not reflected in the calculated standard errors.

7.4.3.3 MULTIPLE IMPUTATION APPROACHES

A number of options for incorporating longitudinal data into multiple imputation are available. One approach involves the inclusion of a derived summary measure of the available longitudinal measurements. An alternative approach involves the inclusion of measurements recorded in a number of distinct time-blocks into a single multiple imputation model. Most recently, the two-fold fully conditional specification (FCS) multiple imputation approach has been proposed to handle longitudinal missing data (Nevalainen et al., 2009).

Derivatives of these three approaches were applied and were each found to result in very similar estimates. In the first approach, a multiple imputation model including extrapolated BMI was used. In the second approach, measurements recorded before and after the diagnosis time-frame were included in the imputation model. Unfortunately, this approach was not able to account for the temporal ordering of the data and therefore gave equal weight to BMI measurements regardless of the timing of the recording. This assumption is unlikely to be correct. Unfortunately, collinearity issues prevented the inclusion of further BMI measurements. These issues have been observed in previous studies incorporating longitudinal data in standard multiple imputation approaches (Welch et al., 2014). In their simulation study, Welch et al. (2014) reported that their imputation model which conditioned on weight measurements from numerous time-blocks failed to run successfully in 25% of datasets due to collinearity problems.

An additional problem for this technique, and indeed the two-fold FCS approach is the need to arbitrarily define the width of the time intervals. The specification of the time-window width is a balance between available measurements within time-periods and the limiting of missing data in each window. For example, ideally the width of the time blocks would be small in order to incorporate a number of measurements that are likely to be highly correlated to the value of interest. However this would result in large proportions of missing data within time-windows. It is possible that estimates could change depending on the width of the time-interval. Time constraints restricted the further exploration of this issue, though it does appear that estimates have been fairly robust to changes in the specification of the imputation model.

The final approach applied was the two-fold FCS method. Again, this method produced similar estimates for the relationship between BMI at diagnosis and all-cause mortality, despite being considerably more time-consuming.

This approach is comparatively new and has only recently became available as a STATA command (Welch et al., In print 2014). Due to the relative newness of this approach, it has yet to be evaluated in a wide range of settings. Despite this, Welch et al. (2014) were able to present largely unbiased estimates using two-fold FCS in a simulation study using a time-to-event model. In their study, the greatest gain in efficiency through using

this approach was observed when imputing weight compared to the imputation of BP and smoking status measurements. This was probably due to the high within-patient correlation between weight measurements. However, these analyses were performed on generated data which had MCAR data and therefore this approach may yield biased estimates when data are MAR. Accordingly the use of this approach requires further investigation in a wider range of settings.

Relative to the other multiple imputation approaches, two fold FCS was considerably more computationally intensive, taking a considerably longer time to complete the imputation process than previous multiple imputation methods. This problem is likely to increase in larger, more complicated models (Nevalainen et al., 2009).

Despite this, it is clear that this technique has some very attractive features which may ensure its widespread use in the future. Indeed, it is likely that this approach might be used to improve investigations describing patterns of change in BMI after diagnosis of DM in Scotland. Current research investigating weight changes after diagnosis with type 2 DM and their effect on metabolic and cardiovascular outcomes has been hindered by the presence of missing data at different time-points (Aucott et al., 2013). Using CCA, the researchers have identified that the majority of patients (>50%) of patients exhibited a stable pattern of weight (-2.5% to 2.5% weight change) in the first year following diagnosis with DM. Unfortunately, these analyses included only 29,316 cases out of 81,990 cases due to the presence of missing data. Two-fold FCS could therefore be more useful in imputing missing values at these different time-points, rather than being limited to imputing BMI at a single time-point.

Each multiple imputation approach had the benefit of overcoming the problem associated with underestimated standard errors. Nonetheless they produced largely comparable results to single imputation approaches, despite taking a considerably time longer to implement.

7.4.4 LIMITATIONS

This investigation of available approaches for incorporating longitudinal measurements to improve imputations had a number of potential limitations. Firstly, patients diagnosed in 2008 may have fewer measurements as BMI readings were not available

after 2008 at the time of writing up. Therefore these patients typically will have had less information from which to estimate missing values with. However, the numbers of patients diagnosed in 2008 were low since data extraction occurred in June 2008, and amongst these patients BMI at diagnosis completeness was higher than observed in earlier years (88%). Secondly, these approaches heavily depend on the definition of cut-off points that may or may not be correct. For example, the width and definition of the time-window for measurements of BMI at diagnosis was arbitrarily defined and may not truly reflect BMI at diagnosis. It may well have been beneficial to explore how much estimates for the association between BMI and mortality changed by altering the definition of BMI at diagnosis. This could certainly be investigated in the future. Similarly, the width of the time-blocks for the BMI measurements was also arbitrarily defined. Unfortunately the width of the time-blocks needed to be the same in two-fold FCS whether the measurements were recorded before or after diagnosis. This is despite there being considerable more available measurements after diagnosis during shorter time-frames. However, using smaller time-frames would have resulted in considerably more missing data in the BMI variables recorded before diagnosis (t-3, t-2, t-1).

Adjustment for hyperglycaemic treatment is a further limiting factor of these analyses. BMI measurements recorded after diagnosis were likely to have been affected by hyperglycaemic therapy (Yki-Jarvinen, 2001; Aucott et al., 2004; Wing et al., 2013).

7.4.5 CONCLUSIONS

Incorporating longitudinal data into analyses with missing data has the potential to increase the precision of imputed values whilst also reducing the risk of MNAR. Unfortunately, this process is not straightforward using routine data where data tends to have intermittent patterns of missing data.

Several missing data methods which were able to incorporate longitudinal data were applied. Despite each respective method making different assumptions, estimates of the association between BMI and mortality were robust. Though the absolute risk of mortality increased across all BMI categories upon inclusion of patients with missing data, it is unlikely that these increases would affect overall inferences of the strength of relative associations. In this situation, using more advanced but more time-consuming

methods for handling unobserved data did not provide any additional benefit. However, these results may not be generalisable to different settings with different patterns of missing data and mechanisms of missingness.

8. MISSING DATA IN MULTIPLE VARIABLES

8.1 INTRODUCTION

In previous sections, the pattern of missingness was univariate with only one variable, BMI at diagnosis containing missing data. Patients with missing data on any other variable to be included in either the imputation model or the model of substantive interest were excluded. However, missing data are rarely clustered in a single variable in epidemiological research. More usually, missing data are present in multiple variables and exhibit a non-monotone pattern of missingness, whereby the configuration of missing data are in no clear pattern. Additionally, it is not uncommon for the form of the analysis model to be unknown, particularly which variables should be included and their functional form. In previous chapters, only variables that have been established as important confounders of the relationship between BMI and mortality have been included in survival models in an effort to simplify the issues. However, it is probable that other available variables may confound this relationship and ought to be included in the analysis model to prevent biased estimates. Typically, variable selection for unestablished analysis models is implemented using backwards, forwards or stepwise selection among the complete cases. Unfortunately, these methods are impractical in this dataset for two reasons. Firstly, the sample size is so large that the presence of even very small differences away from the null lead to highly significant p-values ($p<0.000$). Secondly, the identification of predictors in the presence of missing data may lead to the selection of unimportant variables or the failure to select important variables due to biased parameter estimates when data are not MCAR. Fortunately, the relationship investigated here has been explored extensively and therefore it was possible to identify appropriate confounders using these studies rather than data exploration exercises.

When researchers are confronted with a missing data problem, they are faced with a number of decisions regarding the method for handling missing data. If the researcher is concerned that CCA may result in biased estimates or reduce the sample size to an unacceptable level then alternative methods for handling missing data are required. Due to its accessibility in popular statistical software, multiple imputation is often used when data are thought to be MAR (Horton & Kleinman, 2007). There are a number of multiple imputation models to choose from including Schafers (1997) multivariate

normal model (MVN), the imputation using chained equations (MICE) method first implemented by van Buuren et al. (1999) and predictive mean matching (PMM). Each of these approaches has been recommended for use in the presence of non-monotone missing data. These approaches rest on different theoretical assumptions yet it remains unclear how these differences may affect a study's overall findings.

The following chapter presents a comparison of results obtained from the three multiple imputation models mentioned above. A comparison of the issues raised by using each of these methods is presented. Unlike in previous sections, this section will demonstrate the use of these techniques in a considerably more realistic setting whereby an investigation into the association between BMI and mortality is complicated by the presence of multiple incomplete variables. Two separate analysis models will be presented. The first will be adjusted for well-known confounders, whilst the second will be further adjusted for additional confounders that have been included in analysis models in previous studies.

8.2 METHODS

8.2.1 DATA

The data used will be the cohort of patients who were diagnosed with type 2 DM between 2004 and 2008 and were alive two years after diagnosis (n=73,890).

In previous sections, biochemical measurements observed closest to diagnosis date were used to describe patient characteristic at diagnosis. In this section, to ensure measurements accurately reflect biochemical measurements at diagnosis, recordings taken within 1 year of diagnosis (10 months prior to or 2 months after) were used.

Changes to the definition of biochemical measurements at diagnosis were applied to: smoking status, systolic BP, diastolic BP, glycated haemoglobin, creatinine, total cholesterol, HDL-cholesterol and triglycerides.

Additionally, smoking status at diagnosis was divided into three groups; current, former and never smokers. In previous sections smoking had been grouped into two distinct groups: never smoker vs. ever smoker. This dichotomisation may have biased adjusted

estimates for the association between BMI at diagnosis and all-cause mortality. Weight gain is known to be strongly associated with smoking cessation (Williamson et al., 1991). Grouping together former and current smokers together may therefore have diluted the relationship between BMI and smoking status.

8.2.2 DATA AND EXTENT OF MISSINGNESS

Amongst the cohort of patients diagnosed with DM between 2004 and 2008, the extent of and patterns of completeness in variables to be included in the models of substantive interest were observed.

8.2.3 PREDICTORS OF MISSINGNESS

To ascertain whether patients with missing data were a random subsample of the full cohort, a comparison of demographic and biochemical characteristics between patients with and without complete data was conducted. To achieve this, a missing indicator variable was generated to indicate presence of missing data in any variable to be included in the final model. Using this variable, mean and percentage differences between patients with and without complete data for all variables were observed. Logistic regression analyses were used to identify characteristics that were predictive of missing data. The presence of multiplicative interactions was investigated graphically and by the inclusion of interaction terms in logistic regression models. To reduce the risk of identifying spurious interaction effects, only those interactions considered plausible were investigated. These are listed below:

a. Age at diagnosis vs. Sex

b. Age at diagnosis vs. Deprivation status

c. Deprivation status vs. Sex

d. Charlson comorbidity index vs. Age at diagnosis

e. Charlson comorbidity index vs. Vital status

Cox proportional hazards models were used to identify differences in survival between patients with and without complete data. Absolute mortality rates (deaths/1,000

person-years) by missing data indicator were also produced. This process was then repeated for each incomplete variable individually.

8.2.4 Analysis Model

Cox proportional hazards models were used to assess the association between categories of BMI and all-cause mortality. In Model I, estimates were adjusted for age, sex, deprivation status, Charlson comorbidity index and smoking status. In Model II, estimates were adjusted for, in addition to those variables listed above, systolic BP, glycated haemoglobin, total cholesterol and HDL-cholesterol. Due to the large sample size, typical methods for predictor selection such as backward selection would lead to the inclusion of unimportant predictor variables. These additional variables included in Model II were chosen on the basis of having the propensity to confound the observed relationship and have regularly been included in adjusted models in previous studies. The proportional hazards assumption was examined using log minus log plots. Hypothesis tests for proportionality were not used due to the large sample size. HRs and 95% confidence intervals are presented. Estimates were stratified according to smoking status to assess potential residual confounding and by Charlson comorbidity index (No comorbidity vs. Comorbidity) to reduce the potential for reverse causation. Effect modification was investigated by the inclusion of the multiplicative interaction term between BMI at diagnosis and the effect modifier in the main effects model. The proposed effect modifiers were age at diagnosis (<65 vs. ≥ 65years) and smoking status.

8.2.5 Missing data methods

CCA were conducted to serve as comparison estimates. Three separate multiple imputation approaches were then applied; MICE, multivariate normal model and predictive mean matching. The theoretical underpinnings of these techniques are described in Section 2.3.

The imputation of smoking status posed unique challenges. The variable smoking status cannot transition from either being a current smoker or an ex-smoker to being a never smoker. This feature of the categorical variable was difficult to accommodate in missing data methods. To overcome this problem in part, patients with more than one mention of being a never smoker prior to or after diagnosis were assumed to be never smokers at

diagnosis. This was deemed a reasonable assumption given that smoking initiation typically occurs in adolescents and young adults, whilst levels of initiation are low amongst people over 30 years of age (Oh et al., 2010). In this sample, the mean age at diagnosis was 62 years. In addition, for patients who were missing a smoking status at diagnosis but had consistent smoking status records before and after diagnosis, it was assumed that their smoking status did not change during this time.

8.2.5.1 MULTIPLE IMPUTATION USING CHAINED EQUATIONS

STATA'S *ice* command was used to generate imputed datasets. This iterative procedure begins by replacing missing values with randomly selected observed values. It then imputes the variable with the least amount of missing data first using the appropriate imputation model (linear for continuous, logistic for binary, multinomial logistic regression for categorical), regressing on the other variables in the imputation model. The variable with the second least missing data is imputed next. This continues until all variables with missing data have had their values imputed. This concludes a cycle and in order to produce one multiply imputed dataset, the process is iterated for the pre-selected number of cycles.

Rubin (1996) recommends that imputation models should be at least as rich as the analysis model. Accordingly, an imputation model was developed with analysis Model II in mind. Model II is adjusted for age at diagnosis, sex, smoking status, deprivation status, Charlson Comorbidity Index, smoking status, systolic blood pressure, glycated haemoglobin, total cholesterol and HDL-cholesterol. The imputation model therefore included variables which were predictive of missing data and also variables that were predictive of missing values themselves.

In chapter 7, it was observed that the inclusion of repeated variables of longitudinal may improve the predictive power of imputation models. For this reason, new variables were generated which represented closest measurement of the variable recorded after the diagnosis date. The following variables were therefore included in the imputation model:

- Age at diagnosis (continuous)
- Sex (binary)
- BMI at date of diagnosis (ordinal)
- Closest BMI recorded after diagnosis (ordinal)

- Deprivation status (ordinal)
- Vital status (binary)
- Follow-up time (log transformed continuous)
- Smoking status at diagnosis (categorical)
- Charlson comorbidity index (binary – [Comorbidities Y/N])
- Year of diagnosis (ordinal)
- Health board area (categorical)
- Closest smoking status recorded after diagnosis (categorical)
- Systolic BP at diagnosis (continuous)
- Closest systolic BP recorded after diagnosis (continuous)
- Glycated haemoglobin at diagnosis (log-transformed continuous)
- Closest glycated haemoglobin recorded after diagnosis (log-transformed continuous)
- Total Cholesterol at diagnosis (continuous)
- Closest Total Cholesterol recorded after diagnosis (continuous)
- HDL-cholesterol at diagnosis (continuous)
- Closest HDL-cholesterol recorded after diagnosis (continuous)

In addition, the interaction terms investigated in the analysis model were included in the imputation model. This approach has been proposed to result in smaller downward bias than simply generating interaction terms from imputed values, i.e. passive imputation (Von Hippel, 2009).

Evidence of non-normality was observed for follow-up time and glycated haemoglobin variables. Failure to correctly handle incomplete non-normal variables would result in the distribution of imputed values differing from the distribution of the observed values. These variables were therefore log-transformed and imputed values were back-transformed to the original scale.

Since missingness occurred in both smoking status at diagnosis and closest smoking status after diagnosis variables it was necessary to set rules to ensure patients who were smokers at diagnosis could not later become never smokers. To achieve this, the *ice* option *conditional* was specified. The *interval* command was used to ensure imputes for continuous variables remained within plausible bounds, i.e non-negative for biochemical measurements.

White and Royston (2009) recommend the inclusion of the Nelson-Aalen estimator of the cumulative baseline hazard rather than follow-up time. Unfortunately, this was not

feasible since the follow-up time variable was incomplete. In chapter 5 it was shown that the difference between estimates when using the Nelson-Aalen estimator and when using follow-up time in the imputation model was negligible in this large sample. For this reason, follow-up time was included in its basic form.

To ensure convergence was achieved, the number of cycles was set to 20. This was more than the default 10 in STATA since there were a number of variables which were highly associated (e.g. BMI at diagnosis and closest BMI recorded after diagnosis). When this is the case, White et al. (2011) recommend increasing the number of cycles. The *trace* command was used to observe the mean of imputed values after each cycle and these were observed to stabilise after between 10 and 20 cycles.

As recommended by Bodner (2008), the number of imputations was chosen based on the fraction of incomplete cases. Since analysis Model II includes only 40% complete cases, 60 imputed datasets were created.

8.2.5.2 PREDICTIVE-MEAN MATCHING (PMM)

PMM was used to impute missing values of continuous variables, whilst the regression models to impute missing values in ordinal, categorical and ordinal variables remained the same. The imputation model was specified as above using STATA's *ice* command with the inclusion of the *pmm* option. Non-normal variables were not transformed since PMM is a suggested means of overcoming issues with non-Normality. Despite the default match-pool being 3 in *ice*, this was set to 10 in accordance with recent recommendations (Morris et al., 2014). Sixty imputed datasets were created.

8.2.5.3 MULTIVARIATE NORMAL (MVN) MULTIPLE IMPUTATION

STATA'S *mi impute mvn* was used to implement the multivariate normal model. The process of generating imputations using the MVN model is described in detail in Chapter 2. Briefly however, the MVN normal model uses an iterative Markov chain Monte Carlo (MCMC) method to impute missing values and assumes that all variables in the model jointly follow a multivariate distribution. There are two main phases of this model. The first phase involves obtaining initial values using the Expectation maximisation method, and the second performs the MCMC procedure to obtain imputations.

The imputation model was specified as above with the inclusion of the same variables. Since this approach assumes normality of continuous variables, non-normal variables were log-transformed. In this setting, binary and ordinal variables were imputed as dummy indicator variables and rounded to 0 or 1 based on the value of the other indicator variables. The number of iterations for the MCMC to converge was set to 200 (default is 100) due to the high fraction of missing information. Convergence was assessed using trace plots of the MCMC parameter estimates. Sixty imputed datasets were created.

8.3 RESULTS

8.3.1 EXTENT OF MISSINGNESS

Of the 73,890 patients, 52,980 patients were complete cases for Model I and 27,706 patients were complete cases for Model II (Table 8.1). For patients with missing data for Model I, there were 5,099 patients who were without both values for BMI at diagnosis and smoking status, suggesting some clustering of missing values amongst patients. Three percent of patients were missing five variables, though these variables were those which required a valid date of diagnosis in their construction.

For model II, there were 11,107 patients who were only without an HDL-cholesterol value alone. A further 4,430 patients were missing smoking status, BMI, systolic BP, glycated haemoglobin, total cholesterol and HDL-cholesterol values.

Table 8.1: Completeness of variables included in models of substantive interest (n=73,890)

Variables to be included in substantive model	Observed (%)
Age at diagnosis	71,405 (96.64)
Sex	73,834 (99.92)
Deprivation status	73,365 (99.29)
Vital status	73,890 (100.00)
Follow-up time	71,414 (96.65)
Smoking status at diagnosis	65,215 (88.26)
Charlson comorbidity index	71,414 (96.65)
BMI at diagnosis	54,534 (73.80)
Systolic BP	62,674 (84.82)
Glycated haemoglobin	49,992 (67.66)
Total cholesterol	58,829 (79.62)
HDL-cholesterol	39,785 (53.84)

MODEL I: Overall completeness: 52,980 (71.70%)

MODEL II: Overall missingness: 27,706 (37.25%)

8.3.2 PREDICTORS OF MISSINGNESS

8.3.2.1 ALL VARIABLES TO BE INCLUDED IN MODEL I

Higher levels of comorbidity were associated with a lower probability of having complete data for any variable to be included in Model I (Comorbidities vs. no comorbidity: mv OR 1.08 [1.03, 1.12]) (Appendix Table A8.1). Similarly, dying before the end of follow up was associated with having missing data (Deceased vs. Alive: mv OR 1.25 [1.13, 1.39]). In general, there was no evidence to indicate any biochemical measurements were associated with the likelihood of patients being complete cases. Higher levels of total cholesterol were marginally protective against having missing data. When biochemical measurements were redefined to include only measurements taken 10 months prior to or 2 months after diagnosis with DM, there remained no evidence that these biochemical measurements were associated with the likelihood of being a complete case (Appendix Table A8.2). HR estimates indicated that poorer survival was predictive of having incomplete data (mv HR missing vs. complete: 1.27 [1.22, 1.33]) (Appendix Table A8.3). Since missingness was dependent on the outcome of the analysis model, a CCA was likely to produce biased estimates.

Divergent trends in the risk of missing data in variables to be included in Model I across Charlson comorbidity indices were observed by vital status at study endpoint. Among

patients with high levels of comorbid conditions, risk of missing data was highest amongst patients that died before the study endpoint than among survivors (Figure 8.1). This may reflect considerably poorer health statuses of patients with more comorbid conditions who later died compared with patients with more comorbid conditions who were still alive at study-endpoint. Accordingly, the recording of clinically unimportant variables may not have been prioritised in these extremely unwell patients.

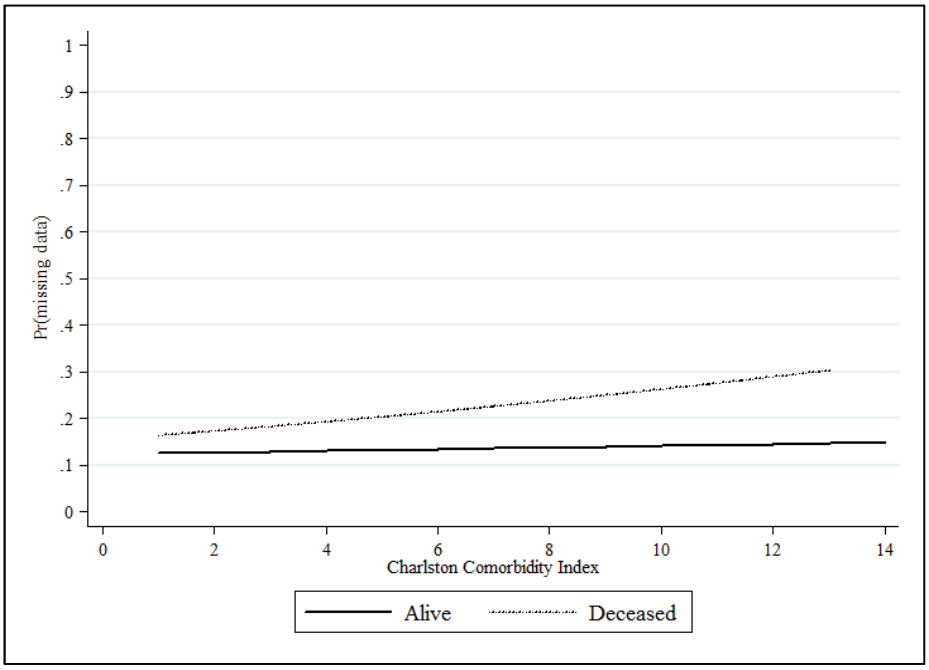

Figure 8.1: Probability of having missing data by Charlson Comorbidity index. Stratified by vital status.

8.3.2.2 *All variables to be included in Model II*

Logistic regressions to explore any differences between patients with (n= 27,706) and without (n=46,184) complete data identified similar trends in missingness to those observed for missingness in Model I variables (Appendix Table A8.4 & A8.5). As before, better survival was associated with complete cases, as was the absence of comorbidities. Interestingly, evidence pointed towards a relationship between year of diagnosis and the probability of being a complete case (2008 vs. 2004 mv OR: 0.42 [0.36, 0.49]); an association that was not observed amongst complete cases for Model I. Since complete cases in this analysis included patients who had complete data for additional biochemical variables such as total cholesterol, this trend may reflect an increased emphasis on recording these variables between 2004 and 2008. This may account for the absence of this trend amongst complete cases in Model I whose

completeness of total cholesterol is not considered. Lower glycated haemoglobin levels at diagnosis were associated with a higher probability of having missing data (mv OR: 0.92 [0.91, 0.93]). No evidence of any multiplicative interaction was observed.

8.3.2.3 *IN INDIVIDUAL VARIABLES (INCLUDED IN MODEL I AND/OR MODEL II)*

Next, predictors of missingness in each individual variable with incomplete data were observed (Appendix tables A8.7, A8.9, A8.11, A8.13, A8.15, A8.17). The trends in missingness by demographic and clinical characteristics for each individual variable with incomplete data were largely similar. For example, vital status, survival (Appendix tables A8.8, A8.10, A8.12, A8.14, A8.16, A8.19) and HBA were all strongly associated with missingness in each individual variable. Being male was associated with lower levels of missing data in each distinct variables but the relationship was strongest for missingness in BMI (male vs. female mv OR: 0.89 [0.84, 0.93]) and total cholesterol (male vs. female mv OR: 0.86 [0.81, 0.92]). Additionally, the probability of having missing smoking status, total cholesterol and HDL-cholesterol data increased with earlier calendar year of diagnosis, while a U-shaped relationship between year of diagnosis and missingness in BMI and systolic BP was also observed. Longer lengths of follow-up were associated with a higher probability of having unobserved data amongst the glycated haemoglobin, total cholesterol and HDL-cholesterol variables only. Meanwhile, the probability of having incomplete data in the BMI at diagnosis, smoking status and systolic BP variables were highest amongst patients with comorbidities.

8.3.3 COMPLETE CASE ANALYSIS (CCA)

CCA resulted in the exclusion of 20,910 patients for Model I (4,798 deaths and 104,560 person-years, crude mortality: 45.9 deaths per 1,000 person-years) and 46,184 patients in Model II (7,582 deaths and 250,564 person-years, crude mortality: 30.3 deaths per 1,000 person-years). Therefore there were 52,980 patients included in Model I analyses (5,484 deaths and 300,592 person-years, crude mortality: 18.2 deaths per 1,000 person-years) and 27,706 patients included in Model II analyses (2,700 deaths and 154,589 person years, crude mortality: 17.5 deaths per 1,000 person-years).

Estimates for the association between BMI and all-cause mortality are presented in Figures 8.2, Figures 8.3 and Table 8.2. In both Model I and Model II, a U-shaped

relationship was observed with the highest risk of mortality being observed in the lowest and highest BMI categories. The lowest risk of mortality was observed amongst obese patients (30 to <35kg/m^2).

Results were stratified by smoking status to assess for residual confounding (Table 8.3). Estimates were similar for current, ex- and never smokers in Model I. However, relative mortality in the >40kg/m^2 group was higher amongst non-smokers mv OR: 1.59 [1.31, 1.93]) than current smokers (mv OR 1.31 [1.05, 1.65]). In an effort to control for the potential effects of reverse causation by chronic diseases, results were stratified by presence of comorbidities (Figure 8.2 - panels E & F) (Table 8.3). Interestingly, compared with patients with comorbidities, patients without comorbidities had higher relative risk of mortality in the lower BMI categories. No interaction by age at diagnosis was observed. Estimates obtained from Model II (adjusted for age at diagnosis, sex, smoking status, deprivation status, Charlson comorbidity index, systolic BP, glycated haemoglobin, total cholesterol and glycated haemoglobin) were calculated using considerably fewer patients. Differences between estimates from Model I and Model II may therefore be due to these large changes in sample size, rather than the effect of the addition of further potential confounders.

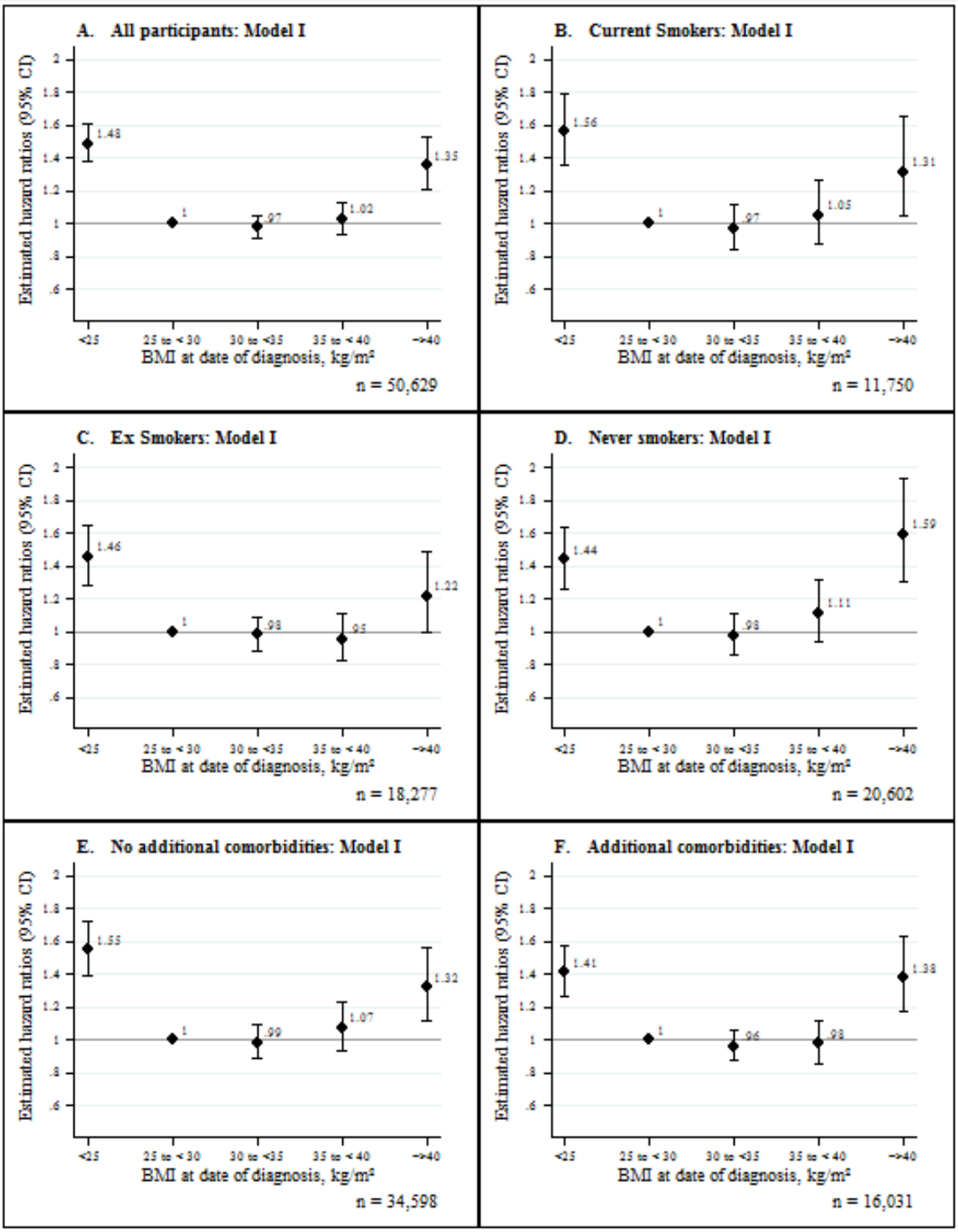

Figure 8.2: Model I estimates for the relationship between categories of BMI and all-cause mortality with stratification. Results for all participants are provided in Panel A. Relative estimates for current smokers are provided in Panel B, for Ex-smokers in Panel C and Never smokers in Panel D. Panel E presents estimates for patients with diabetes only and Panel F presents results for patients with comorbidities.

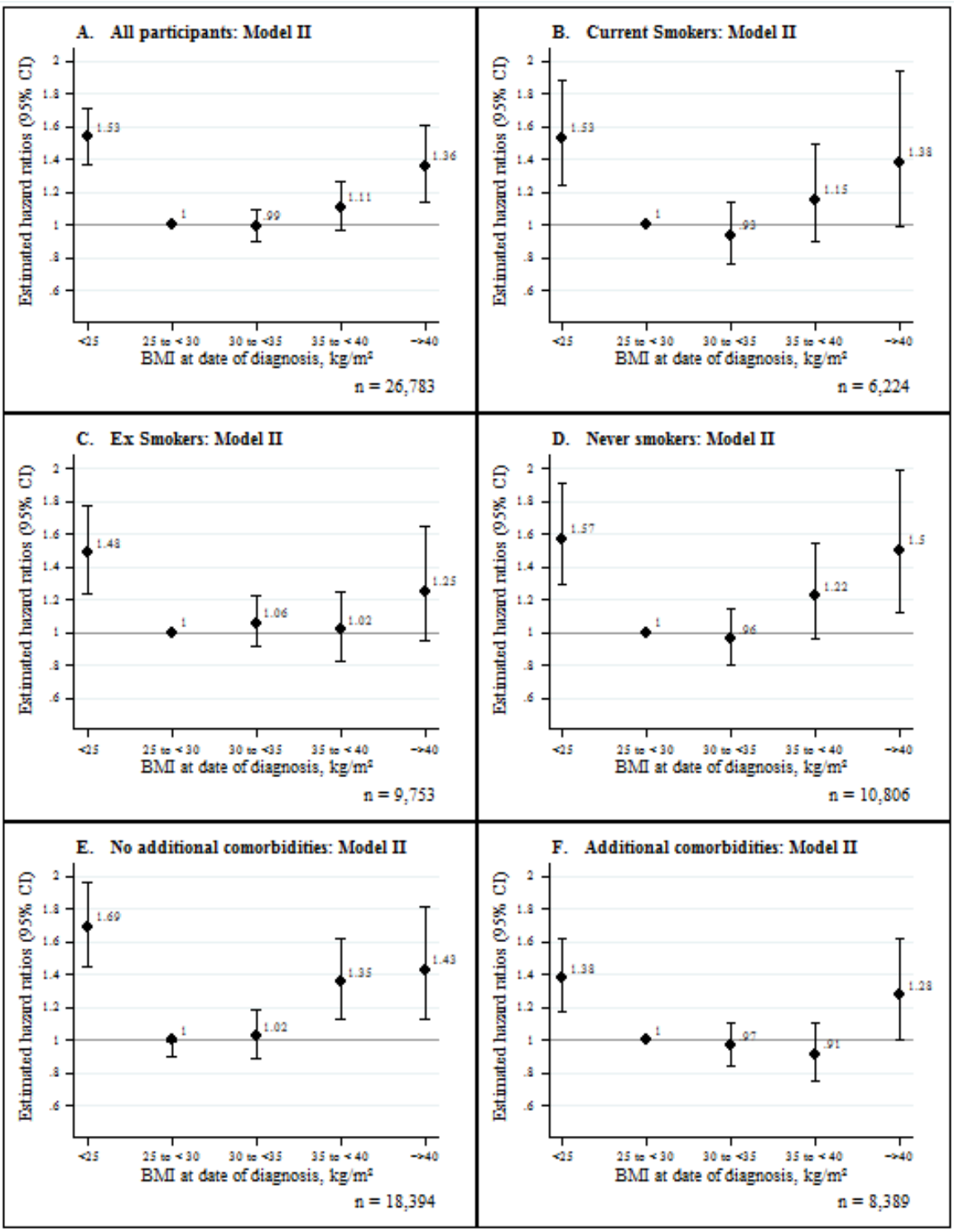

Figure 8.3: Model II estimates for the relationship between categories of BMI and all-cause mortality with stratification.Results for all participants are provided in Panel A. Relative estimates for current smokers are provided in Panel B, for Ex-smokers in Panel C and Never smokers in Panel D. Panel E presents estimates for patients with diabetes only and Panel F presents results for patients with comorbidities.

Table 8.2: Estimates for relationship between categories of BMI at diagnosis and all-cause mortality using CCA (All participants)

		BMI category (kg/m²)				
		< 25	25 to < 30	30 to < 35	35 to < 40	≥ 40
ALL PARTICIPANTS						
MODEL I	Deaths/Person-years	1,047 / 30,094	1,845 / 89,305	1,411 / 89,460	584 / 46,461	359 / 31,599
	Age-standardised mortality	18.9 (17.3, 20.6)	11.7 (11.0, 12.3)	11.1 (10.4, 11.8)	11.4 (10.3, 12.4)	14.9 (11.7, 18.1)
	Multivariable adjusted HR	1.48 (1.37, 1.60)	1.00	0.97 (0.91, 1.04)	1.02 (0.93, 1.13)	1.35 (1.20, 1.51)
MODEL II	Deaths/Person-years	507 / 15,504	905 / 46,779	715 / 46,645	311 / 24,059	172 / 16,349
	Age-standardised mortality	17.9 (15.7, 20.1)	10.9 (10.1, 11.8)	10.6 (9.7, 11.5)	11.7 (10.2, 13.2)	16.7 (9.8, 23.6)
	Multivariable adjusted HR	1.53 (1.37, 1.71)	1.00	0.99 (0.90, 1.10)	1.11 (0.97, 1.26)	1.36 (1.14, 1.61)

Table 8.3: Estimates for relationship between categories of BMI at diagnosis and all-cause mortality using CCA, stratified by smoking status

		BMI category (kg/m²)				
		< 25	25 to < 30	30 to < 35	35 to < 40	≥ 40
CURRENT SMOKERS						
MODEL I	Deaths/Person-years	335 / 9,324	463 / 20,807	355 / 19,815	155 / 10,094	96 / 6,568
	Age-standardised mortality	27.4 (23.9, 31.0)	17.5 (15.5, 19.5)	15.5 (13.4, 17.6)	16.1 (12.1, 18.3)	17.0 (12.4, 21.6)
	Multivariable adjusted HR	1.56 (1.35, 1.79)	1.00	0.97 (0.84, 1.11)	1.05 (0.87, 1.26)	1.31 (1.05, 1.65)
MODEL II	Deaths/Person-years	152 / 4,740	229 / 10,989	170 / 10,403	85 / 5,426	44 / 3,279
	Age-standardised mortality	25.9 (20.8, 31.1)	15.0 (12.7, 17.3)	14.3 (11.5, 17.0)	15.3 (12.3, 17.7)	16.2 (9.3, 23.0)
	Multivariable adjusted HR	1.53 (1.24, 1.88)	1.00	0.93 (0.76, 1.14)	1.15 (0.89, 1.49)	1.38 (0.99, 1.94)
EX-SMOKERS						
MODEL I	Deaths/Person-years	361 / 8,784	798 / 32,887	636 / 34,519	236 / 16,531	123 / 10,177
	Age-standardised mortality	16.7 (13.6, 19.8)	12.0 (10.9, 13.0)	11.1 (10.0, 12.3)	11.1 (9.2, 12.9)	17.5 (6.6, 28.3)
	Multivariable adjusted HR	1.46 (1.28, 1.65)	1.00	0.98 (0.88, 1.09)	0.95 (0.82, 1.11)	1.22 (1.00, 1.48)
MODEL II	Deaths/Person-years	179 / 4,485	401 / 17,407	345 / 18,144	127 / 8,544	63 / 5,397
	Age-standardised mortality	14.7 (11.8, 17.6)	11.8 (10.3, 13.4)	11.1 (9.6, 12.5)	11.7 (9.0, 14.3)	11.1 (8.1, 14.2)
	Multivariable adjusted HR	1.48 (1.24, 1.77)	1.00	1.06 (0.91, 1.22)	1.01 (0.83, 1.24)	1.25 (0.95, 1.65)
NEVER SMOKERS						
MODEL I	Deaths/Person-years	351 / 11,984	584 / 35,609	420 / 35,125	193 / 19,835	140 / 14,853
	Age-standardised mortality	11.7 (9.9, 13.5)	8.5 (7.6, 9.4)	8.4 (7.5, 9.4)	9.1 (7.7, 10.6)	14.2 (9.4, 19.1)
	Multivariable adjusted HR	1.44 (1.26, 1.64)	1.00	0.98 (0.86, 1.11)	1.11 (0.94, 1.32)	1.59 (1.31, 1.93)

MODEL II	Deaths/Person-years	176 / 6,279	275 / 18,383	200 / 18,098	99 / 10,089	65 / 7,673
	Age-standardised mortality	11.4 (9.1, 13.8)	8.2 (7.0, 9.5)	7.8 (6.5, 9.1)	9.2 (7.2, 11.3)	14.4 (7.4, 21.5)
	Multivariable adjusted HR	1.57 (1.30, 1.91)	1.00	0.96 (0.80, 1.15)	1.22 (0.96, 1.54)	1.50 (1.13, 1.99)

Table 8.4: Estimates for relationship between categories of BMI at diagnosis and all-cause mortality using CCA, stratified by presence of comorbid conditions

		BMI category (kg/m²)				
		< 25	25 to < 30	30 to < 35	35 to < 40	≥ 40
NO COMORBIDITIES						
MODEL I	Deaths/Person-years	562 / 20,650	901/61,153	672 /61,241	288 / 32,407	176 /22,893
	Age-standardised mortality	15.3 (13.7, 16.9)	9.2 (8.5, 10.0)	8.5 (7.7, 9.3)	8.4 (7.3, 9.6)	12.1 (8.1, 16.2)
	Multivariable adjusted HR	1.55 (1.39, 1.72)	1.00	0.99 (0.89, 1.09)	1.07 (0.94, 1.23)	1.32 (1.12, 1.56)
MODEL II	Deaths/Person-years	281 / 10,884	424 /32,363	323 / 31,869	168 / 16,772	88 / 11,823
	Age-standardised mortality	14.3 (12.3, 16.4)	8.1 (7.2, 9.0)	7.8 (6.8, 8.9)	9.5 (7.8,11.2)	13.8 (6.8, 20.7)
	Multivariable adjusted HR	1.69 (1.45, 1.97)	1.00	1.02 (0.88, 1.18)	1.35 (1.13, 1.63)	1.43 (1.13, 1.82)
COMORBIDITIES						
MODEL I	Deaths/Person-years	485 / 9,443	944 /28,150	739 / 28,219	296 / 14,053	183 / 8,706
	Age-standardised mortality	28.4 (23.4, 33.3)	17.0 (15.3, 18.8)	16.2 (14.6, 17.9)	17.4 (14.7, 20.1)	21.2 (15.3, 27.2)
	Multivariable adjusted HR	1.41 (1.26, 1.58)	1.00	0.96 (0.87, 1.06)	0.98 (0.86, 1.12)	1.38 (1.17, 1.63)
MODEL II	Deaths/Person-years	226 / 4,620	481 /14,417	392 /17,776	143 /7,288	84 / 4,526
	Age-standardised mortality	28.6 (21.3, 35.8)	17.5 (14.7, 20.2)	16.4 (14.0, 18.8)	16.1 (12.5, 19.8)	17.5 (13.4, 21.7)
	Multivariable adjusted HR	1.38 (1.17, 1.62)	1.00	0.97 (0.84, 1.10)	0.91 (0.76, 1.11)	1.28 (1.00, 1.62)

8.3.4 MULTIPLE IMPUTATION USING CHAINED EQUATIONS

MICE was used to create 60 datasets. Due to the number of auxiliary variables included in the imputation model and the size of the dataset, this took a considerable amount of time to run (>5 hours). The inclusion of predictor variables for incomplete variables introduced even more missing data into the analyses. For example, a closest glycated haemoglobin after diagnosis variable was included to improve imputations of glycated haemoglobin at diagnosis, however this new variable included missing data too.

For each variable with missing data, the distribution of imputed and observed values was compared. Box-plots for each continuous imputed variable and bar charts for categorical variables were observed. Even summarising this data graphically was computationally intensive and took considerable time to generate. No gross discrepancies between observed and imputed data were observed and no implausible values were observed. Similarly, generating plotted residuals against fitted values for each imputed dataset was time-consuming. The pattern of residuals was similar across imputed datasets thus indicating no problems with the imputation model.

Upon the inclusion of patients with incomplete data using multiple imputation, relative estimates for the association between categories of BMI and all-cause mortality were largely the same as those observed using CCA (Figure 8.4) (Table 8.5). A U-shaped relationship was observed with the lowest relative risk of mortality being observed in the overweight (25 to <30kg/m^2) BMI category. Obese patients (30 to <35kg/m^2) had a similar risk of mortality to overweight patients. Across all BMI categories absolute mortality rates were considerably larger upon the inclusion of patients with missing data using MICE suggesting considerably more mortality in patients with excluded data. However, mortality rates continued to exhibit a U-shaped relationship with categories of BMI.

The results from Model I and Model II were almost identical suggesting that adjustment for systolic BP, glycated haemoglobin, cholesterol and HDL-cholesterol provided little additional benefit (Table 8.5). Stratification by smoking status had little impact on estimates (Table 8.6). Nonetheless, amongst never smokers, risk of all-cause mortality in the <25kg/m^2 group was lowered (mv HR: 1.29 [1.14, 1.46]) compared with current

smokers (mvHR: 1.42 [1.25, 1.62]) and ex-smokers (mv HR: 1.34 [1.19, 1.52]). Relative estimates of the association stratified by smoking status were similar when CCA and multiple imputation were used.

Interestingly, upon stratification by presence of comorbidities, patients with no comorbidities in the lowest BMI group had higher risk of mortality (mv HR: 1.39 [1.25, 1.53]) than patients with comorbidities in the lowest BMI category (mv HR: 1.30 [1.18, 1.44]) (Table 8.7). No evidence of multiplicative interaction between BMI and age at diagnosis was observed.

The Monte Carlo errors were reasonably small for both the estimate and standard errors (<0.07), indicating that relative estimates were likely to be similar in repeat imputation model runs.

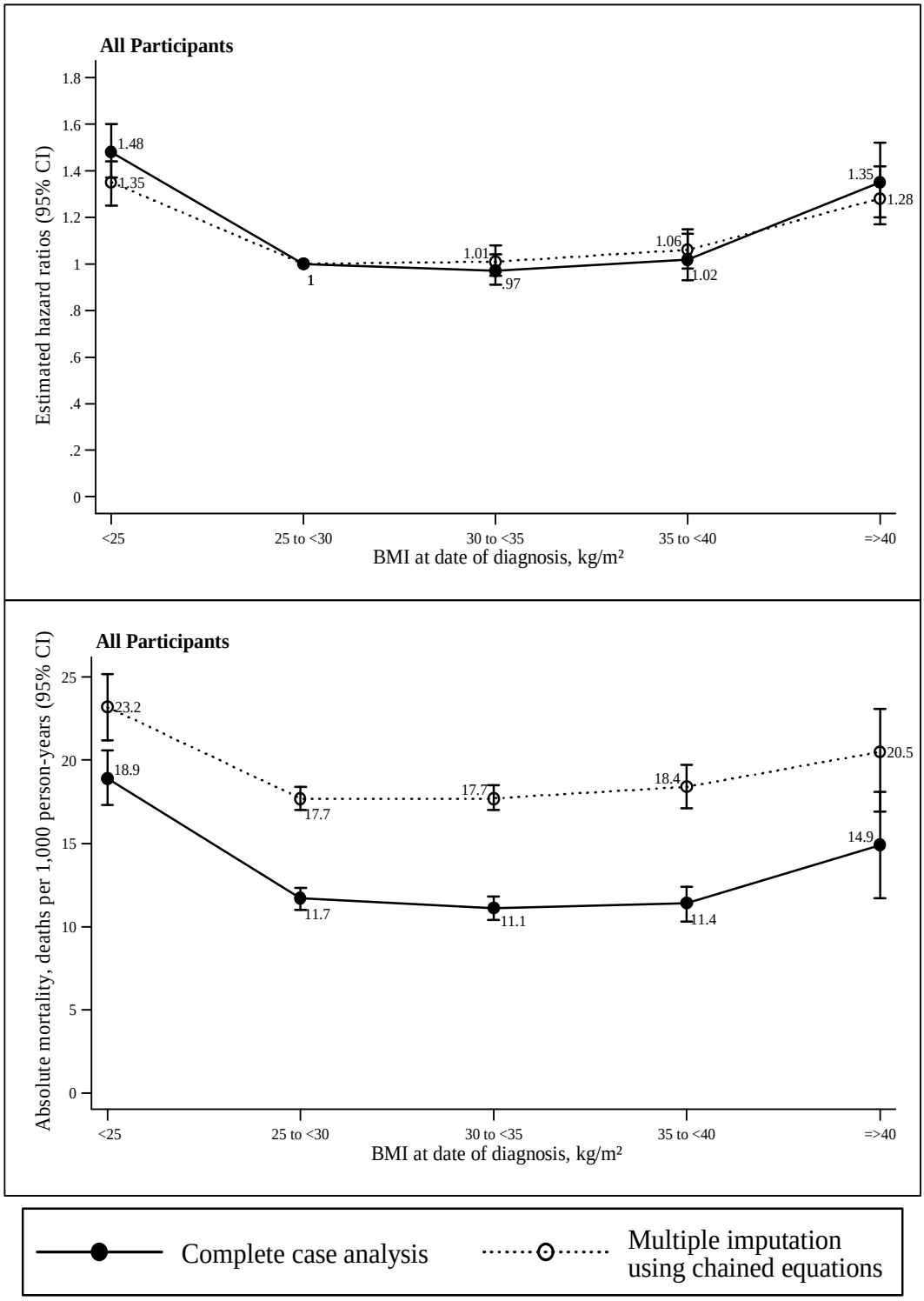

Figure 8.4: A comparison of absolute and relative estimates for the association between categories of BMI and all-cause mortality using CCA and multiple imputation using chained equations. Estimates obtained from analysis Model I.

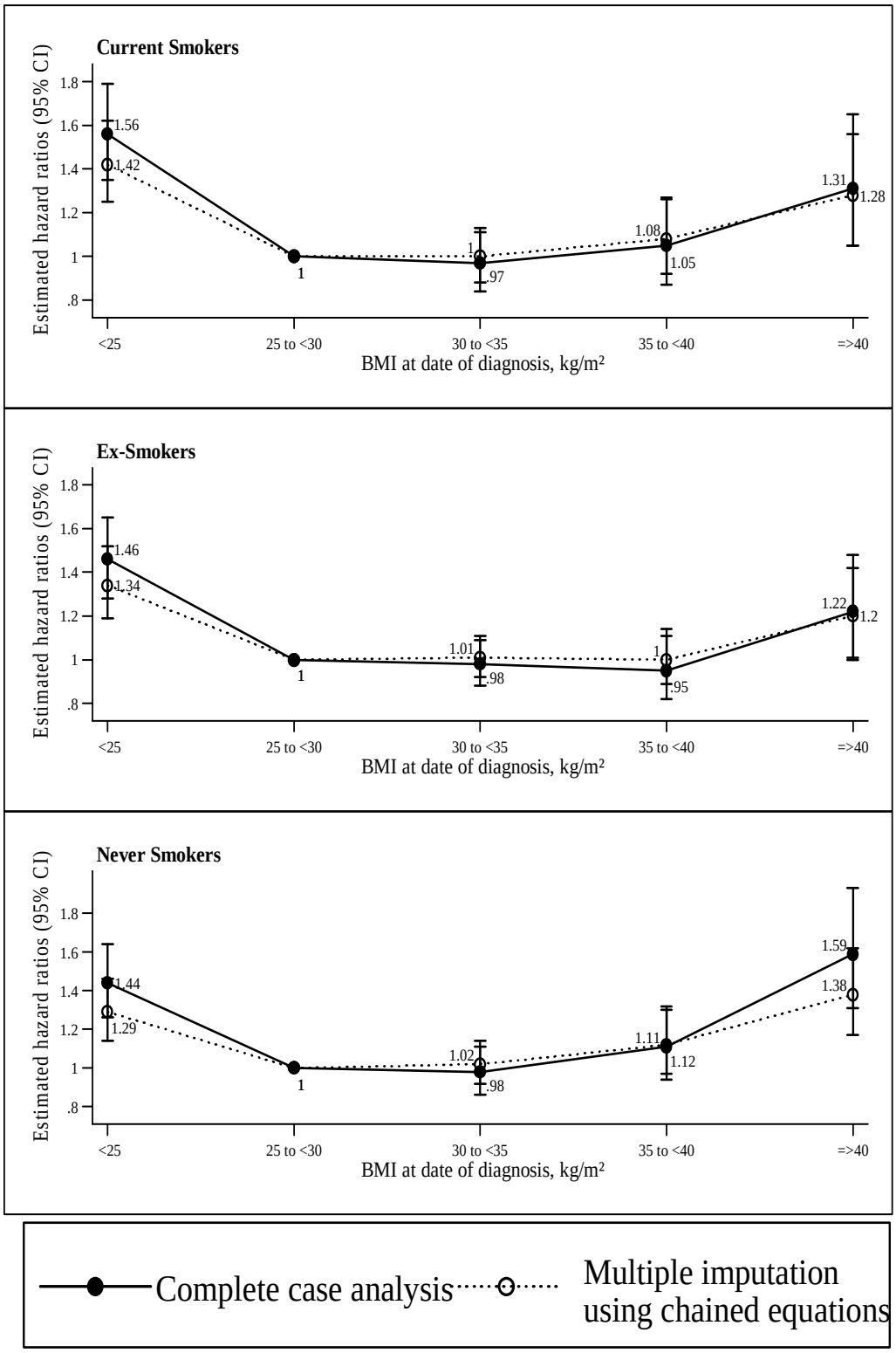

Figure 8.5: A comparison of relative estimates for the association between categories of BMI and all-cause mortality using CCA and multiple imputation using chained equations, stratified by smoking status. Estimates obtained from analysis Model I.

Table 8.5: Estimates for relationship between categories of BMI at diagnosis and all-cause mortality using multiple imputation using chained equations (All participants)

	BMI category (kg/m²)				
	< 25	**25 to < 30**	**30 to < 35**	**35 to < 40**	**≥ 40**
*Deaths/Person-years	1,639/42,887	3,404/125,328	2,955/126,160	1,382/65,753	902/45,025
*Age-standardised mortality	23.2 (21.2, 25.2)	17.7 (17.0, 18.4)	17.7 (17.0, 18.5)	18.4 (17.1, 19.7)	20.5 (16.9, 23.1)
Model I adjusted HR	1.35 (1.25, 1.44)	1.00	1.01 (0.95, 1.08)	1.06 (0.98, 1.15)	1.28 (1.17, 1.42)
Model II adjusted HR	1.33 (1.24, 1.43)	1.00	1.01 (0.95, 1.08)	1.07 (0.98, 1.16)	1.29 (1.17, 1.42)

* Average across 60 imputed datasets

Table 8.6: Estimates for relationship between categories of BMI at diagnosis and all-cause mortality using multiple imputation using chained equations. Estimates stratified by smoking status

	BMI category (kg/m²)				
	< 25	**25 to < 30**	**30 to < 35**	**35 to < 40**	**≥ 40**
CURRENT SMOKERS					
*Deaths/Person-years	493/12,207	821/28,385	703/27,367	337/14,014	218/9,218
*Age-standardised mortality	32.7 (31.0, 34.2)	24.5 (23.0, 26.0)	24.0 (23.1, 24.9)	26.6 (24.9, 28.9)	28.1 (25.0., 31.2)
Model I adjusted HR	1.42 (1.25, 1.62)	1.00	1.00 (0.88, 1.13)	1.08 (0.92, 1.27)	1.28 (1.05, 1.56)
Model II adjusted HR	1.40 (1.23, 1.61)	1.00	1.01 (0.89, 1.14)	1.08 (0.92, 1.27)	1.28 (1.05, 1.56)
EX-SMOKERS					
*Deaths/Person-years	527/12,214	1,309/43,220	1,160/45,294	496/22,035	293/13,821
*Age-standardised mortality	21.1 (18.9, 23.3)	17.7 (15.3, 20.1)	17.6 (15.4, 20.0)	18.2 (15.7, 20.7)	20.2 (16.6, 23.8)
Model I adjusted HR	1.34 (1.19, 1.52)	1.00	1.01 (0.92, 1.11)	1.00 (0.89, 1.14)	1.20 (1.01, 1.42)
Model II adjusted HR	1.33 (1.18, 1.50)	1.00	1.01 (0.92, 1.11)	1.00 (0.88, 1.15)	1.20 (1.01, 1.42)
NEVER SMOKERS					
*Deaths/Person-years	619/18,466	1,273/53,724	1,092/53,499	549/29,705	392/21,986
*Age-standardised mortality	17.2 (15.8, 18.6)	14.7 (13.0, 16.4)	14.9 (13.1, 16.7)	15.9 (14.5, 17.3)	18.0 (15.0, 21.1)
Model I adjusted HR	1.29 (1.14, 1.46)	1.00	1.02 (0.92, 1.14)	1.12 (0.97, 1.30)	1.38 (1.17, 1.62)
Model II adjusted HR	1.28 (1.13, 1.45)	1.00	1.03 (0.92, 1.15)	1.13 (0.98, 1.30)	1.38 (1.18, 1.63)

* Average across 60 imputed datasets

Table 8.7: Estimates for relationship between categories of BMI at diagnosis and all-cause mortality using multiple imputation using chained equations. Estimates stratified by presence of comorbid conditions

	BMI category (kg/m²)				
	< 25	**25 to < 30**	**30 to < 35**	**35 to < 40**	**≥ 40**
NO COMORBIDITIES					
*Deaths/Person-years	730/29,351	1,320/85,841	1,069/86,208	488/45,616	320/32,191
*Age-standardised mortality	14.5 (12.8, 16.2)	10.0 (8.6, 11.4)	9.7 (8.2, 11.2)	10.1 (8.7, 11.5)	11.9 (8.5, 15.3)
Model I adjusted HR	1.39 (1.25, 1.53)	1.00	1.01 (0.92, 1.11)	1.08 (0.96, 1.22)	1.28 (1.11, 1.48)
Model II adjusted HR	1.37 (1.24, 1.52)	1.00	1.02 (0.92, 1.12)	1.09 (0.96, 1.22)	1.29 (1.11, 1.48)
COMORBIDITIES					
*Deaths/Person-years	909/13,535	2,083/39,487	1,887/39,953	895/20,137	582/12,834
*Age-standardised mortality	47.8 (46.0, 49.6)	39.4 (37.6, 41.2)	38.8 (37.0, 40.6)	40.1 (39.1, 41.1)	42.7 (40.1, 45.3)
Model I adjusted HR	1.30 (1.18, 1.44)	1.00	1.01 (0.93, 1.10)	1.05 (0.94, 1.17)	1.28 (1.11, 1.48)
Model II adjusted HR	1.28 (1.16, 1.42)	1.00	1.01 (0.93, 1.10)	1.05 (0.94, 1.17)	1.29 (1.11, 1.48)

* Average across 60 imputed datasets

8.3.5 PREDICTIVE MEAN MATCHING (PMM)

PMM was used to generate 60 imputed datasets. This process took considerably longer than the MICE approach (>15 hours). The model contained twenty variables, of which seven were represented by dummy variables. This feature of the imputation model became problematic and the model failed to run. To overcome this problem, a subset of variables was selected to predict each incomplete variable. Within each custom imputation model, those variables to be included in analysis model II were included. Model diagnostics were carried out and deemed reasonable. Again, this was a lengthy process.

Estimates from PMM were identical to those obtained from MICE. Compared with CCA, PMM resulted in a slight reduction in relative risk of mortality in the highest and lowest BMI categories (Table 8.8) (Figure 8.6). Nonetheless, highest relative mortality remained in the lowest BMI category (<25kg/m² vs. 25 to 30kg/m² mv HR: 1.36 [1.26, 1.46]) compared with the reference category. Relative mortality risk was similar amongst overweight and obese patients (30 to 35kg/m²).

Upon stratification, relative estimates for smokers, ex-smokers and never smokers were similar when obtained from CCA and PMM (Figure 8.7 & Table 8.9). The flattening out of the U-shaped relationship upon using PMM was observed in each stratified analysis with the risk of mortality attenuating in the lowest and highest BMI categories. Similarly, few differences were observed between estimates obtained from CCA and PMM upon stratification by comorbidity status, defined using the Charlson comorbidity index (Figure 8.8 & Table 8.10). Model II produced identical estimates to those obtained in Model II indicating that systolic BP, total cholesterol, glycated haemoglobin and HDL-cholesterol did not confound this relationship.

Changing the size of the match-pool in the imputation model from ten to three made no difference to the estimates observed.

Table 8.8: Estimates for relationship between categories of BMI at diagnosis and all-cause mortality using PMM (All participants)

	BMI category (kg/m²)				
	< 25	**25 to < 30**	**30 to < 35**	**35 to < 40**	**≥ 40**
*Deaths/Person-years	1,650 / 42,915	3,401 / 125,384	2,955 / 126,072	1,374 / 65,826	902 / 44,957
*Age-standardised mortality	23.6 (21.4, 25.8)	17.9 (16.4, 19.4)	17.9 (16.5, 19.4)	18.3 (16.7, 19.9)	20.6 (17.7, 23.5)
Model I adjusted HR	1.36 (1.26, 1.46)	1.00	1.01 (0.95, 1.07)	1.05 (0.97, 1.14)	1.27 (1.14, 1.41)
Model II adjusted HR	1.34 (1.25, 1.45)	1.00	1.01 (0.95, 1.08)	1.05 (0.97, 1.14)	1.27 (1.14, 1.41)

* Average across 60 imputed datasets

Table 8.9: Estimates for relationship between categories of BMI at diagnosis and all-cause mortality using PMM, estimates stratified by smoking status

	BMI category (kg/m²)				
	< 25	**25 to < 30**	**30 to < 35**	**35 to < 40**	**≥ 40**
CURRENT SMOKERS					
*Deaths/Person-years	473 / 12,031	809 / 28,308	703 / 27,447	336 / 14,025	219 / 9,244
*Age-standardised mortality	31.6 (30.1, 33.1)	23.8 (22.6, 25.0)	23.5 (21.2, 25.8)	25.0 (23.5, 26.5)	27.1 (23.7, 30.5)
Model I adjusted HR	1.44 (1.26, 1.65)	1.00	1.00 (0.87, 1.15)	1.07 (0.90, 1.28)	1.25 (1.01, 1.54)
Model II adjusted HR	1.42 (1.24, 1.63)	1.00	1.01 (0.88, 1.15)	1.07 (0.90, 1.28)	1.25 (1.01, 1.54)
EX-SMOKERS					
*Deaths/Person-years	545 / 12,400	1,295 / 43,224	1,126 / 45,078	484 / 22,085	293 / 13,856
*Age-standardised mortality	23.0 (21.5, 24.6)	18.0 (16.8, 19.3)	17.6 (16.3, 18.8)	18.0 (16.6, 19.3)	20.9 (17.5, 22.3)
Model I adjusted HR	1.35 (1.20, 1.52)	1.00	0.99 (0.90, 1.10)	0.98 (0.86, 1.13)	1.17 (0.99, 1.39)
Model II adjusted HR	1.33 (1.18, 1.50)	1.00	1.00 (0.90, 1.10)	0.99 (0.86, 1.13)	1.17 (0.99, 1.39)
NEVER SMOKERS					
*Deaths/Person-years	632 / 18,484	1,297 / 53,852	1,126 / 53,547	554 / 29,815	390 / 21,857
*Age-standardised mortality	18.1 (16.8, 19.4)	15.2 (14.1, 16.4)	15.4 (14.3, 16.6)	16.1 (14.9, 15.3)	18.1 (15.8, 21.4)
Model I adjusted HR	1.30 (1.15, 1.48)	1.00	1.04 (0.93, 1.15)	1.12 (0.98, 1.28)	1.39 (1.18, 1.64)
Model II adjusted HR	1.29 (1.13, 1.47)	1.00	1.04 (0.93, 1.16)	1.12 (0.98, 1.28)	1.39 (1.18, 1.64)

Table 8.10: Estimates for relationship between categories of BMI at diagnosis and all-cause mortality using PMM, estimates stratified by presence of comorbid conditions.

	BMI category (kg/m²)				
	< 25	**25 to < 30**	**30 to < 35**	**35 to < 40**	**≥ 40**
NO COMORBIDITIES					
*Deaths/Person-years	727 / 29,331	1,318 / 85,902	1,074 / 11,163	490 / 45,655	319 / 32,157
*Age-standardised mortality	14.4 (12.5, 16.3)	10.0 (8.9, 11.1)	9.8 (8.7, 10.9)	10.1 (8.7, 11.5)	11.8 (8.7, 14.9)
Model I adjusted HR	1.39 (1.26, 1.54)	1.00	1.01 (0.93, 1.11)	1.08 (0.96, 1.22)	1.27 (1.09, 1.48)
Model II adjusted HR	1.38 (1.24, 1.52)	1.00	1.02 (0.93, 1.12)	1.09 (0.96, 1.23)	1.28 (1.10, 1.48)
COMORBIDITIES					
*Deaths/Person-years	923 / 13,584	2,083 / 39,481	1,882 / 39,909	884 / 20,171	583 / 12,800
*Age-standardised mortality	50.2 (48.2, 52.1)	41.7 (40.5, 42.9)	40.0 (38.8, 41.2)	40.4 (39.0, 41.8)	44.1 (41.4, 47.7)
Model I adjusted HR	1.32 (1.19, 1.46)	1.00	1.00 (0.92, 1.09)	1.03 (0.92, 1.15)	1.27 (1.09, 1.47)
Model II adjusted HR	1.30 (1.18, 1.44)	1.00	1.01 (0.92, 1.10)	1.03 (0.92, 1.15)	1.27 (1.09, 1.47)

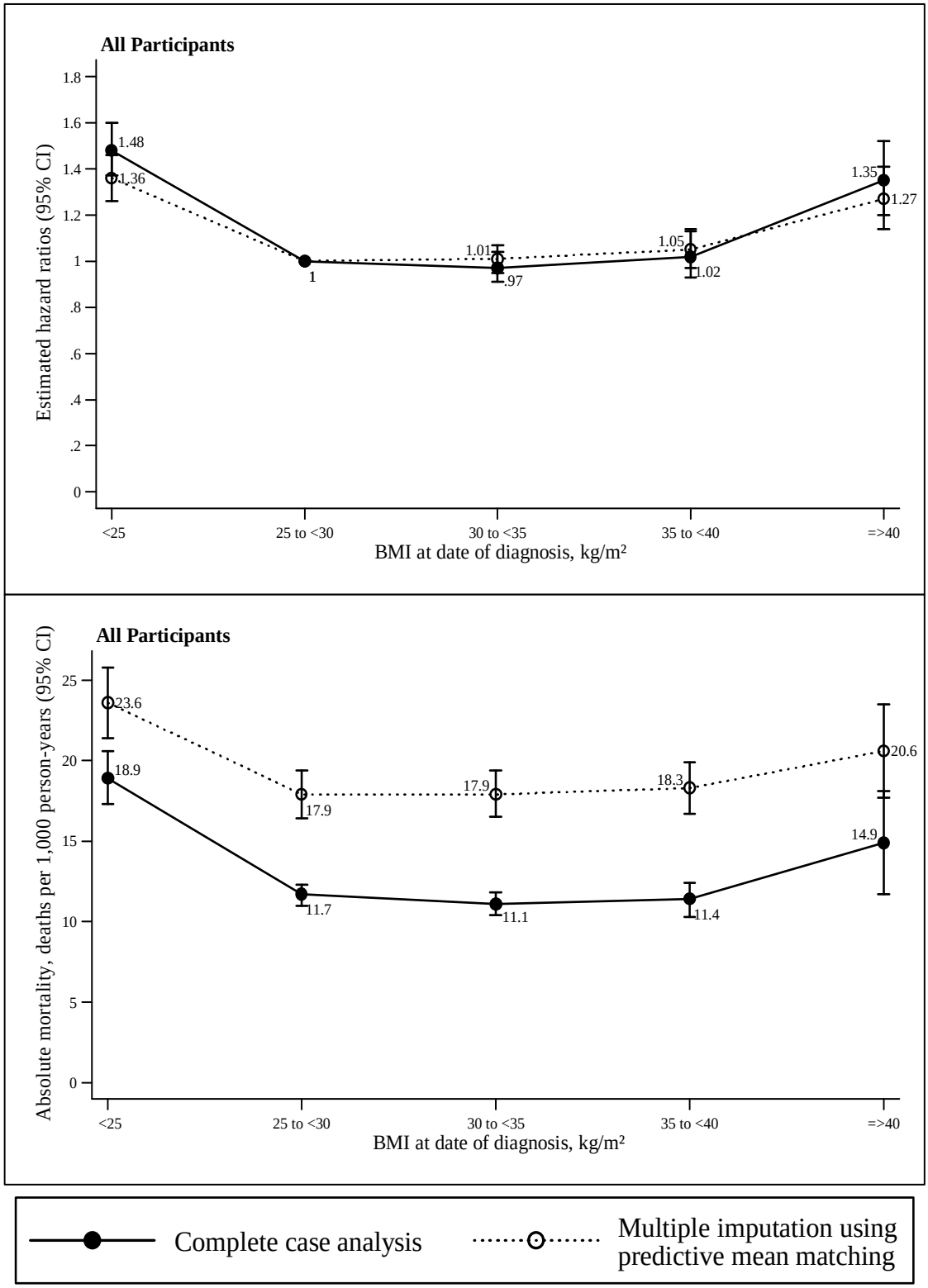

Figure 8.6: A comparison of relative and absolute estimates for the association between categories of BMI and all-cause mortality using CCA and multiple imputation using PMM. Relative estimates obtained from Model I.

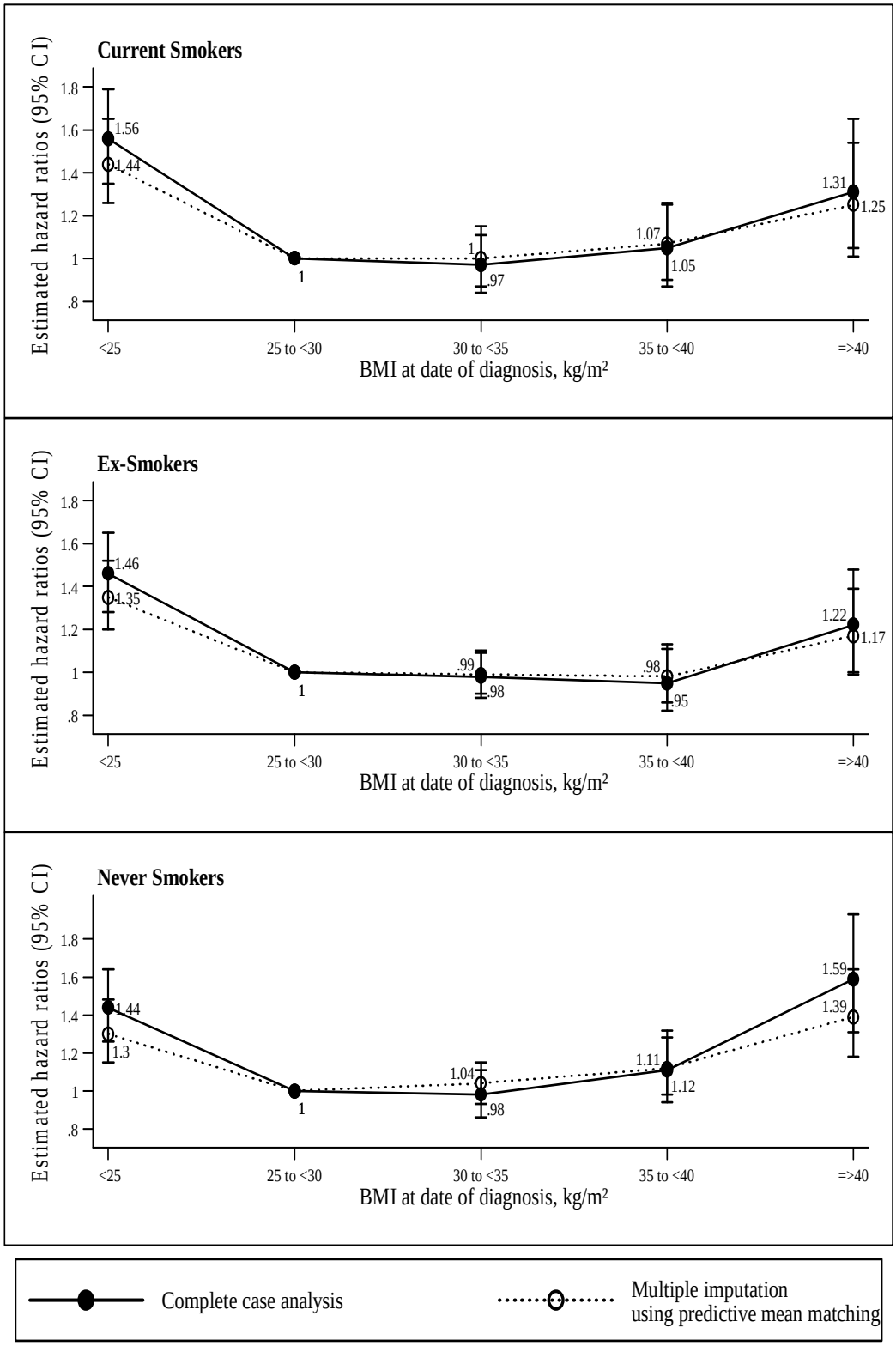

Figure 8.7: A comparison of relative estimates for the association between categories of BMI and all-cause mortality using CCA and multiple imputation using PMM, stratified by smoking status at diagnosis. Estimates were obtained from Model I

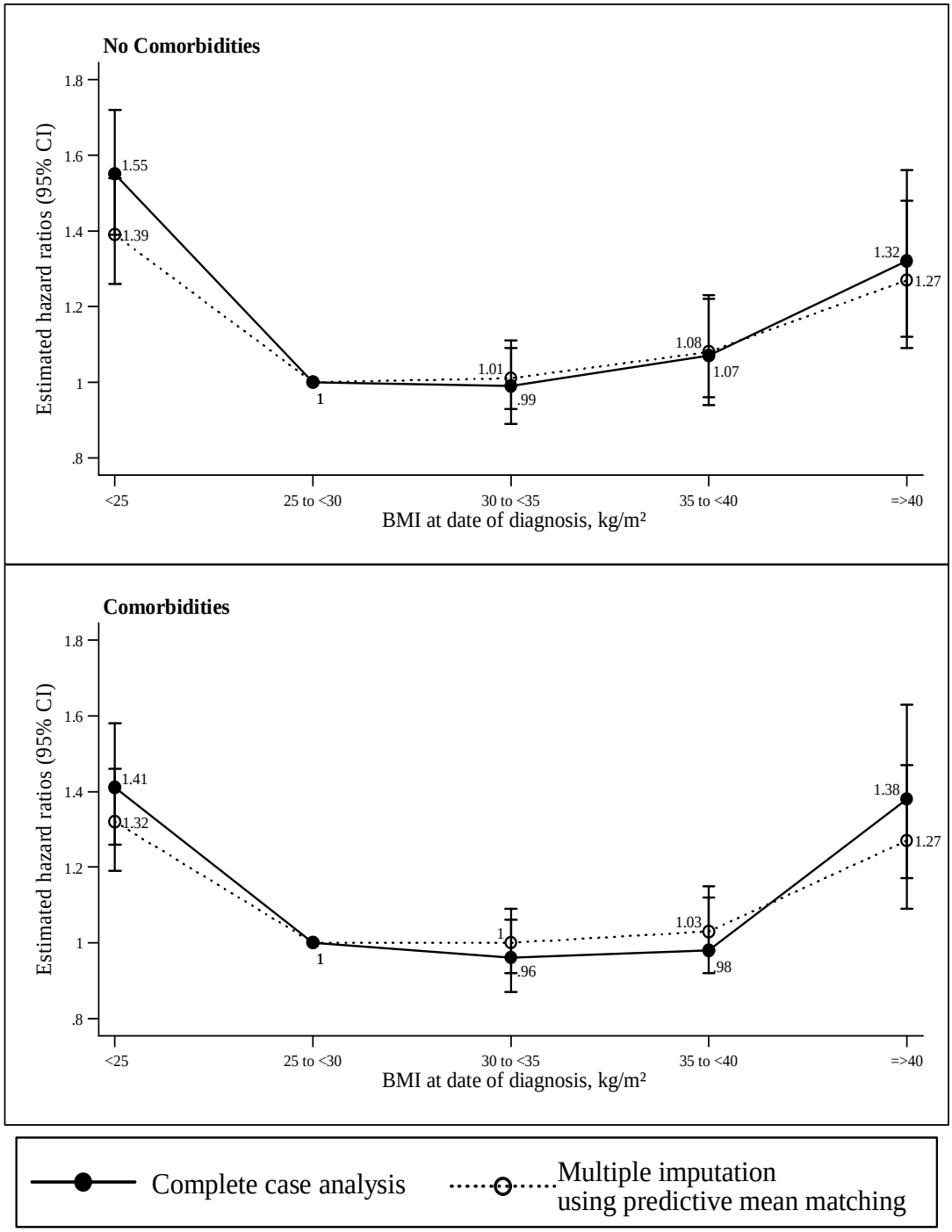

Figure 8.8: A comparison of relative estimates for the association between categories of BMI and all-cause mortality using CCA and multiple imputation using PMM, stratified by comorbidities. Estimates obtained from Model I

8.3.6 MULTIVARIATE NORMAL (MVN) MULTIPLE IMPUTATION

The MVN model was used to create 60 imputed datasets. This model ran considerably quicker than the previous two imputation models.

However, additional steps were required to successfully run the analysis model on each imputed dataset. To impute categorical variables it was necessary to include each dummy variable excluding the referent category and these were imputed assuming they were continuous. Accordingly imputed values fell between 0 and 1. After imputation, to calculate the predicted value for the referent category the values from the dummy variables were subtracted from one, as recommended by Enders (2010). Rounding was then used in each imputed dummy variable to replace values with zero or one.

Table 8.11 presents the results when the MVN method was applied. Relative and absolute estimates for all participants are presented in Figure 8.9. Relative estimates are almost identical to those obtained using CCA with only slight reductions in risk of all-cause mortality being observed in the highest and lowest BMI categories.

Upon stratification by smoking status, an error message stated that the estimation sample varied for each imputation model. When using the *ice* command for the chained equations and PPM matching approach, this potential problem went undetected. Differing estimation samples may occur when stratifying by an imputed variable and can be a source of bias. The distribution of smoking status was observed across imputed datasets and deemed reasonable (See appendix Table A8.18). Stratified relative results are presented in Figure 8.10 & Table 8.12 and were similar when obtained using CCA and the MVN model. Amongst never smokers however, the relative risk of mortality in the >40kg/m^2 group was considerably lower than observed using CCA.

Table 8.11: Estimates for relationship between categories of BMI at diagnosis and all-cause mortality using MVN model (All participants)

	BMI category (kg/m²)				
	< 25	**25 to < 30**	**30 to < 35**	**35 to < 40**	**≥ 40**
*Deaths/Person-years	1,803 / 42,872	3,637 / 130,874	2,875 / 130,692	6,548 / 67,048	742 / 44,547
*Age-standardised mortality	24.6 (22.5, 26.7)	17.3 (16.2, 18.4)	16.0 (15.9, 18.2)	17.7 (16.5, 18.9)	21.1 (17.8, 24.5)
Model I adjusted HR	1.43 (1.33, 1.54)	1.00	0.98 (0.92, 1.05)	1.03 (0.94, 1.12)	1.33 (1.21, 1.47)
Model II adjusted HR	1.41 (1.32, 1.52)	1.00	0.99 (0.92, 1.05)	1.03 (0.94, 1.13)	1.34 (1.21, 1.48)

* Average across 60 imputed datasets

Table 8.12: Estimates for relationship between categories of BMI at diagnosis and all-cause mortality using MVN model, estimates are stratified by smoking status.

	BMI category (kg/m²)				
	< 25	**25 to < 30**	**30 to < 35**	**35 to < 40**	**≥ 40**
CURRENT SMOKERS					
*Deaths/Person-years	533 / 12,393	893 / 29,621	700 / 28,408	307 / 14,304	189 / 9,087
*Age-standardised mortality	32.1 (30.3, 34.8)	22.1 (20.8, 23.5)	22.1 (20.7, 23.5)	24.0 (22.3, 25.8)	27.4 (24.3, 30.4)
Model I adjusted HR	1.55 (1.36, 1.77)	1.00	1.00 (0.87, 1.14)	1.07 (0.89, 1.28)	1.38 (1.04, 1.72)
Model II adjusted HR	1.54 (1.35, 1.77)	1.00	0.99 (0.87, 1.13)	1.05 (0.87, 1.26)	1.39 (1.05, 1.73)
EX-SMOKERS					
*Deaths/Person-years	595 / 12,232	1,425 / 45,341	1,163 / 47,257	458 / 22,415	247 / 13,556
*Age-standardised mortality	21.2 (19.5, 22.9)	15.4 (14.1, 16.6)	14.8 (13.6, 16.1)	15.4 (14.0, 16.7)	19.0 (16.4, 22.6)
Model I adjusted HR	1.40 (1.28, 1.51)	1.00	0.98 (0.88, 1.09)	1.00 (0.87, 1.14)	1.29 (1.04, 1.54)
Model II adjusted HR	1.37 (1.26, 1.49)	1.00	0.98 (0.87, 1.09)	1.01 (0.87, 1.15)	1.28 (1.04, 1.55)
NEVER SMOKERS					
*Deaths/Person-years	675 / 18,247	1,333 / 55,912	1,012 / 55,027	461 / 30,329	306 / 21,905
*Age-standardised mortality	15.9 (13.6, 18.3)	12.1 (10.9, 13.2)	12.0 (10.8, 13.2)	13.0 (11.8, 14.3)	16.3 (13.8, 19.7)
Model I adjusted HR	1.30 (1.17, 1.43)	1.00	1.00 (0.89, 1.12)	1.04 (0.93, 1.15)	1.26 (1.01, 1.49)
Model II adjusted HR	1.30 (1.16, 1.43)	1.00	1.00 (0.89, 1.12)	1.05 (0.96, 1.17)	1.25 (1.01, 1.48)

* Average across 60 imputed datasets

Table 8.13: Estimates for relationship between categories of BMI at diagnosis and all-cause mortality using MVN model, estimates are stratified by presence of comorbidities.

	BMI category (kg/m²)				
	< 25	**25 to < 30**	**30 to < 35**	**35 to < 40**	**≥ 40**
NO COMORBIDITIES					
*Deaths/Person-years	767 / 28,210	1,333 / 55,912	1,012 / 55,027	461 / 30,329	306 / 21,905
*Age-standardised mortality	15.4 (13.1, 17.7)	10.0 (8.9, 11.2)	9.5 (8.3, 10.6)	9.8 (8.6, 11.0)	12.6 (9.1, 15.0)
Model I adjusted HR	1.45 (1.31, 1.61)	1.00	0.98 (0.89, 1.08)	1.06 (0.93, 1.20)	1.33 (1.15, 1.53)
Model II adjusted HR	1.44 (1.30, 1.59)	1.00	0.99 (0.89, 1.09)	1.06 (0.94, 1.21)	1.33 (1.15, 1.54)
COMORBIDITIES					
*Deaths/Person-years	1,036 / 14,663	2,255 / 44,024	1,832 / 43,657	774 / 21,407	458 / 13,018
*Age-standardised mortality	39.7 (37.6, 41.8)	25.8 (24.4, 27.2)	24.8 (23.4, 26.1)	25.9 (24.3, 25.4)	30.3 (27.6, 33.9)
Model I adjusted HR	1.40 (1.26, 1.55)	1.00	0.98 (0.90, 1.07)	1.00 (0.88, 1.14)	1.33 (1.14, 1.54)
Model II adjusted HR	1.38 (1.25, 1.53)	1.00	0.99 (0.90, 1.08)	1.01 (0.89, 1.15)	1.34 (1.15, 1.56)

* Average across 60 imputed datasets

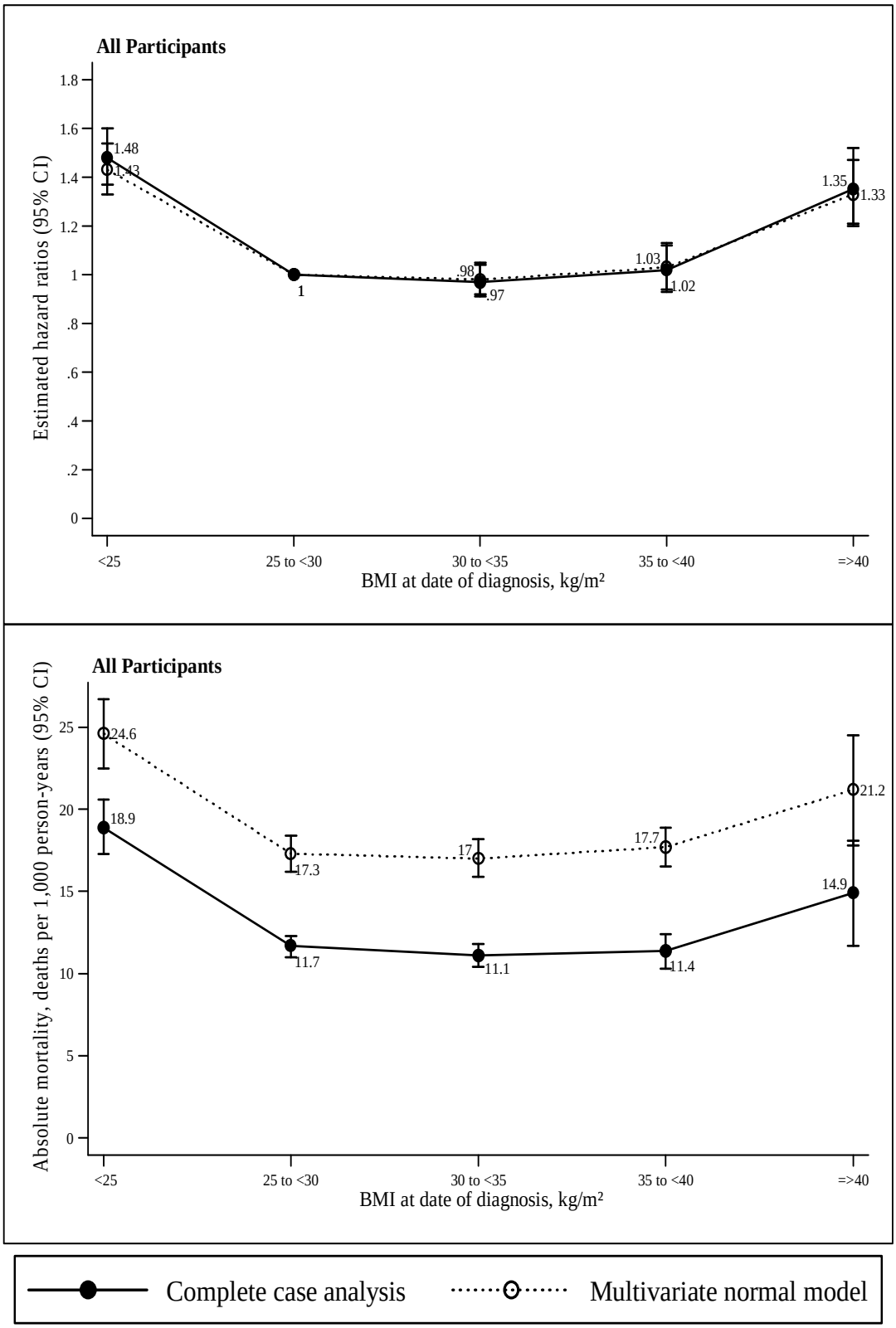

Figure 8.9: comparison of relative and absolute estimates for the association between categories of BMI and all-cause mortality using CCA and MVN model. Relative estimates obtained from Model I.

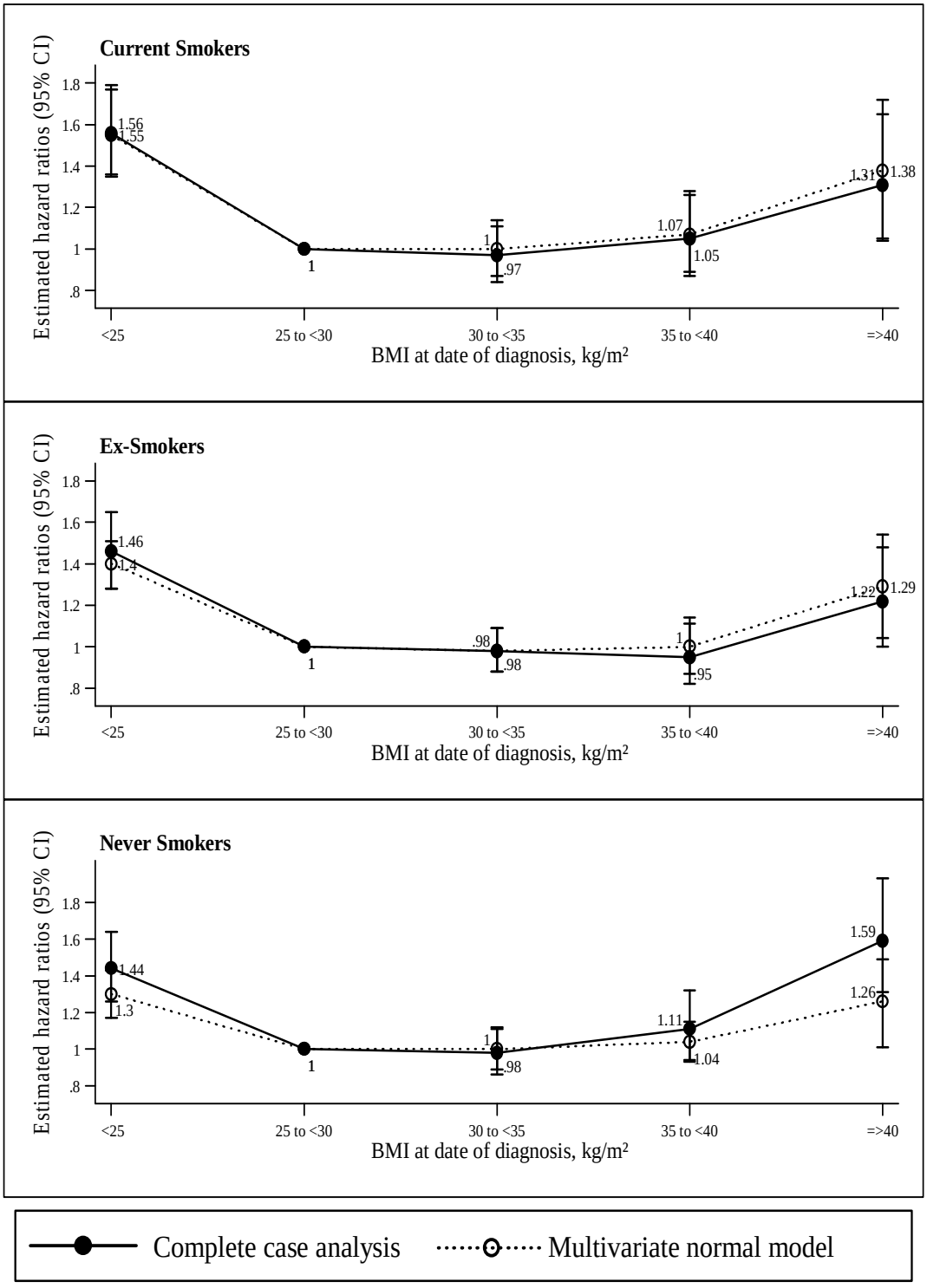

Figure 8.10: A comparison of relative estimates for the association between categories of BMI and all-cause mortality using CCA and multiple imputation using MVN model, stratified by smoking status. Estimates obtained from Model I

8.3.7 ALL METHODS

Differences in relative differences between the four methods utilised are presented in Figure 8.11. From this Figure it is clear that estimates obtained when using MICE and PMM are very similar whilst estimates from the MVN model were most similar to estimates from CCA. Figure 8.12 presents absolute estimates from each method for handling missing data. Each multiple imputation model produced largely comparable results and observed differences would be unlikely to affect overall inferences. Clearly, failure to include patients who had unobserved data in the variables of interest would produce invalid absolute estimates.

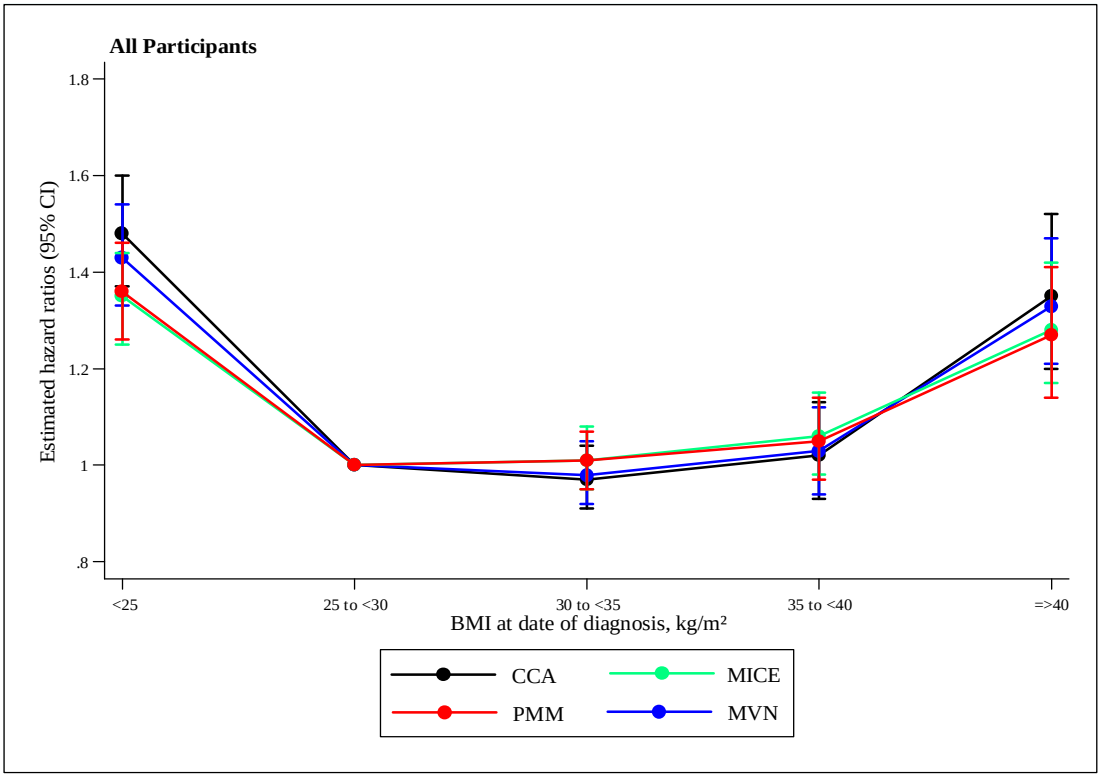

Figure 8.11: A comparison of relative estimates for the association between categories of BMI and all-cause mortality using four different methods for handling incomplete data. All estimates were obtained from Model I.

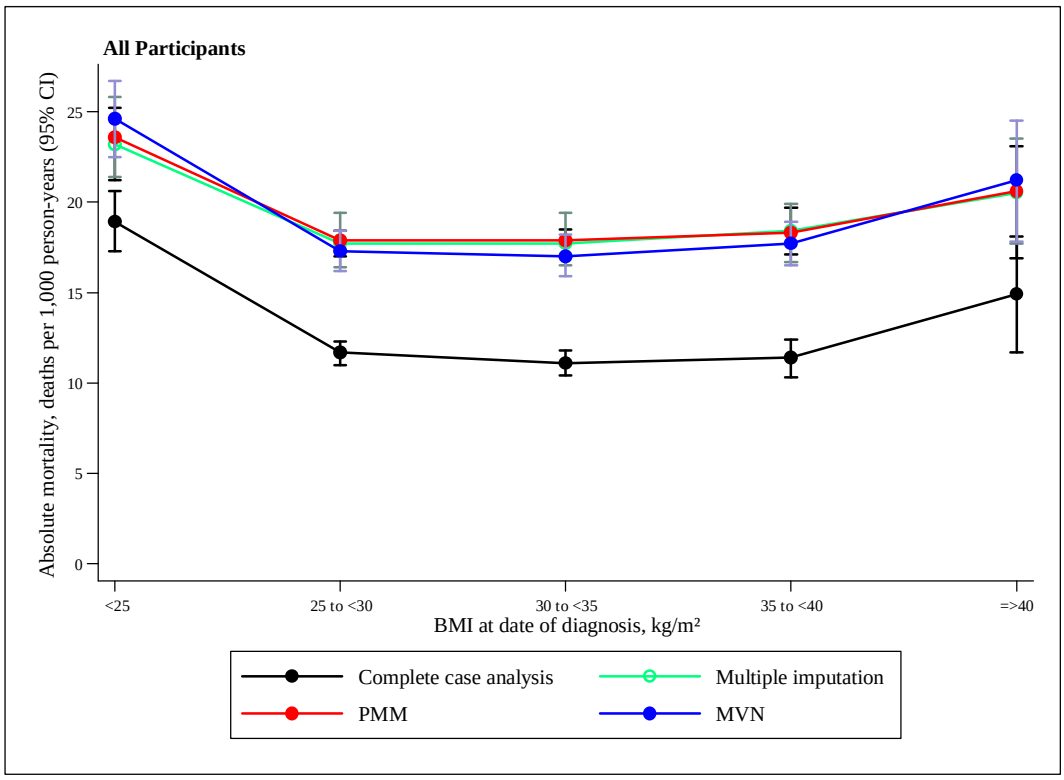

Figure 8.12: A comparison of absolute estimates for the association between categories of BMI and all-cause mortality using four different methods for handling incomplete data. All estimates were obtained from Model I.

8.4 DISCUSSION

8.4.1 INTRODUCTION

The use of a number of multiple imputation approaches has been demonstrated in this large retrospective cohort study investigating the association between BMI at diagnosis and all-cause mortality amongst patients with incident type 2 DM. Regardless of the imputation method used, the observed relationship was U-shaped. This section is subdivided into two sections with the first presenting details on the missing data issues investigated whilst the second presents details on the findings from the analysis model.

8.4.2 MISSING DATA IN SCI-DC

Unobserved data in each of the variables to be included in Model I and Model II analysis model combined to result in complete data for only seventy and thirty-seven percent of patients, respectively. These levels of completeness are undoubtedly higher

than earlier years (See Section 5.4), suggesting a positive impact by the QOF on risk factor recording. Improvements in the recording of important risk factor recording, such as smoking status (Taggar et al., 2012), ethnicity (Mathur et al., 2013) and BMI (Bhaskaran et al., 2013) have been observed in multiple other UK-based electronic health databases. Despite these improvements, clinician work-load, irregular GP attendance and the respective study's definition of complete data mean that missing data are an inevitable problem in medical research using routine data.

Missing data occurred in a non-monotone pattern, though the presence of incomplete data in some variables was more likely amongst patients with unobserved data in other variables. Data appeared to be, but not definitely, MAR with the main predictor of missing data in all variables being patient survival. Patients with missing data had considerably poorer survival than those with complete data resulting in the underestimation of absolute mortality in categories of BMI. This setting whereby completeness of covariates is related to the outcome of interest has been reported in other observational studies (van Buuren et al., 1999; Hotchkiss & Leyland, 2011). Clearly, the risk of bias in this setting is high and therefore the choice of appropriate methods for handling missing data is essential in ensuring the study's validity.

Due to its ability to generate unbiased parameter estimates when data are MAR, multiple imputation was selected as a potentially appropriate method for handling the unobserved data. In this study three separate multiple imputation models were applied; MICE, PMM and the MVN model. These methods can be implemented in commonly used statistical software packages, yet there remains little practical guidance as to which imputation model is the most appropriate for any given missing data problem. Those studies which have provided guidance have done so using simulation studies and few have presented comparisons using real-data with real missing data problems (van Buuren, 2007; Lee & Carlin, 2010; Kropko et al., 2014). Furthermore, only one study was identified which compared the use of PMM with MICE and MVN (Lee & Carlin, 2010). No study was identified which compared the performance of these approaches in survival analyses.

Typically the two main concerns regarding the use of MVN is the assumption regarding the distribution of the joint data and secondly the rounding process for non-continuous

data. The MVN model assumes a joint normal distribution of the data, an assumption that is regularly violated in the presence of binary or categorical variables. Nonetheless, it has been observed that this model is relatively robust to departures from the normal distribution (Schafer & Graham, 2002; Demirtas et al., 2008; Kropko et al., 2014). In their study which tested MVN and MICE in a political research setting, Kropko et al., (2014) observed that MVN and MICE performed similarly well when continuous variables were imputed which did not exhibit a MVN distribution. Demirtas et al., (2008) also reported reasonable performance of MVN even when the normality assumption did not hold, particularly when the sample size was large. In the analysis presented here, non-normal variables were transformed in the MVN analysis and back-transformed when included in the imputation model. PMM was used as an alternative means of handling non-Normal variables within the MICE framework. However, results obtained from MICE and PMM were almost identical suggesting that the transformation of variables in MICE was an adequate method for handling non-Normality. Nonetheless, the performance of MICE may not be comparable to PMM in all contexts. For example, in the presence of non-linear relationships the performance of PMM is likely to be superior to MICE (Morris et al., 2014).

Using the MVN model, categorical variables can be imputed as a set of indicator variables whilst ordinal variables can be imputed in the same way or as a single continuous variable. However, it has been shown that imputing ordinal variables as continuous variables can distort non-linear relationships between imputed covariates variables and the outcome of interest (Lee et al., 2012). In this study, ordinal and categorical variables were imputed as a set of indicator variables and then rounded according to the combination of indicators (e.g. for each patient, the indicator with the highest imputed value was categorised as having the characteristic, so set to 1), an approach recommended by Enders (2010). However, this approach to re-categorising imputed values into appropriate categories by simple rounding has been criticised and other approaches such as adaptive rounding and calibration have been suggested (Bernaards et al., 2007; Lee et al., 2012). The issue of rounding in MVN analyses has been extensively studied and it has been observed that inappropriate rounding approaches can introduce considerable bias (Horton et al., 2003; van Buuren, 2007). For this reason van Buuren (2007) recommends the use of MICE approaches when the joint

distribution is not easily specified in the presence of missing data. Similarly, Kropko et., (2014) advised against using MVN for ordinal and categorical data. Nonetheless, even using a very simplistic rounding method in MVN, estimates from this study were not dissimilar to those obtained from MICE and PMM which have more natural means of handling non-continuous data. Similarly, Lee and Carlin (2010) presented results which indicated that comparable results can be expected from MICE and MVN approaches in the context of linear regression with a mixture of variable types. Due to time-constraints, alternative means of rounding were not investigated. Clearly however, it would be worthwhile to check whether estimates were robust to changes in the rounding approach employed.

In conclusion, three multiple imputation methods for handling unobserved data produced similar estimates. Despite strict assumptions made by the MVN model it appears that this model can be relatively robust to violations to the joint distribution assumption. However, this has yet to be vigorously investigated in the context of Cox regression analyses and therefore further work may be required. Ultimately the choice between the three is likely to be based upon practical considerations and potentially personal preference. In this large dataset, each approach took considerable time to run particularly PMM; a problem that is likely to be common in studies utilising large routinely collected databases. In addition, imputation model specification is a challenging undertaking and this in itself is an arduous activity. For example, in the context of a study with multiple incomplete variables, a balance needs to be struck as to how to improve the predictive power of the imputation model by including more variables whilst restricting the amount of missing data.

Finally, the likely biggest source of bias when using these methods is incorrect model specification. Some of the issues faced upon specifying imputation models are presented in Chapter 8.

8.4.3 Analysis of Association Between BMI and All-Cause Mortality

Relative estimates of the association between BMI and all-cause mortality did not differ significantly by method used to handle missing data. However, absolute estimates from

the multiple imputation approaches were considerably different, highlighting the problems of estimating absolute estimates when patients are excluded from analyses.

The relationship between BMI and all-cause mortality remained U-shaped, providing more evidence of an obesity paradox amongst patients with chronic disease.

When analyses were restricted to patients who had never-smoked, the relationship between BMI and all-cause mortality remained U-shaped, though the risk of mortality did lower slightly for patients in the <25kg/m^2 group. This finding indicates that at least some of the excess risk of mortality in the <25kg/m^2 group may be explained by smoking status. The reliability of smoking status records may also be questionable. Self-reported smoking status records have regularly been shown to be inaccurate with the under-reporting of tobacco-use, particularly when reporting to clinicians (Gorber et al., 2009). If this were the case in this dataset then misclassification bias may have masked the true relationship amongst non-smokers. A previous study conducted by Tobias et al. (2014) in the Nurses' Health Study identified that upon exclusion of current and former smokers, the U-shaped relationship diminished considerably and became a monotonically increasing trend. This effect may have been a result of the inverse relation between BMI and smoking and the increased mortality risk amongst smokers. Nonetheless, many other studies have not observed any differences in trends among smokers and non-smokers (Khalangot et al., 2009; Carnethon et al., 2012). In fact, the practice of excluding patients to investigate the relationship between BMI and mortality is becoming more controversial as large-scale exclusions may result in changes in findings that may simply be due to chance (Cooper, 2008) In addition, these findings cannot be extrapolated to the whole population with DM.

Upon stratification by the presence of comorbidities, estimates for all-cause mortality in the lowest BMI category were highest amongst patients without comorbidities. This is an interesting finding given that the presence of comorbid conditions is likely to contribute to both weight loss and increased mortality. For this reason, upon stratification it was thought that relative risk of mortality in this BMI category would actually decrease. There are a few potential explanations for this finding. Firstly, this may reflect the presence of undiagnosed disease amongst patient with no comorbid conditions. However, it is unlikely that undiagnosed disease contributed to such a large

change in relative estimates and absolute estimates of mortality in this group remain low.

It has been suggested that the validity of BMI as a measure of adiposity decreases amongst elderly people through sarcopenia. This process lowers body weight through the reduction in lean muscle mass but not body adiposity. In addition, older age confers an increased prevalence of comorbid conditions thus resulting in an increased likelihood of reverse causation amongst older patients. Accordingly, it has been observed elsewhere that amongst patients below 65 years the relationship between BMI and all-cause mortality is comparatively linear whilst U-shaped amongst patients over 65 years (Zoppini et al., 2003; Tobias et al., 2014). Nonetheless, age (<65 vs. ≥ 65years) was not an effect modifier of the relationship between BMI and all-cause mortality in this study, nor in a recent pooled analysis of five longitudinal studies (Carnethon et al., 2012). Future analyses could investigate higher cut-offs since 65 years may be too low in patients with type 2 DM as the effect may appear at later ages.

Despite the World Health Organisation advising against adjustment for additional confounders that are closely related to BMI such as hypertension and hyperlipidaemia due to the potential for downward bias, additionally adjusting for systolic BP, total cholesterol and glycated haemoglobin had very little impact on relative estimates (World Health Organisation, 2000). This may be since this study was amongst people with DM and therefore the differences in lipids and blood pressures between obese and non-obese patients may not be as marked than for people without DM. Mulnier et al., (2006) suggested that the finding of increased mortality amongst patients with low BMIs could be a consequence of poorly controlled DM. Although this potential explanation was not investigated in this study, controlling for glycated haemoglobin at diagnosis had no effect on estimates.

In conclusion, this study provides additional evidence of the existence of an obesity paradox. Even after controlling for residual confounding by excluding smokers and controlling for reverse causation by excluding patients who had died within two years of diagnosis and restricting analyses to patient with no comorbid conditions, risk of mortality remained lowest in overweight or obese participants. The presence of an obesity paradox amongst patients with DM may have important implications for future

practice. Currently, patients with DM are encouraged to lose weight, however these findings suggest that this may not be appropriate for all patients.

Further work is required to assess whether patients diagnosed at a lower BMI have a greater disease severity than patients diagnosed with higher BMI categories. This could be assessed by comparing occurrence of diabetes complications such as retinopathy and nephropathy by BMI at diagnosis. This may assist in the interpretation of the relationship between BMI and mortality.

Furthermore, a trial which observes weight changes following diagnosis with type 2 DM and their association with a number of health outcomes may also be beneficial. Given the observational nature of the studies conducted to date, it has not been possible to distinguish between intentional and unintentional weight loss, a trial could overcome this problem. This trial could provide justification for weight loss advice following diagnosis with type 2 DM.

9. DISCUSSION

9.1 INTRODUCTION

Missing data are an inevitable problem across epidemiological research and their presence can lead to inefficient and sometimes biased analyses. The most effective way of managing missing data is to collect complete data in the first place. Unfortunately, this approach is not tenable when data are obtained from electronic healthcare records whose recording is typically clinically driven. These sources of data are information-rich and offer numerous opportunities for clinical research including the long-term follow-up of traditionally hard-to-reach populations. Despite these obvious benefits, concerns surrounding data quality are unavoidable and so appropriate methods for handling missing data are required to ensure analyses remain valid despite the presence of missing data.

Various approaches have been developed and utilised to handle missing data. Complete case analysis, the practice of excluding incomplete cases is the most commonly applied method despite some well-established limitations. Other approaches include single imputation methods whereby missing values are replaced with some value obtained from the observed data such as the variable mean. Most recently, multiple imputation methods have been incorporated into most standard statistical software packages. Compared to CCA, multiple imputation makes more relaxed assumptions regarding the mechanism of missingness and mitigates the limited efficiency of CCA. This approach aims to generate several plausible imputed datasets using the observed data and then results from each imputed dataset are appropriately combined to produce final estimates. The appropriateness of any given method for handling unobserved data is largely dependent on the mechanism by which the missing data came about. Accordingly, any analysis with missing data should be supplemented with a sensitivity analysis which tests the robustness of the methods assumptions regarding the mechanism of missingness.

This discussion is subdivided into two sections. The first presents a review of the missing data methods applied and the second section assesses the findings from the investigation of the association between BMI at diagnosis and mortality in patients with type 2 DM.

9.2 FINDINGS FROM MISSING DATA

In this large population-based register of patients with DM in Scotland, unobserved data have presented difficult challenges for research. Despite high levels of completeness in demographic characteristics, problems persist in repeatedly measured health indicators such as BMI, BP and cholesterol.

In this study we were interested in the completeness of health indicators, particularly BMI recorded at date of diagnosis with type 2 DM. Apart from fixed variables such as ethnicity and sex, missing data patterns vary through time, and the precise proportion of missing data depends on the definition used in its calculation. Typically, a time-frame is specified around an event of interest, such as date of registration at a GP, and it is observed whether patients have a measurement within this time-frame. Since our analyses were concerned with estimating the effect of BMI at diagnosis, it was necessary to specify a time-frame during which BMI recordings would accurately reflect BMI at diagnosis. Since missing data issues in routine data have not been widely explored, little guidance has been provided in choosing the width of these time-frames. In this study, a time-frame was chosen that was wide enough to limit the extent of missing data but was narrow enough to provide recordings of BMI that reasonably accurately reflected BMI at diagnosis.

Even with a relatively wide time-interval, the proportion of patients diagnosed with type 2 DM without a BMI at diagnosis was high (58%). When analyses were restricted to patients diagnosed between 2004 and 2008, BMI incompleteness reduced to 27% but this was still moderately high. This attenuation in missingness was observed in other important health indicators and possibly reflects the introduction of the QOF. Improvements in completeness levels in routine clinical databases have been reported elsewhere (Bhaskaran et al., 2013; Mathur et al., 2013). However, since most health indicators are recorded based on their clinical relevance and may not be available during the time-period of interest, missing data problems are likely to persist in analyses using the SCI-DC dataset.

Whilst it was not possible to speak to every clinician and data-entry clerk involved with gathering data in this population register, discussions with a selection of healthcare

workers provided useful information regarding recording practices and data entry processes. This information was used to postulate the potential reasons for missing data. Data exploration activities identified key differences between patients with and without missing data. Patients with missing data had considerably worse survival, had an earlier year of diagnosis and were less likely to have comorbid conditions than patients without missing data. This therefore indicated that data were MAR conditional primarily on survival, a finding that has been reported in observational data before (van Buuren et al., 1999; Hippisley-Cox et al., 2007). In the development of the QRISK score using the QRESEARCH database, Hippisley-Cox and colleagues (2007) reported that men with missing cholesterol ratios had a 10 year risk of a cardiovascular event of 10.9% compared to 4.9% for patients with a recording. This observation possibly reflects patients who are very unwell at the date of measurement and so the recording of health indicators such as BMI may not be possible or a priority for these patients. There may be important implications for future research which use this and other routinely collected databases as a result of this finding. Many studies are interested in the impact of an exposure on survival. The exclusion of a large number of events may mean the study will lack power, absolute estimates may be underestimated and relative estimates could be biased if the missing data are clustered around particular values of the exposure.

Nevertheless, using data exploration techniques, it was not possible to rule out the possibility that missing data in the SCI-DC register were MNAR, dependent on the missing value. However, the introduction of QOF in 2004 and the subsequent standardisation of patient care should have ensured the regular recording of health indicators in patients with type 2 DM. Furthermore, interviewed clinicians stated that the recording of health indicators was universal, irrespective of their perceived value.

In Chapter 6, CCA was used to handle incompleteness in the BMI at diagnosis variable only. This approach reduced the sample size by 25% and the number of events by 32%. In this large dataset, exclusions such as these had limited impact on the study's power since the sample size remained very large. It is unlikely that the exclusion of this proportion of patients would be acceptable in many other studies, particularly if the excluded participants were mostly those with a rare outcome such as amputations.

We could find little evidence that this approach introduced substantial bias in this case. Population imputation illustrated the problem of underestimated survival by increasing the absolute and relative risk of mortality in the category in which patients with missing data were added, but cannot be advocated as a method for handling unobserved data as it relies on highly unrealistic assumptions.

Estimates from stochastic imputation and multiple imputation using chained equations produced almost identical estimates. Compared to CCA, absolute estimates of mortality from stochastic imputation and multiple imputation increased uniformly across all categories of BMI, whilst relative estimates indicated a weakening of the association between BMI and all-cause mortality, with the HR decreasing across all categories of BMI towards the null. Standard errors were smaller using stochastic imputation than multiple imputation though this difference was negligible, potentially due to the sample size. In this setting, it was less time-consuming than using multiple imputation to run the model in STATA. Handling 25 datasets for 72,000 patients was computationally demanding and modelling diagnostics took a significant length of time to run. This may deter epidemiologists with time constraints from applying multiple imputation in large routine clinical databases.

The finding that multiple imputation approaches may produce similar results to single imputation approaches and CCA is not uncommon. For example, multiple imputation was used to impute >10% (exact percentage not provided) unobserved data and produced estimates comparable to those obtained using CCA in a case-control study investigating the association between obesity and myocardial infarction in the routinely collected General Practice Research Database (Delaney et al., 2007). Similarly, a study using the same database to investigate the effect of hormone therapy on coronary heart disease reported estimates that were similar to CCA estimates when multiple imputation was applied to handle incomplete BMI, Systolic BP and smoking data for 20% of cases (Weiner et al., 2008).

Meanwhile, evidence from a study which used a range of imputation methods to handle unobserved serum cholesterol in 28 cohort studies indicated that single imputation approaches, including stochastic imputation produced similar results to multiple imputation when the percentage of missing data was less than 10% (Barzi & Woodward,

2004). For studies with more than 10% missing data, large differences in estimated mean cholesterol values and HRs for risk of coronary heart disease were reported for single and multiple imputation approaches. In a simulation study, estimates from stochastic imputation and multiple imputation were generally consistent until the proportion of missing data exceeded 15% (Marshall et al., 2010). However, stochastic imputation underestimated the variability even with 10% missingness. In our study, over 20% of cases had missing data, thus supporting the notion that the results from different studies cannot be generalised to other studies. The performance of these methods will not only depend on the proportion of unobserved data but also on the distribution of the covariates, the mechanisms of missingness and the relationship between the covariates and outcomes.

A selection-based sensitivity analysis was also applied to explore the robustness of estimates obtained from multiple imputation. Results from this approach were remarkably similar to those observed using stochastic imputation, suggesting estimates were robust to departures from the MAR assumption.

In Chapter 7 a range of methods were used to incorporate earlier or later patient BMI measurements in an effort to improve predictions of BMI at diagnosis. The majority of patients (83%) with missing BMI data at diagnosis had BMI measurements taken after the BMI time-frame. These measurements were used to improve the prediction of missing data at diagnosis and also minimise the risk that BMI recording was dependent on the value of BMI itself. The identification of approaches capable of imputing BMI values will be hugely beneficial to colleagues in Aberdeen who are currently investigating the effect of BMI trajectories after diagnosis on a number of health outcomes such as cardiovascular disease events. Despite the increasing complexity of each imputation approach, relative estimates of the effect of BMI on total mortality remained similar.

In Chapter 8, missing data methods were applied to a more complicated setting where missing data occurred in multiple variables. The proportion of patients with complete data for all variables to be included in the model of substantive interest was low (37%). Despite having different statistical underpinnings, three separate imputation models were applied and produced similar results. Like previous imputation models, the methods used in this work attenuated the risk of mortality in each of the BMI

categories. The multivariate normal multiple imputation model had the lowest effect on the observed estimates from CCA.

A study by Lee and Carlin (2012) reported that multiple imputation has the greatest value when missing data occur in confounders rather than the predictor of interest. When missing data occur in confounding variables only, the exclusion of these cases will lead to the loss of available information regarding the association of interest. However, when missing data occurred in the potential confounders in this analysis, estimates did not change considerably and results from multiple imputation approaches were largely similar for the risk of mortality whether missing data occurred in the predictor of interest only or missing data occurred in confounders and the risk of mortality.

9.2.1 STRENGTHS & LIMITATIONS

There are a number of strengths to this work. This project was able to use a large, population-based register of patients with DM which contain a large number of variables which can be used to provide explanations for the missing data and also to impute the unobserved data. Methods were applied to a real clinical database with real missing data problems and therefore the encountered issues are likely to be applicable to analyses using other clinical databases.

The primary limitation of this work is the fact that the true values of the observed data were unknown. Accordingly it was not possible to ascertain which method produced estimates that were closest to the truth. However, this approach ensured that problems faced were realistic and typical of those faced by epidemiologists. Nonetheless, access to a contemporary external dataset of patients with DM with complete data would have been beneficial to identify whether the distribution of BMI and other incomplete variables following imputation were comparable to distributions in this external dataset. Instead, comparisons were made with similar studies investigating this relationship.

In these analyses, the choice of missing data method was based upon each method's accessibility in standard statistical software and the availability of practical guidance in the literature. For this reason maximum likelihood methods for missing data were not applied in this project. Some authors recommend these approaches over multiple imputation since it is less labour intensive since all analyses are confined to a single

analysis. Unfortunately, maximum likelihood techniques require problem specific programmes and are therefore not readily available. As a consequence, these methods remain inaccessible to many epidemiologists and are rarely applied.

Similarly, the accessibility of MNAR methods prevented the thorough exploration of the robustness of MAR estimates to departures from this assumption. Unfortunately, it is not possible to rule out the possibility of data being MNAR using the data alone. Nonetheless, the inclusion of earlier and later measurements of BMI at diagnosis in multiple imputation approaches in Chapters 7 and 8 may have reduced the potential for missing data to be dependent on the value of the data point itself. Furthermore, according to the clinicians spoken to, there were no common characteristics of patients who failed to attend appointments. However, these clinicians were based in the Lothian HBA only and therefore this observation may not be generalisable to all patients with DM in Scotland.

9.2.2 IMPLICATIONS & FUTURE RESEARCH

Regardless of the complexity of the method used for handling missing data, relative estimates did not change substantially. Where relative estimates did change, the changes were generally not large enough to affect the overall conclusions.

Within this dataset, patients with unobserved values for important biochemical variables such as blood pressure, BMI and glycated haemoglobin had considerably worse survival than those who had observed data. This finding could have important implications for future research using the SCI-DC dataset. In particular, CCA may not be an appropriate method for handling unobserved data when absolute measures are reported. Though this finding is likely to have the largest impact on analyses investigating mortality trends, it may also introduce survival bias in analyses investigating other health outcomes. Furthermore, STROBEs reporting guidelines encourage the reporting of both absolute and relative measures of association and therefore this finding may have wider implications (Altman et al., 2007).

Whilst missing data occurring in variables involved in the relationship between BMI and all-cause mortality did not affect relative estimates, this may not be true for all analyses conducted using this dataset. In future analyses, due to the extent of the missing data

problem and the likelihood that data are not MCAR, it remains sensible to explore the missing data problem and report a secondary analysis investigating the robustness of estimates. This could be in the form of expanding the time-frame for measurement, as shown in Chapter 7 or by applying a more advanced approach for handling missing data such as imputation.

Expanding the time-frame for measurements of interest may be a simple and viable approach for handling unobserved data in repeatedly recorded variables but not for variables that are recorded once only or are only typically recorded due to their clinical relevance. This approach is also unlikely to be appropriate where measurements are likely to change considerably and will also over-estimate the precision of estimates. Further research is required to assess whether this approach may be an appropriate and applicable method that may be used as an alternative to more complicated missing data methods. Simulations within routinely collected datasets may be useful in informing the applicability of this approach for repeatedly measured variables.

Where this approach is not feasible or where there are concerns regarding precision estimates, more advanced approaches to handling unobserved data such as multiple imputation may be applied. Unfortunately, there remains a series of evidence gaps regarding the implementation of multiple imputation that may limit their use. Further work needs to assess the appropriateness of different imputation models in routine data. For example, two-fold FCS was designed to handle the complicated missing data problems observed in routine datasets. This approach has only recently been implemented in standard statistical software and the reliability of estimates from this approach requires further assessment, particularly whether this approach produces valid estimates when data are MAR. Data from this research has indicated that two-fold FCS produces largely similar estimates to MICE, MVN and PMM. It is necessary to identify whether this finding is typical of research using large routinely collected datasets.

Clarity also needs to be sought regarding the best practice for handling non-linear relationships and interactions in imputation models. In the analyses presented here, the continuous variable BMI was categorised prior to imputation to prevent bias due to non-linearity issues. In analyses investigating the association between BMI and mortality, BMI is traditionally categorised using the World Health Organisation's

classification system and therefore categorisation was deemed acceptable in this setting. In many other analyses, categorisation of an imputed variable may not be appropriate. There is currently no consensus on the method for handling non-linearities which will result in the least amount of downward bias. Currently, the Just Another Variable method has been proposed as the best approach for handling non-linear data but has been shown to produce invalid estimates when data are MAR (Seaman et al., 2012). Further simulation work is necessary to identify the most appropriate method.

9.2.3 SUMMARY

In this large routinely collected dataset, the choice of method for handling missing data had limited effect on the overall relative estimates. However, though alternative methods for CCA can be computationally intensive with many important practical considerations, it remains valuable to explore the robustness of estimates to departures from assumptions made by CCA. Where correctly applied, these methods are able to preserve the sample size whilst also reducing bias.

9.3 FINDINGS FROM THE ASSOCIATION BETWEEN BMI & ALL-CAUSE MORTALITY

In 2012, data from the Scottish Health Survey suggested that 62% of people in Scotland aged between 16 to 64 years were overweight or obese. Excess adiposity is an established risk factor for a number of adverse health outcomes including cardiovascular disease and type 2 DM. In the general population, obesity is associated with increased all-cause mortality (Adams et al., 2006; Pischon et al., 2008; Prospective Studies Collaboration, 2009; Chen et al., 2012). The relationship between BMI and mortality is far less well-established amongst patients with type 2 DM. Whilst most recent studies have reported an increased risk of mortality in people with BMI above 35 kg/m² (Carnethon et al., 2012; Logue et al., 2013), a number of early studies found no association between BMI and all-cause mortality (Morrish et al., 1990; Knuiman et al., 1992). However, each these studies failed to investigate the impact of missing data on these estimates.

The analyses presented in this thesis, after applying multiple different methods for handling unobserved data, are generally consistent with previous observations of lower mortality among overweight and obese patients relative to normal weight, thus providing further evidence of an obesity paradox. After adjustment for smoking status, presence of comorbid conditions, age at diagnosis, sex and deprivation status, risk of mortality in overweight and moderately obese patients were generally lower relative to normal weight patients, even after the exclusion of deaths occurring in the first two years of follow-up. These findings were consistent regardless of the method applied for handling missing data.

9.3.1 THE OBESITY PARADOX

The obesity paradox refers to the situation whereby normal weight patients have higher mortality than overweight and moderately obese patients (those with BMI of 25 to 35kg/m^2), in spite of the well-known association between elevated BMI and a number of adverse health outcomes. It has primarily been reported amongst adults with pre-existing chronic diseases such as heart failure, though a recent systematic review described decreased mortality in overweight people relative to normal weight in the general population (Flegal et al., 2013). In this review of prospective, observation cohort studies, overweight patients had lower all-cause mortality than normal weight participants (summary HR: 0.94 [0.91-0.96]). The main explanations for this observation included earlier presentation of overweight and obese patients and more rigorous medical treatment for non-normal weight patients. However, it may also reflect secular trends in BMI distribution and typically mortality is lowest in the group that forms the largest proportion of any population (Mehta & Chang, 2011).

In patients with DM, there is growing evidence of an obesity paradox (Hu et al., 2005; Mulnier et al., 2006; Carnethon et al., 2012; Logue et al., 2013; Thomas et al., 2014). Whilst plausible mechanisms for the obesity paradox in patients with some chronic diseases have been put forward including that for chronic kidney failure whereby weight losses are common as this disease progresses, a plausible mechanism for the observation amongst patients with DM is less obvious (Carnethon et al., 2014). Nonetheless, there have been several put forward.

- *Triggers for type 2 DM diagnosis*

The current procedures for the identification and diagnosis of people with Type 2 DM may play a role in the presence of the obesity paradox. Within the current NICE guidance document "Preventing type 2 diabetes: risk identification and interventions for individuals at high risk" it is recommended that GPs and other health professionals use a risk assessment to identify people of high risk of type 2 diabetes. According to this document, this risk assessment should be offered to:

- Adults aged 40 and above (excluding pregnant women)
- People aged 25-39 of South Asian, Chinese, African-Caribbean, black African and other high risk black and minority ethnic groups
- Adults with conditions that increase the risk of type 2 diabetes, including cardiovascular disease, hypertension, obesity, stroke, polycystic ovarian syndrome, a history of gestational diabetes and mental health problems.

The risk assessment is conducted using a validated risk assessment tool, such as the Diabetes UK risk score. This risk score uses information regarding patient age, sex, ethnicity, family history of DM, waist circumference, BMI and history of hypertension. Patients with a high risk score are then offered either a fasting blood glucose test or a glycated haemoglobin blood test to determine type 2 DM status. This strategy for identifying patients with type 2 DM, and in particular the recommendation that patients with obesity should be offered a risk assessment is likely to ensure overweight or obese patients to be diagnosed early. Whilst it is recommended that all adults aged 40 and above should be offered a risk assessment, GP time-constraints make it unlikely that this recommendation is routinely applied. Individuals with a lower BMI and without the risk factors described above would therefore only be offered an assessment and diagnostic tests if they were exhibiting symptoms suggestive of type 2 DM. As a consequence, patients diagnosed at lower BMIs may have a more severe form of DM or may have been diagnosed at a later stage of DM. These patients may therefore have a worse prognosis. For example, it has been shown that longer duration of type 2 DM is independently associated with elevated mortality and non-fatal DM complications (Huang et al., 2014). The obesity paradox may be a consequence of a longer period of

undiagnosed DM or more severe disease amongst low BMI patients. To test this hypothesis and to help in the interpretation of this relationship, the development of a tool for measuring DM severity at diagnosis may be beneficial. Accordingly, this score would identify those patients with the worst prognosis at diagnosis and analyses of the association between BMI and all-cause mortality could be adjusted for disease severity. Currently, there are several scores which attempt to classify DM severity, however these tools were not designed to measure disease severity at diagnosis and therefore may not include important factors, such as prescription data (Selby et al., 2001; Clarke et al., 2004; Young et al., 2008). Alternatively, the identification of patients with a more progressive form of type 2 DM may be achieved by observing rates of DM complications by BMI at diagnosis.

- *Heterogeneous DM*

Firstly, patients with type 2 DM are phenotypically heterogeneous. Previous studies have identified that lean and normal-weight patients with DM have different genetic profiles to overweight or obese patients with DM (Perry et al., 2012). These genetic profiles or other disease processes may predispose these individuals to DM at lower BMIs through a number of mechanisms including more rapid visceral fat accumulation, increased insulin resistance and earlier beta cell failure (Logue et al., 2013). If these genetic differences also confer increased susceptibility to other diseases or are a more aggressive form of DM, patients may exhibit worse survival. One such phenotype that has been identified is the metabolically obese normal weight (MONW) phenotype (Ruderman et al., 1998). MONW individuals have been observed to have higher visceral adiposity, higher blood pressure, lower physical activity expenditure, altered insulin sensitivity and are predisposed to cardiovascular diseases (Conus et al., 2007). Similarly, it is possible that some type 2 DM patients may have been misdiagnosed and actually have late onset autoimmune diabetes (LADA) which is phenotypically similar to type 1 DM. In a population of Swedish patients with clinically determined type 2 DM, it was observed that between 8 to 10% of patients may have LADA. The presence of these more aggressive pathophysiologies may have important management implications for people with newly-diagnosed DM. Treatment advice for normal-weight but metabolically unhealthy is likely to differ to advice offered to overweight patients.

- *Body composition & poor performance of BMI as a measure of obesity*

Secondly, patient body composition and the distribution of fat may be important contributors to the obesity paradox. BMI is a crude anthropometric measure of obesity which is unable to discriminate between adiposity and lean body mass. Though some patients may have low BMIs, these patients may have low muscle mass and high fat mass. Lean muscle mass is more insulin sensitive than fat mass and therefore sarcopenic obesity is metabolically disadvantageous (Hainer & Aldhoon-Hainerová, 2013). Sarcopenia also increases the frailty of patients which in turn is heavily associated with poor health outcomes including survival (Shamliyan et al., 2013). Therefore the use of BMI as a measure of adiposity may not be suitable amongst elderly populations as used here. Indeed, Zoppini et al. (2003) observed an obesity paradox only when analyses were restricted to patients over 65 years of age in a cohort of type 2 DM patients (n= 3,398) in Italy. However, no evidence of an interaction by age was observed in our study or in a pooled analysis of five longitudinal studies by Carnethon et al. (2012).

A related issue is BMIs inability to reflect the distribution of body fat. Visceral fat is known to be more strongly associated with insulin resistance than subcutaneous fat. When Carmethon et al. (2012) used waist circumference as the measure of visceral fat, the obesity paradox was no longer apparent, suggesting that the obesity paradox may, in part be explained by the limitations of BMI as a measure of adiposity. Similarly, a recent meta-analysis of studies investigating mortality amongst patients with coronary heart disease identified an obesity paradox when using BMI as a predictor but not when waist circumference was used (Coutinho et al., 2011). Unfortunately, waist circumference measurements are not routinely collected in SCI-DC (or in clinical practice and many routine datasets) and therefore this alternative hypothesis could not be explored using the current data.

BMI also does not provide any indication of physical fitness. A previous study from the Aerobics Center Longitudinal Study observed that obese but fit men with DM had lower cardiovascular disease mortality than lean and unfit patients (Church et al., 2005). Furthermore, an earlier study by the same group noted that cardiorespiratory fitness is a stronger determinant of mortality than weight status (Church et al., 2004). However, this study may be biased by the likely recruitment of people with an interest in healthy

lifestyles and fitness. Nonetheless, Kokkinos et al (2012) found that fitness was a strong predictor of all-cause mortality in patients with DM independent of BMI. Again, as for waist circumference, this information was not available in the SCI-DC register and is rarely available in clinical databases.

Other explanations of the obesity paradox have centred on methodological flaws of the studies investigating this relationship. These flaws include the likelihood of selection bias, treatment bias, residual confounding and reverse causation.

- *Treatment Bias*

Intensive DM management amongst overweight and obese patients may be contributing to a treatment bias which may contribute to reduced mortality in these groups (Standl et al., 2013). This form of bias is difficult to control for and may have arisen through a greater emphasis on stringent DM management of cardiovascular risk factors in overweight or obese patients. Unfortunately, assessing the potential for treatment bias is not straightforward. For example, using prescription data as an indicator of intensive risk factor recording is likely to produce biased estimates since prescriptions for statins or anti-hypertensive agents by categories of BMI are more likely to relate to the increased prevalence of hypertension or dyslipidemia in higher BMI categories.

- *Selection bias*

Collider stratification bias is a form of selection bias which has also been proposed as a potential explanation for the apparent protective influence of obesity amongst patients with chronic diseases (Dahabreh & Kent, 2011). Collider bias occurs when an association between two variables is distorted by conditioning on a common effect or collider, in this case, DM. By restricting analyses to patients with DM, a collision between obesity and smoking status occurs leading to a strengthening of the inverse association between smoking and obesity in this subpopulation. This in turn increases the likelihood that the association between obesity and mortality will be underestimated because of a failure to fully control for the effect of smoking. Preston and Stokes (2014) used data from the National Health and Nutritional Examination Survey to test this hypothesis in a population of DM patients. The authors presented an obesity paradox which was attenuated when smoking behaviour was controlled for but was eliminated

when former or current smokers were excluded. In my analyses presented in Chapter 8, when analyses were conducted on never smokers the obesity paradox was not eliminated when either CCA or multiple imputation were applied.

- *Residual confounding by poorly controlled smoking status*

Smoking status is strongly, inversely related to BMI and mortality (positively) and adjusting for this important confounder is difficult. Information on dose such as number of cigarettes smoked per day was not available and information regarding time since smoking cessation or smoking initiation amongst former and current smokers, respectively could not be incorporated into analyses. However, restricting analyses to people described as never smokers should have helped to overcome this problem.

- *Cross Sectional Studies*

Another proposed explanation for the paradox and a clear limitation of our study here is the cross-sectional nature of the exposure variable. This study and many others before it ignored the time-varying nature of the exposure and make the unlikely assumption that BMI remains constant throughout the entirety of follow-up. Upon diagnosis with DM, overweight and obese patients are more likely to make beneficial weight changes than normal weight patients and therefore exhibit healthier trajectories (Ferreira & Stehouwer, 2012). By making these changes, mortality risk may be underestimated for people who remain in the overweight and obese categories and fail to make such changes. Similarly, this approach of defining variables at a specific time-point meant that temporal changes in confounding factors such as patient smoking status could not be accounted for. Changes in these confounding factors may have contributed to the observed paradoxical finding in this study. Since repeated measurements are available in the SCI-DC dataset, further work using time-dependent variables should be focussed upon identifying trends in BMI following diagnosis with DM. Latent class trajectory analysis, as applied in the Whitehall II Cohort study to model pre-diagnosis BMI trajectories could be used to identify patterns of BMI change after diagnosis (Vistisen et al., 2014). The relationship between BMI trajectories and mortality could be explored. However, this may be difficult to achieve using data that are not measured at fixed periods.

- *Comorbidity*

As discussed previously, the presence of comorbid conditions may have influenced the findings of this study. Low BMI at diagnosis may have been a direct effect of co-morbidity such as chronic kidney disease which is well-known to cause muscle wasting and weight loss. These in turn are likely to hasten mortality. Attempts to control for reverse causation in this study included the exclusion of patients who had died within two years of BMI determination and those with comorbid conditions as defined by the Charlson comorbidity index. Despite these efforts, mortality risk remained high in the lowest weight categories. Nonetheless, this approach does not account for comorbidity not measured by hospital admission and depends on the reliability of the hospital admissions dataset (SMR01). In a recent quality assurance assessment conducted by the Information Services Division (ISD) it was reported that the accuracy rate of recordings was typically around 80% (Information Services Division, 2012). However, a number of issues with this dataset have been identified including its inability to register more than six comorbid conditions and the coding of signs and symptoms of conditions even if there is no diagnosis. Nonetheless, the binary coding of the Charlson comorbidity index should reduce the impact of failure to record all comorbid conditions. Furthermore, a systematic review of studies investigating the association between all-cause mortality and BMI also noted that exclusions based on the presence of comorbid conditions and deaths within 2 years of follow-up rarely have an effect on estimates (Flegal et al., 2013).

9.3.2 STRENGTHS & LIMITATIONS

There were a number of strengths to this work. Firstly, this is a population level dataset and therefore the sample was diverse and large enough with a sufficient number of events for detailed sub-group analyses. This feature also ensured the risk of selection biases were minimised. Additionally, SCI-DC is linked to numerous other clinical databases which were able to provide detailed information on some aspects of comorbidity status. This information in combination with relatively long term follow-up allowed for a reduction in the potential impact of pre-existing diseases. Death data were retrieved by register linkage with death registrations which are considerably more reliable than hospital discharge notes.

In addition, these analyses were adjusted for by deprivation status, the SIMD. This is an area-based measure of deprivation whose data-zones have an average population of less than 800 people. These data-zones should be relatively homogenous.

Along with those already discussed, the study had other potential limitations which may have obscured the true relationship between BMI and all-cause mortality. Firstly, this study was not able to adjust for a number of potentially important confounders such as ethnicity, alcohol intake, physical activity, diet, fat distribution and prescription data.

Ethnicity is very poorly recorded in the SCI-DC database and was not included in the analyses. Typically, Asian patients develop DM at lower BMIs than European populations, possibly through the high prevalence of the MONW phenotype in these populations (Florez & Castillo-Florez, 2012). Failure to stratify by ethnicity may therefore have influenced our studies findings. Nonetheless, according to the Scottish Diabetes Survey 2013, less than 4% of patients with type 2 DM in Scotland are from ethnic minority groups (Scottish Diabetes Survey Monitoring Group, 2013). The effect of ethnicity on estimates is therefore likely to be negligible.

Using all-cause mortality as the outcome of interest may have masked different patterns for different causes of death. By investigating cause-specific mortality it may have been possible to identify which specific causes of death were driving this obesity paradox. For example, the Prospective Studies Collaboration observed a U-shaped relationship between BMI and all-cause mortality in a general population, but different patterns of mortality by BMI were observed when cause-specific mortality was investigated (Prospective Studies Collaboration, 2009). Unsurprisingly, the excess mortality observed in the higher BMI categories was primarily driven by vascular deaths, particularly caused by ischaemic heart disease. In contrast, lung and aerodigestive cancer deaths and respiratory deaths exhibited strong inverse associations with BMI. Though it remains uncertain if these patterns would be observed in a DM population it is likely that the obesity paradox may be driven by certain diseases only.

The definition of BMI at diagnosis may also have led to incorrect estimates. Measurements taken within 10 months prior to or 2 months after diagnosis may not have truly reflected BMI at diagnosis. For example, body weight is influenced by treatment many patients may lose weight following diagnosis treatment effects within

this small time-period following diagnosis were deemed likely to be small. There is a small amount of evidence to suggest patient BMI typically changes in the year prior to diagnosis. Looker et al. (2001) observed that mean BMI rose between 0.43 and 0.71kg/m² in a population of Pima Indians in central Arizona though this was in a relatively small population (n=816) which was likely to differ considerably from the Scottish population. Vistisen et al. (2014) found that during 18 years of follow-up, patient BMIs typically increased steadily before decreasing a year prior to diagnosis. In contrast, De Fine Olivarius et al. (2008) observed an average weight loss of 2.3kg/m² in the year preceding diagnosis amongst patients enrolled in the Diabetes Care in General Practice study based in Denmark. In the same population, Drivsholm et al. (2005) reported that 61% of patients experienced unintentional weight loss in the three months preceding diagnosis with DM. Hyperglycaemia prior to the diagnosis of diabetes can result in weight loss so BMI at diagnosis of DM may not reflect pre-morbid BMI and duration of undiagnosed diabetes and severity of hyperglycaemia may confound the association between BMI and mortality. The majority of "BMI at diagnosis" measurements used in CCA analyses were recorded a short time after diagnosis (median time from diagnosis = 0 days (IQR: -11, 21 days). The definition of BMI at diagnosis was chosen on the basis of limiting the amount of incomplete data since missing data methods have been shown to produce progressively more biased results as the proportion of missing data increases (Graham, 2009). The short time between recording and diagnosis may also have limited the potential biases associated with using date of diagnosis rather than date of BMI determination as the starting point for follow-up. Furthermore, the stability of patient measurements during this time-frame was observed and deemed reasonable. In addition, estimates obtained using closest BMI values to replace unobserved measurements resulted in similar estimates to those obtained using CCA and multiple imputation. Using wider time-intervals therefore may offer a more straightforward approach to handling unobserved data.

A further limitation of the work was grouping together of patients with a BMI below 25kg/m². This limitation is discussed at length in section 6.4.3. Briefly, this categorisation strategy may have obscured the dose-response relationship between BMI and all-cause mortality since this group also included underweight patients. However, these categories were chosen due to concerns surrounding the number of patients with

a BMI below 20kg/m². Future work using this dynamic routinely collected data source will have larger numbers to explore this dose-response relationship in greater detail.

9.3.3 IMPLICATIONS & FUTURE RESEARCH

Weight loss in people with DM is thought to be desirable since it is assumed that weight reductions improve morbidity and mortality. Findings from this study suggest that for patients with a BMI above 35kg/m², weight loss should be encouraged. Nonetheless, current research on the effect of weight loss interventions on health outcomes amongst patients with type 2 DM have reported mixed findings (Williamson et al., 2000; Gregg et al., 2004; Bodegard et al., 2013; Wing et al., 2013). Further work is required to clarify the impact of weight changes following diagnosis with DM on patient health outcomes. A randomised controlled trial of weight loss interventions such as bariatric surgery or intensive weight management treatments is likely to be the most effective but costliest means of identifying the impact of weight changes following diagnosis with type 2 DM. As a cheaper alternative, an observational study using the SCI-DC dataset may be useful in assessing the association between BMI trajectories following diagnosis and mortality. To achieve this, methods used by Vistisen et al (2014) in the Whitehall II study could be applied. SCI-DCs data linkage with prescribing records will also allow adjustment for pharmacotherapies, an important confounder in this relationship.

Conversely, results presented here and in several other studies indicate that normal weight status at diagnosis with DM may be an important marker for raised mortality risk. In contrast, overweight and obese appear to have a protective effect on all-cause mortality. Further work is required to assess potential reasons for this obesity paradox and whether changes need to be made to current treatment plans for patients diagnosed at lower BMIs. More specifically, observing rates of type 2 DM complications by BMI at diagnosis may help in interpreting this relationship. If patients with lower BMIs at diagnosis develop complications earlier, this may be indicative of a more aggressive disease phenotype or a diagnosis at a later stage of the disease. Alternatively a measure of disease severity at diagnosis using baseline biochemical measures such as glycated haemoglobin could be a useful tool in explaining the paradox. If the paradox is a consequence of a later disease diagnosis adjustment for baseline disease severity in analyses of the association between BMI and mortality may attenuate risk of mortality at

lower BMIs. This measure of disease severity could subsequently be used to identify patients who would benefit from more aggressive DM treatment earlier.

Where no differences in disease severity by BMI at diagnosis are observed, alternative explanations for the obesity paradox should be investigated. For example, the observational study assessing the impact of BMI trajectories would also be able to identify whether the obesity paradox is a consequence of the cross-sectional study design.

9.3.4 SUMMARY

In summary, investigations of the association between BMI at diagnosis of type 2 DM and all-cause mortality in patients revealed a U-shaped relationship that remained even after exclusion deaths in the first two years, smokers and patients with comorbid conditions. Previous studies investigating this association have applied used a varied of approaches to handle missing data, such as CCA which may have been unsuitable and likely to introduce bias. Given the presence of an obesity paradox, it seemed bizarre that earlier studies have not explored the possibility that this paradox may have arisen due to inappropriate methods for handling unobserved data. This is the only study investigating this relationship in patients with DM where the robustness of estimates to different methods for handling missing data has been explored. In these analyses, the missing data method utilised did not have a large impact on the overall conclusions indicating other explanations for the paradox.

Further work is required to identify whether patients with different baseline BMI would benefit from different weight management plans.

10. BIBLIOGRAPHY

Abraham, W. T. and D. W. Russell (2004). "Missing data: a review of current methods and applications in epidemiological research." Current Opinion in Psychiatry **17**(4): 315-321.

Adams, K. F., A. Schatzkin, T. B. Harris, V. Kipnis, T. Mouw, R. Ballard-Barbash, A. Hollenbeck and M. F. Leitzmann (2006). "Overweight, Obesity, and Mortality in a Large Prospective Cohort of Persons 50 to 71 Years Old." New England Journal of Medicine **355**(8): 763-778.

Alberti, G., P. Zimmet, J. Shaw, Z. Bloomgarden, F. Kaufman and M. Silink (2004). "Type 2 Diabetes in the Young: The Evolving Epidemic: The International Diabetes Federation Consensus Workshop." Diabetes Care **27**(7): 1798-1811.

Alley, D. E., E. J. Metter, M. E. Griswold, T. B. Harris, E. M. Simonsick, D. L. Longo and L. Ferrucci (2010). "Changes in weight at the end of life: characterizing weight loss by time to death in a cohort study of older men." Am J Epidemiol **172**(5): 558-565.

Allison, P. D. (2000). Missing Data. Thousand Oaks, CA, Sage.

Allison, P. D. (2012). Handling Missing Data by Maximum Likelihood. Statistics and Data Analysis S. G. F. 2012. Haverford, PA.

Altman, D., M. Egge, P. C. Gotzsche, S. Pocock, J. P. Vandenbroucke and E. von Elm (2007). Strobe Statement.

Aucott, L., A. Poobalan, W. C. Smith, A. Avenell, R. Jung, J. Broom and A. M. Grant (2004). "Weight loss in obese diabetic and non-diabetic individuals and long-term diabetes outcomes--a systematic review." Diabetes Obes Metab **6**(2): 85-94.

Aucott, L., S. Philip, A. Avenell, N. Sattar and S. Wild (2013). "Patterns of Weight Change after diagnosis in patients with Type 2 diabetes and their relationship with metabolic and cardiovascular outcomes." Unpublished

Banack, H. R. and J. S. Kaufman (2013). "The "Obesity Paradox" Explained." Epidemiology **24**(3): 461-462 410.1097/EDE.1090b1013e31828c31776c.

Baraldi, A. N. and C. K. Enders (2010). "An introduction to modern missing data analyses." Journal of School Psychology **48**(1): 5-37.

Barnard, J. and X. L. Meng (1999). "Applications of multiple imputation in medical studies: from AIDS to NHANES." Statistical Methods In Medical Research **8**(1): 17-36.

Bartlett, J. W., S. R. Seaman, I. R. White, J. R. Carpenter and f. t. A. s. D. N. Initiative* (2014). "Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model." Statistical Methods in Medical Research.

Barzi, F. and M. Woodward (2004). "Imputations of missing values in practice: results from imputations of serum cholesterol in 28 cohort studies." Am J Epidemiol **160**(1): 34-45.

Bennett, D. A. (2001). "How can I deal with missing data in my study?" Australian and New Zealand Journal of Public Health **25**(5): 464-469.

Bernaards, C. A., T. R. Belin and J. L. Schafer (2007). "Robustness of a multivariate normal approximation for imputation of incomplete binary data." Stat Med **26**(6): 1368-1382.

Berrington de Gonzalez, A., P. Hartge, J. R. Cerhan, A. J. Flint, L. Hannan, R. J. MacInnis, S. C. Moore, G. S. Tobias, H. Anton-Culver, L. B. Freeman, W. L. Beeson, S. L. Clipp, D. R. English, A. R. Folsom, D. M. Freedman, G. Giles, N. Hakansson, K. D. Henderson, J. Hoffman-Bolton, J. A. Hoppin, K. L. Koenig, I.-M. Lee, M. S. Linet, Y. Park, G. Pocobelli, A. Schatzkin, H. D. Sesso, E. Weiderpass, B. J. Willcox, A. Wolk, A. Zeleniuch-Jacquotte, W. C. Willett and M. J. Thun (2010). "Body-Mass Index and Mortality among 1.46 Million White Adults." New England Journal of Medicine **363**(23): 2211-2219.

Bessonova, L., S. F. Marshall, A. Ziogas, J. Largent, L. Bernstein, K. D. Henderson, H. Ma, D. W. West and H. Anton-Culver (2011). "The association of body mass index with mortality in the California Teachers Study." Int J Cancer **129**(10): 2492-2501.

Bhaskaran, K., H. J. Forbes, I. Douglas, D. A. Leon and L. Smeeth (2013). "Representativeness and optimal use of body mass index (BMI) in the UK Clinical Practice Research Datalink (CPRD)." BMJ Open **3**(9).

Bodegard, J., J. Sundstrom, B. Svennblad, C. J. Ostgren, P. M. Nilsson and G. Johansson (2013). "Changes in body mass index following newly diagnosed type 2 diabetes and risk of cardiovascular mortality: a cohort study of 8486 primary-care patients." Diabetes Metab **39**(4): 306-313.

Bodner, T. E. (2008). "What Improves with Increased Missing Data Imputations?" Structural Equation Modeling: A Multidisciplinary Journal **15**(4): 651-675.

Cao, C., R. Wang, J. Wang, H. Bunjhoo, Y. Xu and W. Xiong (2012). "Body Mass Index and Mortality in Chronic Obstructive Pulmonary Disease: A Meta-Analysis." PLoS ONE **7**(8): e43892.

Carnethon, M., L. Rasmussen-Torvik and L. Palaniappan (2014). "The Obesity Paradox in Diabetes." Current Cardiology Reports **16**(2): 1-7.

Carnethon, M. R., P. J. De Chavez, M. L. Biggs, C. E. Lewis, J. S. Pankow, A. G. Bertoni, S. H. Golden, K. Liu, K. J. Mukamal, B. Campbell-Jenkins and A. R. Dyer (2012). "Association of weight status with mortality in adults with incident diabetes." Jama **308**(6): 581-590.

Carpenter, J., M. Kenward, S. Evans and I. White (2004). "Last observation carry-forward and last observation analysis." Statistics in Medicine **23**(20): 3241-3242.

Carpenter, J. R. and M. G. Kenward (2007). Missing data in randomised controlled trials - a practical guide.

Carpenter, J. R., M. G. Kenward and I. R. White (2007). "Sensitivity analysis after multiple imputation under missing at random: a weighting approach." Statistical Methods In Medical Research **16**(3): 259-275.

Chan, J. M., E. B. Rimm, G. A. Colditz, M. J. Stampfer and W. C. Willett (1994). "Obesity, Fat Distribution, and Weight Gain as Risk Factors for Clinical Diabetes in Men." Diabetes Care **17**(9): 961-969.

Charlson, M. E., P. Pompei, K. L. Ales and C. R. MacKenzie (1987). "A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation." Journal of Chronic Diseases **40**(5): 373-383.

Chaudhry, Z. W., M. C. Gannon and F. Q. Nuttall (2006). "Stability of Body Weight in Type 2 Diabetes." Diabetes Care **29**(3): 493-497.

Chen, Z., G. Yang, A. Offer, M. Zhou, M. Smith, R. Peto, H. Ge, L. Yang and G. Whitlock (2012). "Body mass index and mortality in China: a 15-year prospective study of 220 000 men." Int J Epidemiol **41**(2): 472-481.

Chiu, C. J., L. A. Wray, F. H. Lu and E. A. Beverly (2013). "BMI change patterns and disability development of middle-aged adults with diabetes: a dual trajectory modeling approach." J Gen Intern Med **28**(9): 1150-1156.

Church, T. S., Y. J. Cheng, C. P. Earnest, C. E. Barlow, L. W. Gibbons, E. L. Priest and S. N. Blair (2004). "Exercise capacity and body composition as predictors of mortality among men with diabetes." Diabetes Care **27**(1): 83-88.

Church, T. S., M. J. LaMonte, C. E. Barlow and S. N. Blair (2005). "Cardiorespiratory fitness and body mass index as predictors of cardiovascular disease mortality among men with diabetes." Archives of Internal Medicine **165**(18): 2114-2120.

Clarke, P. M., A. M. Gray, A. Briggs, A. J. Farmer, P. Fenn, R. J. Stevens, D. R. Matthews, I. M. Stratton and R. R. Holman (2004). "A model to estimate the lifetime health outcomes of patients with type 2 diabetes: the United Kingdom Prospective Diabetes Study (UKPDS) Outcomes Model (UKPDS no. 68)." Diabetologia **47**(10): 1747-1759.

Connolly, V. M. and C. M. Kesson (1996). "Socioeconomic Status and Clustering of Cardiovascular Disease Risk Factors in Diabetic Patients." Diabetes Care **19**(5): 419-422.

Conus, F., R. Rabasa-Lhoret and F. Peronnet (2007). "Characteristics of metabolically obese normal-weight (MONW) subjects." Appl Physiol Nutr Metab **32**(1): 4-12.

Cooper, R. S. (2008). "Which factors confound or modify the relationship between body weight and mortality[quest]." Int J Obes **32**(S3): S47-S51.

Corrada, M. a. M., C. H. Kawas, F. Mozaffar and A. Paganini-Hill (2006). "Association of Body Mass Index and Weight Change with All-Cause Mortality in the Elderly." American Journal of Epidemiology **163**(10): 938-949.

Coutinho, T., K. Goel, D. Corrêa de Sá, C. Kragelund, A. M. Kanaya, M. Zeller, J.-S. Park, L. Kober, C. Torp-Pedersen, Y. Cottin, L. Lorgis, S.-H. Lee, Y.-J. Kim, R. Thomas, V. L. Roger, V. K. Somers and F. Lopez-Jimenez (2011). "Central Obesity and Survival in Subjects With Coronary Artery Disease: A Systematic Review of the Literature and Collaborative Analysis With Individual Subject Data." Journal of the American College of Cardiology **57**(19): 1877-1886.

Dahabreh, I. J. and D. M. Kent (2011). "Index event bias as an explanation for the paradoxes of recurrence risk research." Jama **305**(8): 822-823.

de Fine Olivarius, N., A. H. Andreasen, V. Siersma, B. Richelsen and H. Beck-Nielsen (2006). "Changes in patient weight and the impact of antidiabetic therapy during the first 5 years after diagnosis of diabetes mellitus." Diabetologia **49**(9): 2058-2067.

De Fine Olivarius, N., B. Richelsen, V. Siersma, A. H. Andreasen and H. Beck-Nielsen (2008). "Weight history of patients with newly diagnosed Type 2 diabetes." Diabetic Medicine **25**(8): 933-941.

Delaney, J. A., S. S. Daskalopoulou, J. M. Brophy, R. J. Steele, L. Opatrny and S. Suissa (2007). "Lifestyle variables and the risk of myocardial infarction in the general practice research database." BMC Cardiovasc Disord **7**: 38.

Demirtas, H., S. A. Freels and R. M. Yucel (2008). "Plausibility of multivariate normality assumption when multiply imputing non-Gaussian continuous outcomes: a simulation assessment." Journal of Statistical Computation and Simulation **78**(1): 69-84.

Diabetes UK Scotland (2010). Diabetes Action Plan 2010. Quality Care for Diabetes in Scotland.

Donders, A. R., G. J. van der Heijden, T. Stijnen and K. G. Moons (2006). "Review: a gentle introduction to imputation of missing values." J Clin Epidemiol **59**(10): 1087-1091.

Drivsholm, T., N. de Fine Olivarius, A. B. S. Nielsen and V. Siersma (2005). "Symptoms, signs and complications in newly diagnosed type 2 diabetic patients, and their relationship to glycaemia, blood pressure and weight." Diabetologia **48**(2): 210-214.

Dugas, M., M. Lange, C. Müller-Tidow, P. Kirchhof and H.-U. Prokosch (2010). "Routine data from hospital information systems can support patient recruitment for clinical studies." Clinical Trials **7**(2): 183-189.

Eeg-Olofsson, K., J. Cederholm, P. M. Nilsson, B. Zethelius, L. Nunez, S. Gudbjornsdottir and B. Eliasson (2009). "Risk of cardiovascular disease and mortality in overweight and obese patients with type 2 diabetes: an observational study in 13,087 patients." Diabetologia **52**(1): 65-73.

Elliott, P. and G. Hawthorne (2005). "Imputing missing repeated measures data: how should we proceed?" Australian and New Zealand Journal of Psychiatry **39**(7): 575-582.

Enders, C. K. (2010). Applied Missing Data Analysis. New York, Guildford Press.

Engels, J. M. and P. Diehr (2003). "Imputation of missing longitudinal data: a comparison of methods." Journal of Clinical Epidemiology **56**(10): 968-976.

European Medicines Agency (2010). Guideline on missing data in confirmatory clinical trials.

Faris, P. D., W. A. Ghali, R. Brant, C. M. Norris, P. D. Galbraith and M. L. Knudtson (2002). "Multiple imputation versus data enhancement for dealing with missing data in observational health care outcome analyses." Journal of Clinical Epidemiology **55**(2): 184-191.

Ferreira, I. and C. D. Stehouwer (2012). "Obesity paradox or inappropriate study designs? Time for life-course epidemiology." J Hypertens **30**(12): 2271-2275.

Fisher, M. (2006). "Macrovascular disease in diabetes." Medicine **34**(3): 101-103.

Flegal, K. M., B. I. Graubard, D. F. Williamson and M. H. Gail (2005). "Excess deaths associated with underweight, overweight, and obesity." Jama **293**(15): 1861-1867.

Flegal, K. M., B. K. Kit, H. Orpana and B. I. Graubard (2013). "Association of all-cause mortality with overweight and obesity using standard body mass index categories: A systematic review and meta-analysis." Jama **309**(1): 71-82.

Florez, H. and S. Castillo-Florez (2012). "Beyond the obesity paradox in diabetes: Fitness, fatness, and mortality." Jama **308**(6): 619-620.

Fonarow, G. C., P. Srikanthan, M. R. Costanzo, G. B. Cintron and M. Lopatin (2007). "An obesity paradox in acute heart failure: analysis of body mass index and inhospital mortality for 108,927 patients in the Acute Decompensated Heart Failure National Registry." Am Heart J **153**(1): 74-81.

Ford, I., H. Murray, C. J. Packard, J. Shepherd, P. W. Macfarlane and S. M. Cobbe (2007). "Long-Term Follow-up of the West of Scotland Coronary Prevention Study." New England Journal of Medicine **357**(15): 1477-1486.

Gadbury, G. L., C. S. Coffey and D. B. Allison (2003). "Modern statistical methods for handling missing repeated measurements in obesity trial data: beyond LOCF." Obesity Reviews **4**(3): 175-184.

Gorber, S. C., S. Schofield-Hurwitz, J. Hardt, G. Levasseur and M. Tremblay (2009). "The accuracy of self-reported smoking: A systematic review of the relationship between self-reported and cotinine-assessed smoking status." Nicotine & Tobacco Research **11**(1): 12-24.

Gorber, S. C., M. Tremblay, D. Moher and B. Gorber (2007). "A comparison of direct vs. self-report measures for assessing height, weight and body mass index: a systematic review." Obesity Reviews **8**(4): 307-326.

Graham, J. W. (2009). "Missing Data Analysis: Making It Work in the Real World." Annual Review of Psychology **60**(1): 549-576.

Graham, J. W. (2012). Missing data: Analysis and design, Springer.

Graham, J. W., A. E. Olchowski and T. D. Gilreath (2007). "How many imputations are really needed? Some practical clarifications of multiple imputation theory." Prev Sci **8**(3): 206-213.

Gregg, E. W., R. B. Gerzoff, T. J. Thompson and D. F. Williamson (2004). "Trying to lose weight, losing weight, and 9-year mortality in overweight U.S. adults with diabetes." Diabetes Care **27**(3): 657-662.

Hainer, V. and I. Aldhoon-Hainerová (2013). "Obesity Paradox Does Exist." Diabetes Care **36**(Supplement 2): S276-S281.

Hardt, J., M. Herke and R. Leonhart (2012). "Auxiliary variables in multiple imputation in regression with missing X: a warning against including too many in small sample research." BMC Medical Research Methodology **12**(1): 184.

He, Y. (2010). "Missing data analysis using multiple imputation: getting to the heart of the matter." Circ Cardiovasc Qual Outcomes **3**(1): 98-105.

Hemminki, E., K. Heikkilä, T. Sevón and P. Koponen (2008). "Special features of health services and register based trials - experiences from a randomized trial of childbirth classes." BMC Health Services Research **8**(1): 1-8.

Heraud-Bousquet, V., C. Larsen, J. Carpenter, J. C. Desenclos and Y. Le Strat (2012). "Practical considerations for sensitivity analysis after multiple imputation applied to epidemiological studies with incomplete data." BMC Med Res Methodol **12**: 73.

Hippisley-Cox, J., C. Coupland, Y. Vinogradova, J. Robson, M. May and P. Brindle (2007). Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study.

Hippisley-Cox, J., C. Coupland, Y. Vinogradova, J. Robson, M. May and P. Brindle (2007). "Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study." BMJ **335**(7611): 136.

Horton, N. J. and K. P. Kleinman (2007). "Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models." Am Stat **61**(1): 79-90.

Horton, N. J., S. R. Lipsitz and M. Parzen (2003). "A Potential for Bias When Rounding in Multiple Imputation." The American Statistician **57**(4): 229-232.

Hotchkiss, J. W. and A. H. Leyland (2011). "The relationship between body size and mortality in the linked Scottish Health Surveys: cross-sectional surveys with follow-up." Int J Obes (Lond) **35**(6): 838-851.

Hu, G., P. Jousilahti, N. C. Barengo, Q. Qiao, T. A. Lakka and J. Tuomilehto (2005). "Physical Activity, Cardiovascular Risk Factors, and Mortality Among Finnish Adults With Diabetes." Diabetes Care **28**(4): 799-805.

Huang, E. S., N. Laiteerapong, J. Y. Liu, P. M. John, H. H. Moffet and A. J. Karter (2014). "Rates of complications and mortality in older patients with diabetes mellitus: The diabetes and aging study." JAMA Internal Medicine **174**(2): 251-258.

Hughes, V. (2013). "The big fat truth." Nature **497**(7450): 428-430.

Information Services Division (2012). Assessment of SMR01 Data 2010-2011. Scotland Report.

Janghorbani, M. and M. Amini (2009). "Patterns and predictors of long-term weight change in patients with type 2 diabetes mellitus." Ann Nutr Metab **54**(2): 111-118.

Janssen, K. J. M., A. R. T. Donders, F. E. Harrell Jr, Y. Vergouwe, Q. Chen, D. E. Grobbee and K. G. M. Moons (2010). "Missing covariate data in medical research: To impute is better than to ignore." Journal of Clinical Epidemiology **63**(7): 721-727.

Jeličić, H., E. Phelps and R. Lerner (2010). "Why Missing Data Matter in the Longitudinal Study of Adolescent Development: Using the 4-H Study to Understand the Uses of Different Missing Data Methods." Journal of Youth and Adolescence **39**(7): 816-835.

Johnson, D. R. and R. Young (2011). "Toward Best Practices in Analyzing Datasets with Missing Data: Comparisons and Recommendations." Journal of Marriage and Family **73**(5): 926-945.

Kannel, W. B., N. Brand, J. J. J. Skinner, T. R. Dawber and P. M. McNamara (1967). "The Relation of Adiposity to Blood Pressure and Development of HypertensionThe Framingham Study." Annals of Internal Medicine **67**(1): 48-59.

Kenward, M. G. and J. Carpenter (2007). "Multiple imputation: current perspectives." Statistical Methods In Medical Research **16**(3): 199-218.

Kenward, M. G. and G. Molenberghs (2009). "Last observation carried forward: a crystal ball?" Journal of Biopharmaceutical Statistics **19**(5): 872-888.

Khalangot, M., M. Tronko, V. Kravchenko, J. Kulchinska and G. Hu (2009). "Body mass index and the risk of total and cardiovascular mortality among patients with type 2 diabetes: a large prospective study in Ukraine." Heart **95**(6): 454-460.

Klebanoff, M. A. and S. R. Cole (2008). "Use of Multiple Imputation in the Epidemiologic Literature." American Journal of Epidemiology **168**(4): 355-357.

Knuiman, M. W., T. A. Welborn and D. E. Whittall (1992). "An analysis of excess mortality rates for persons with non-insulin-dependent diabetes mellitus in Western

Australia using the Cox proportional hazards regression model." Am J Epidemiol **135**(6): 638-648.

Kokkinos, P., J. Myers, C. Faselis, M. Doumas, R. Kheirbek and E. Nylen (2012). "BMI–Mortality Paradox and Fitness in African American and Caucasian Men With Type 2 Diabetes." Diabetes Care.

Kropko, J., B. Goodrich, A. Gelman and J. Hill (2014). "Multiple Imputation for Continuous and Categorical Data: Comparing Joint Multivariate Normal and Conditional Approaches." Political Analysis.

Langkamp, D. L., A. Lehman and S. Lemeshow (2010). "Techniques for Handling Missing Data in Secondary Analyses of Large Surveys." Academic Pediatrics **10**(3): 205-210.

Lean, M. E., C. Katsarou, P. McLoone and D. S. Morrison (2012). "Changes in BMI and waist circumference in Scottish adults: use of repeated cross-sectional surveys to explore multiple age groups and birth-cohorts." Int J Obes (Lond).

Lee, K. J. and J. B. Carlin (2010). "Multiple Imputation for Missing Data: Fully Conditional Specification Versus Multivariate Normal Imputation." American Journal of Epidemiology **171**(5): 624-632.

Lee, K. J. and J. B. Carlin (2012). "Recovery of information from multiple imputation: a simulation study." Emerg Themes Epidemiol **9**(1): 3.

Lee, K. J., J. C. Galati, J. A. Simpson and J. B. Carlin (2012). "Comparison of methods for imputing ordinal data using multivariate normal imputation: a case study of non-linear effects in a large cohort study." Statistics in Medicine **31**(30): 4164-4174.

Lewsey, J. D., A. H. Leyland, G. D. Murray and F. A. Boddy (2000). "Using routine data to complement and enhance the results of randomised controlled trials." Health Technol Assess **4**(22): 1-55.

Lipscombe, L. L. and J. E. Hux (2007). "Trends in diabetes prevalence, incidence, and mortality in Ontario, Canada 1995-2005: a population-based study." Lancet **369**(9563): 750-756.

Little, R. J. A. and D. B. Rubin (1987). Statistical analysis with missing data. New Jersey, John Wiley & Sons.

Little, R. J. A. and D. B. Rubin (2002). Statistical Analysis with Missing Data. New York, John Wiley.

Logue, J. and N. Sattar (2013). "Body weight and mortality in type 2 diabetes: weighing up the evidence." Diabetes Metab **39**(4): 287-288.

Logue, J., J. J. Walker, H. M. Colhoun, G. P. Leese, R. S. Lindsay, J. A. McKnight, A. D. Morris, D. W. Pearson, J. R. Petrie, S. Philip, S. H. Wild and N. Sattar (2011). "Do men develop type 2 diabetes at lower body mass indices than women?" Diabetologia **54**(12): 3003-3006.

Logue, J., J. J. Walker, G. Leese, R. Lindsay, J. McKnight, A. Morris, S. Philip, S. Wild and N. Sattar (2013). "Association between BMI measured within a year after diagnosis of type 2 diabetes and mortality." Diabetes Care **36**(4): 887-893.

Looker, H. C., W. C. Knowler and R. L. Hanson (2001). "Changes in BMI and Weight Before and After the Development of Type 2 Diabetes." Diabetes Care **24**(11): 1917-1922.

Mackinnon, A. (2010). "The use and reporting of multiple imputation in medical research – a review." Journal of Internal Medicine **268**(6): 586-593.

Mallinckrodt, C. H., T. M. Sanger, S. Dubé, D. J. DeBrota, G. Molenberghs, R. J. Carroll, W. Z. Potter and G. D. Tollefson (2003). "Assessing and interpreting treatment effects in longitudinal clinical trials with missing data." Biological Psychiatry **53**(8): 754-760.

Manson, J. E., M. J. Stampfer, C. H. Hennekens and W. C. Willett (1987). "Body weight and longevity: A reassessment." Jama **257**(3): 353-358.

Marshall, A., D. Altman, P. Royston and R. Holder (2010). "Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study." BMC Medical Research Methodology **10**(1): 1-16.

Marston, L., J. R. Carpenter, K. R. Walters, R. W. Morris, I. Nazareth and I. Petersen (2010). "Issues in multiple imputation of missing data for large general practice clinical databases." Pharmacoepidemiology and Drug Safety **19**(6): 618-626.

Marston, L., J. R. Carpenter, K. R. Walters, R. W. Morris, I. Nazareth, I. R. White and I. Petersen (2014). "Smoker, ex-smoker or non-smoker? The validity of routinely recorded smoking status in UK primary care: a cross-sectional study." BMJ Open **4**(4).

Mathur, R., K. Bhaskaran, N. Chaturvedi, D. A. Leon, T. Vanstaa, E. Grundy and L. Smeeth (2013). "Completeness and usability of ethnicity data in UK-based primary care and hospital databases." J Public Health (Oxf).

McKnight, P. E., K. M. McKnight, S. Sidani and A. J. FFigueredo (2007). Missing Data. A Gentle Introduction. New York, The Guildford Press.

McLaren, L. (2007). "Socioeconomic status and obesity." Epidemiol Rev **29**: 29-48.

Mehta, N. K. and V. W. Chang (2011). "Secular declines in the association between obesity and mortality in the United States." Popul Dev Rev **37**(3): 435-451.

Molenberghs, G., H. Thijs, I. Jansen, C. Beunckens, M. G. Kenward, C. Mallinckrodt and R. J. Carroll (2004). "Analyzing incomplete longitudinal clinical trial data." Biostatistics **5**(3): 445-464.

Morris, T. P., I. R. White and P. Royston (2014). "Tuning multiple imputation by predictive mean matching and local residual draws." BMC Med Res Methodol **14**: 75.

Morris, T. P., I. R. White, P. Royston, S. R. Seaman and A. M. Wood (2014). "Multiple imputation for an incomplete covariate that is a ratio." Stat Med **33**(1): 88-104.

Morrish, N. J., L. K. Stevens, J. Head, J. H. Fuller, R. J. Jarrett and H. Keen (1990). "A prospective study of mortality among middle-aged diabetic patients (the London Cohort of the WHO Multinational Study of Vascular Disease in Diabetics) I: Causes and death rates." Diabetologia **33**(9): 538-541.

Mulnier, H. E., H. E. Seaman, V. S. Raleigh, S. S. Soedamah-Muthu, H. M. Colhoun and R. A. Lawrenson (2006). "Mortality in people with Type 2 diabetes in the UK." Diabetic Medicine **23**(5): 516-521.

National Records of Scotland (2013). Mid 2011 and Mid-2012 Population Estimates Scotland. G. R. O. f. Scotland. Edinburgh.

Nevalainen, J., M. G. Kenward and S. A. Virtanen (2009). "Missing values in longitudinal dietary data: A multiple imputation approach based on a fully conditional specification." Statistics in Medicine **28**(29): 3657-3669.

Newman, D. A. (2003). "Longitudinal Modeling with Randomly and Systematically Missing Data: A Simulation of Ad Hoc, Maximum Likelihood, and Multiple Imputation Techniques." Organizational Research Methods **6**(3): 328-362.

Nichols, G. A. and A. Gomez-Caminero (2007). "Weight changes following the initiation of new anti-hyperglycaemic therapies." Diabetes Obes Metab **9**(1): 96-102.

Nilsson, P. M. (2008). "Is weight loss beneficial for reduction of morbidity and mortality? What is the controversy about?" Diabetes Care **31 Suppl 2**: S278-283.

Norris, C. M., W. A. Ghali, M. L. Knudtson, C. D. Naylor and L. D. Saunders (2000). "Dealing with missing data in observational health care outcome analyses." Journal of Clinical Epidemiology **53**(4): 377-383.

Office of the Chief Statistician (2004). Analysis of Ethnicity in the 2001 Census - Summary Report.

Oh, D., J. Heck, C. Dresler, S. Allwright, M. Haglund, S. Del Mazo, E. Kralikova, I. Stucker, E. Tamang, E. Gritz and M. Hashibe (2010). "Determinants of smoking initiation among women in five European countries: a cross-sectional survey." BMC Public Health **10**(1): 1-11.

Oreopoulos, A., R. Padwal, K. Kalantar-Zadeh, G. C. Fonarow, C. M. Norris and F. A. McAlister (2008). "Body mass index and mortality in heart failure: a meta-analysis." Am Heart J **156**(1): 13-22.

Ovbiagele, B., P. M. Bath, D. Cotton, R. Vinisko and H. C. Diener (2011). "Obesity and recurrent vascular risk after a recent ischemic stroke." Stroke **42**(12): 3397-3402.

Perry, J. R., B. F. Voight, L. Yengo, N. Amin, J. Dupuis, M. Ganser, H. Grallert, P. Navarro, M. Li, L. Qi, V. Steinthorsdottir, R. A. Scott, P. Almgren, D. E. Arking, Y. Aulchenko, B. Balkau, R. Benediktsson, R. N. Bergman, E. Boerwinkle, L. Bonnycastle, N. P. Burtt, H. Campbell, G. Charpentier, F. S. Collins, C. Gieger, T. Green, S. Hadjadj, A. T. Hattersley, C. Herder, A. Hofman, A. D. Johnson, A. Kottgen, P. Kraft, Y. Labrune, C. Langenberg, A. K. Manning, K. L. Mohlke, A. P. Morris, B. Oostra, J. Pankow, A. K. Petersen, P. P. Pramstaller, I. Prokopenko, W. Rathmann, W. Rayner, M. Roden, I. Rudan, D. Rybin, L. J. Scott, G. Sigurdsson, R. Sladek, G. Thorleifsson, U. Thorsteinsdottir, J. Tuomilehto, A. G. Uitterlinden, S. Vivequin, M. N. Weedon, A. F. Wright, F. B. Hu, T. Illig, L. Kao, J. B. Meigs, J. F. Wilson, K. Stefansson, C. van Duijn, D. Altschuler, A. D. Morris, M. Boehnke, M. I. McCarthy, P. Froguel, C. N. Palmer, N. J. Wareham, L. Groop, T. M. Frayling and S. Cauchi (2012). "Stratifying type 2 diabetes cases by BMI identifies genetic risk variants in LAMA1 and enrichment for risk variants in lean compared to obese cases." PLoS Genet **8**(5): e1002741.

Peugh, J. L. and C. K. Enders (2004). "Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement." Review of Educational Research **74**(4): 525-556.

Pischon, T., H. Boeing, K. Hoffmann, M. Bergmann, M. B. Schulze, K. Overvad, Y. T. van der Schouw, E. Spencer, K. G. M. Moons, A. TjÃ¸nneland, J. Halkjaer, M. K. Jensen, J. Stegger, F. Clavel-Chapelon, M. C. Boutron-Ruault, V. Chajes, J. Linseisen, R. Kaaks, A. Trichopoulou, D. Trichopoulos, C. Bamia, S. Sieri, D. Palli, R. Tumino, P. Vineis, S. Panico, P. H. M. Peeters, A. M. May, H. B. Bueno-de-Mesquita, F. J. B. van Duijnhoven, G. Hallmans, L. Weinehall, J. Manjer, B. Hedblad, E. Lund, A. Agudo, L. Arriola, A. Barricarte, C. Navarro, C. Martinez, J.

R. QuirÃ³s, T. Key, S. Bingham, K. T. Khaw, P. Boffetta, M. Jenab, P. Ferrari and E. Riboli (2008). "General and Abdominal Adiposity and Risk of Death in Europe." New England Journal of Medicine **359**(20): 2105-2120.

Preis, S. R., M. J. Pencina, S.-J. Hwang, R. B. D'Agostino, P. J. Savage, D. Levy and C. S. Fox (2009). "Trends in Cardiovascular Disease Risk Factors in Individuals With and Without Diabetes Mellitus in the Framingham Heart Study." Circulation **120**(3): 212-220.

Preston, S. H. and A. Stokes (2014). "Obesity paradox: conditioning on disease enhances biases in estimating the mortality risks of obesity." Epidemiology **25**(3): 454-461.

Prospective Studies Collaboration (2009). "Body-mass index and cause-specific mortality in 900,000 adults: collaborative analyses of 57 prospective studies." The Lancet **373**(9669): 1083-1096.

Qi, L., Y. F. Wang and Y. He (2010). "A comparison of multiple imputation and fully augmented weighted estimators for Cox regression with missing covariates." Stat Med **29**(25): 2592-2604.

Quan, H., V. Sundararajan, P. Halfon, A. Fong, B. Burnand, J. C. Luthi, L. D. Saunders, C. A. Beck, T. E. Feasby and W. A. Ghali (2005). "Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data." Med Care **43**(11): 1130-1139.

Rothwell, P. M. (2005). "External validity of randomised controlled trials: ?To whom do the results of this trial apply??" The Lancet **365**(9453): 82-93.

Roubenoff, R. (1999). "The Pathophysiology of Wasting in the Elderly." The Journal of Nutrition **129**(1): 256S-259S.

Royston, P. (2004). "Multiple Imputation of Missing Values." The Stata Journal **4**: 227-241.

Royston, P. (2005). "Multiple imputation of missing values: Update of ice." Stata Journal **5**(4): 527-536.

Rubin, D. B. (1987). "Multiple Imputation for Nonresponse in Surveys (Wiley Series in Probability and Statistics)."

Rubin, D. B. (1996). "Multiple Imputation After 18+ Years." Journal of the American Statistical Association **91**(434): 473-489.

Rubin, D. B. (2009). Multiple imputation for nonresponse in surveys, Wiley. com.

Ruderman, N., D. Chisholm, X. Pi-Sunyer and S. Schneider (1998). "The metabolically obese, normal-weight individual revisited." Diabetes **47**(5): 699-713.

Sarwar, N., P. Gao, S. R. Seshasai, R. Gobin, S. Kaptoge, E. Di Angelantonio, E. Ingelsson, D. A. Lawlor, E. Selvin, M. Stampfer, C. D. Stehouwer, S. Lewington, L. Pennells, A. Thompson, N. Sattar, I. R. White, K. K. Ray and J. Danesh (2010). "Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies." Lancet **375**(9733): 2215-2222.

Schafer, J. L. (1997). Analysis of Incomplete Multivariate Data. London, UK, Chapman & Hall Ltd.

Schafer, J. L. (1999). "Multiple imputation: a primer." Statistical Methods In Medical Research **8**(1): 3-15.

Schafer, J. L. and J. W. Graham (2002). "Missing data: our view of the state of the art." Psychol Methods **7**(2): 147-177.

Scharfstein, D. O., M. J. Daniels and J. M. Robins (2003). "Incorporating prior beliefs about selection bias into the analysis of randomized trials with missing outcomes." Biostatistics **4**(4): 495-512.

Schenker, N. and J. M. G. Taylor (1996). "Partially parametric techniques for multiple imputation." Computational Statistics & Data Analysis **22**(4): 425-446.

Scottish Diabetes Research Network (2010). Prevalence of type 2 diabetes and relative mortality following incident diabetes in scotland 2001-2007. European Association for the Study of Diabetes.

Scottish Diabetes Survey Monitoring Group (2013). Scottish Diabetes Survey 2013. NHS Scotland.

Scottish Intercollegiate Guidelines Network (2010). 115. Management of Obesity. Scottish Intercollegiate Guidelines Network. Edinburgh.

Scottish Intercollegiate Guidelines Network (2010). 116. Management of Diabetes - A National Clinical Guideline. Scottish Intercollegiate Guidelines Network. Edinburgh.

Seaman, S., J. Bartlett and I. White (2012). "Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods." BMC Medical Research Methodology **12**(1): 1-13.

Seaman, S. R. and I. R. White (2013). "Review of inverse probability weighting for dealing with missing data." Stat Methods Med Res **22**(3): 278-295.

Selby, J. V., A. J. Karter, L. M. Ackerson, A. Ferrara and J. Liu (2001). "Developing a Prediction Rule From Automated Clinical Databases to Identify High-Risk Patients in a Large Population With Diabetes." Diabetes Care **24**(9): 1547-1555.

Shamliyan, T., K. M. C. Talley, R. Ramakrishnan and R. L. Kane (2013). "Association of frailty with survival: a systematic literature review." Ageing research reviews **12**(2): 719-736.

Staa, T.-P. v., B. Goldacre, M. Gulliford, J. Cassell, M. Pirmohamed, A. Taweel, B. Delaney and L. Smeeth (2012). Pragmatic randomised trials using routine electronic health records: putting them to the test.

Stampfer, M. (2005). "Weight Loss and Mortality: What Does the Evidence Show?" PLoS Med **2**(6): e181.

Standl, E., M. Erbach and O. Schnell (2013). "Defending the Con Side: Obesity Paradox Does Not Exist." Diabetes Care **36**(Supplement 2): S282-S286.

StataCorp (2009). Stata Statistical Software: Release 11. College Station, TX, StataCorp LP.

Sterne, J. A. C., I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood and J. R. Carpenter (2009). "Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls." BMJ **338**.

Steyerberg, E. W. and M. van Veen (2007). "Imputation is beneficial for handling missing data in predictive models." Journal of Clinical Epidemiology **60**(9): 979.

Stumvoll, M., N. Nurjhan, G. Perriello, G. Dailey and J. E. Gerich (1995). "Metabolic effects of metformin in non-insulin-dependent diabetes mellitus." N Engl J Med **333**(9): 550-554.

Szatkowski, L., S. Lewis, A. McNeill, Y. Huang and T. Coleman (2011). "Can data from primary care medical records be used to monitor national smoking prevalence?" Journal of Epidemiology and Community Health.

Taggar, J., T. Coleman, S. Lewis and L. Szatkowski (2012). "The impact of the Quality and Outcomes Framework (QOF) on the recording of smoking targets in primary care medical records: cross-sectional analyses from The Health Improvement Network (THIN) database." BMC Public Health **12**(1): 1-11.

The BMI in Diverse Populations Collaborative Group (1999). "Effect of Smoking on the Body Mass Index-Mortality Relation: Empirical Evidence from 15 Studies." American Journal of Epidemiology **150**(12): 1297-1308.

The Diabetes Prevention Program Research Group (2012). "Long-Term Safety, Tolerability, and Weight Loss Associated With Metformin in the Diabetes Prevention Program Outcomes Study." Diabetes Care **35**(4): 731-737.

Thomas, G., K. Khunti, V. Curcin, M. Molokhia, C. Millett, A. Majeed and S. Paul (2014). "Obesity paradox in people newly diagnosed with type 2 diabetes with and without prior cardiovascular disease." Diabetes, Obesity and Metabolism **16**(4): 317-325.

Tobias, D. K., A. Pan, C. L. Jackson, E. J. O'Reilly, E. L. Ding, W. C. Willett, J. E. Manson and F. B. Hu (2014). "Body-Mass Index and Mortality among Adults with Incident Type 2 Diabetes." New England Journal of Medicine **370**(3): 233-244.

Tuthill, A., M. J. McKenna, D. O'Shea and T. J. McKenna (2008). "Weight changes in type 2 diabetes and the impact of gender." Diabetes, Obesity and Metabolism **10**(9): 726-732.

UK Prospective Diabetes Study Group (1998). "Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). UK Prospective Diabetes Study (UKPDS) Group." Lancet **352**(9131): 837-853.

UK Prospective Diabetes Study Group (UKPDS) (1998). "Effect of intensive blood-glucose control with metformin on complications in overweight patients with type 2 diabetes." Lancet **352**(9131): 854-865.

van Buuren, S. (2007). "Multiple imputation of discrete and continuous data by fully conditional specification." Statistical Methods in Medical Research **16**(3): 219-242.

van Buuren, S., H. C. Boshuizen and D. L. Knook (1999). "Multiple imputation of missing blood pressure covariates in survival analysis." Stat Med **18**(6): 681-694.

van der Heijden, G. J., A. R. Donders, T. Stijnen and K. G. Moons (2006). "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.

Vazquez, G., S. Duval, D. R. Jacobs and K. Silventoinen (2007). "Comparison of Body Mass Index, Waist Circumference, and Waist/Hip Ratio in Predicting Incident Diabetes: A Meta-Analysis." Epidemiologic Reviews **29**(1): 115-128.

Vink, G. and S. van Buuren (2013). "Multiple Imputation of Squared Terms." Sociological Methods & Research **42**(4): 598-607.

Vistisen, D., D. R. Witte, A. G. Tabak, C. Herder, E. J. Brunner, M. Kivimaki and K. Faerch (2014). "Patterns of obesity development before the diagnosis of type 2 diabetes: the Whitehall II cohort study." PLoS Med **11**(2): e1001602.

Von Hippel, P. T. (2009). "How to impute interactions, squares and other transformed variables." Sociological Methodology **39**(1): 265-291.

Waller, J. and P. Hodgkin (2000). "Defaulters in general practice: who are they and what can be done about them?" Family Practice **17**(3): 252-253.

Wayman, J. C. (2003). Multiple Imputation for Missing Data: What is it and how can I use it? Annual Meeting of the American Educational Research Association. Chicago.

Weiner, M. G., K. Barnhart, D. Xie and R. L. Tannen (2008). "Hormone therapy and coronary heart disease in young women." Menopause **15**(1): 86-93.

Welch, C. A., J. W. Bartlett and I. Petersen (In print 2014). "Application of multiple imputation using the two-fold fully conditional specification algorithm in longitudinal clinical data." Stata Journal.

Welch, C. A., I. Petersen, J. W. Bartlett, I. R. White, L. Marston, R. W. Morris, I. Nazareth, K. Walters and J. Carpenter (2014). "Evaluation of two-fold fully conditional specification multiple imputation for longitudinal electronic health record data." Statistics in Medicine.

Westphal, S. A. and P. J. Palumbo (2007). "Weight Gain and Management Concerns in Patients on Insulin Therapy." Insulin **2**(1): 31-36.

White, I. R. and J. B. Carlin (2010). "Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values." Statistics in Medicine **29**(28): 2920-2931.

White, I. R., J. Carpenter, S. Evans and S. Schroter (2007). "Eliciting and using expert opinions about dropout bias in randomized controlled trials." Clin Trials **4**(2): 125-139.

White, I. R. and P. Royston (2009). "Imputing missing covariate values for the Cox model." Statistics in Medicine **28**(15): 1982-1998.

White, I. R., P. Royston and A. M. Wood (2011). "Multiple imputation using chained equations: Issues and guidance for practice." Statistics in Medicine **30**(4): 377-399.

Wild, S., F. MacLeod, J. McKnight, G. Watt, C. MacKenzie, I. Ford, A. McConnachie and R. S. Lindsay (2008). "Impact of deprivation on cardiovascular risk factors in people with diabetes: an observational study." Diabetic Medicine **25**(2): 194-199.

Williamson, D. F., J. Madans, R. F. Anda, J. C. Kleinman, G. A. Giovino and T. Byers (1991). "Smoking Cessation and Severity of Weight Gain in a National Cohort." New England Journal of Medicine **324**(11): 739-745.

Williamson, D. F., T. J. Thompson, M. Thun, D. Flanders, E. Pamuk and T. Byers (2000). "Intentional weight loss and mortality among overweight individuals with diabetes." Diabetes Care **23**(10): 1499-1504.

Wing, R. R., P. Bolin, F. L. Brancati, G. A. Bray, J. M. Clark, M. Coday, R. S. Crow, J. M. Curtis, C. M. Egan, M. A. Espeland, M. Evans, J. P. Foreyt, S. Ghazarian, E. W. Gregg, B. Harrison, H. P. Hazuda, J. O. Hill, E. S. Horton, V. S. Hubbard, J. M. Jakicic, R. W. Jeffery, K. C. Johnson, S. E. Kahn, A. E. Kitabchi, W. C. Knowler, C. E. Lewis, B. J. Maschak-Carey, M. G. Montez, A. Murillo, D. M. Nathan, J. Patricio, A. Peters, X. Pi-Sunyer, H. Pownall, D. Reboussin, J. G. Regensteiner, A. D. Rickman, D. H. Ryan, M. Safford, T. A. Wadden, L. E. Wagenknecht, D. S. West, D. F. Williamson and S. Z. Yanovski (2013). "Cardiovascular effects of intensive lifestyle intervention in type 2 diabetes." N Engl J Med **369**(2): 145-154.

Wood, A. M., I. R. White, M. Hillsdon and J. Carpenter (2005). "Comparison of imputation and modelling methods in the analysis of a physical activity trial with missing outcomes." International Journal of Epidemiology **34**(1): 89-99.

World Health Organisation (2000). Obesity: Preventing and Managing the Global Obesity Epidemic. Report of a WHO Consultation WHO Technical Report Series.

Yki-Jarvinen, H. (2001). "Combination Therapies With Insulin in Type 2 Diabetes." Diabetes Care **24**(4): 758-767.

Young, B. A., E. Lin, M. Von Korff, G. Simon, P. Ciechanowski, E. J. Ludman, S. Everson-Stewart, L. Kinder, M. Oliver, E. J. Boyko and W. J. Katon (2008). "Diabetes Complications Severity Index and Risk of Mortality, Hospitalization, and Healthcare Utilization." The American journal of managed care **14**(1): 15-23.

Zheng, W., D. F. McLerran, B. Rolland, X. Zhang, M. Inoue, K. Matsuo, J. He, P. C. Gupta, K. Ramadas, S. Tsugane, F. Irie, A. Tamakoshi, Y.-T. Gao, R. Wang, X.-O. Shu, I. Tsuji, S. Kuriyama, H. Tanaka, H. Satoh, C.-J. Chen, J.-M. Yuan, K.-Y. Yoo, H. Ahsan, W.-H. Pan, D. Gu, M. S. Pednekar, C. Sauvaget, S. Sasazuki, T. Sairenchi, G. Yang, Y.-B. Xiang, M. Nagai, T. Suzuki, Y. Nishino, S.-L. You, W.-P. Koh, S. K. Park, Y. Chen, C.-Y. Shen, M. Thornquist, Z. Feng, D. Kang, P. Boffetta and J. D. Potter (2011). "Association between Body-Mass Index and Risk of Death in More Than 1 Million Asians." New England Journal of Medicine **364**(8): 719-729.

Zoppini, G., G. Verlato, C. Leuzinger, C. Zamboni, E. Brun, E. Bonora and M. Muggeo (2003). "Body mass index and the risk of mortality in type II diabetic patients from Verona." Int J Obes Relat Metab Disord **27**(2): 281-285.

APPENDICES

A5. APPENDICES FOR CHAPTER 5

Table A. 5.1. Survival status by missingness. Stratified by sex.

	Male		Female	
	Complete BMI	Missing BMI	Complete BMI	Missing BMI
Crude HR (95% CI)	1	0.91 (0.89, 0.93)	1	0.98 (0.94, 1.02)
Age adjusted HR (95% CI)	1	0.99 (0.97, 1.01)	1	1.01 (0.99, 1.02)
Deaths	12,349	31,615	10,034	30,974
Person-years	390,655	861,100	325,950	758,310
Crude mortality, deaths per 1,000 person-years	31.6	36.7	30.7	40.8
Age-standardised mortality, per 1,000 person years (95% CI	24.8 (24.3, 25.3)	36.1 (35.7, 36.5)	20.2 (19.8, 20.7)	31.8 (31.4, 32.1)

A6. APPENDICES FOR CHAPTER 6

Table A.6.1 Estimates of relative hazard of all-cause mortality and age-standardised mortality by categories of BMI using CCA for males

MALES	BMI category (kg/m²)				
	< 25	25 to < 30 (REF)	30 to < 35	35 to < 40	≥ 40
N	2,952 (10.20)	9,997 (34.56)	9,657 (33.38)	4,173 (14.420	2,151 (7.44)
Multivariable adjusted HR (95% CI)[a]	1.53 (1.39, 1.70)	1.00	0.99 (0.90, 1.08)	1.02 (0.90, 1.16)	1.40 (1.17, 1.67)
Standard Error	0.0791		0.0445	0.0675	0.1260
No. deaths	558	1,166	768	292	143
Person years	16,156	56,659	51,316	23,473	11,782
Crude mortality rate	34.5	20.6	16.0	12.4	12.1
Age-standardised mortality rate	21.5 (19.2, 23.9)	12.9 (12.0, 13.8)	12.5 (11.5, 13.6)	12.9 (10.9, 14.8)	14.7 (11.7, 17.7)

Table A. 6.2 Estimates of relative hazard of all-cause mortality and age-standardised mortality by categories of BMI using CCA for females

FEMALES	BMI category (kg/m²)				
	< 25	25 to < 30 (REF)	30 to < 35	35 to < 40	≥ 40
N	2,701 (11.41)	6,291 (26.58)	6,711 (28.36)	4,330 (18.30)	3,634 (15.35)
Multivariable adjusted HR (95% CI)[a]	1.48 (1.33, 1.66)	1.00	1.00 (0.89, 1.11)	1.05 (0.92, 1.20)	1.35 (1.16, 1.57)
Standard Error	0.0838		0.0547	0.0716	0.1057
No. deaths	543	752	610	318	232
Person years	15,061	35,853	38,155	24,451	20,679
Crude mortality rate	36.1	21.0	16.0	13.0	11.2
Age-standardised mortality rate	16.3 (13.2, 19.40	9.6 (8.7, 10.4)	10.1 (9.0, 11.1)	10.8 (9.3, 12.3)	14.0 (10.6, 17.5)

Table A. 6.3. Baseline characteristics of patients included in population mean imputation by BMI category (n=66,372)

		BMI category, kg/m²					
		<25	25 to < 30	30 to < 35	35 to < 40	≥ 40	Total
N (%)		5,653 (8.52)	16,288 (24.54)	30,143 (45.42)	8,503 (12.81)	5,785 (8.72)	66,372 (100.00)
Mean age at diagnosis, yrs (SD)		66.23 (12.69)	64.42 (11.75)	61.63 (12.26)	58.08 (11.40)	54.27 (10.89)	61.69 (12.37)
% Male		52.22	61.38	56.23	49.08	37.18	54.59
% Alive		80.49	88.22	88.12	92.80	93.45	97.62
Median no. of years follow-up (IQR)		5.65 (4.53, 6.74)	5.75 (4.70, 6.84)	5.79 (4.68, 6.95)	5.74 (4.69, 6.83)	5.77 (4.75, 6.84)	5.76 (4.68, 6.88)
% SIMD quintile	1 (Most deprived)	23.03	21.83	22.82	26.07	29.99	23.64
	2	21.77	22.27	22.37	23.99	24.32	22.67
	3	19.07	19.93	21.40	21.24	20.31	20.73
	4	18.29	18.56	18.98	16.21	15.23	18.13
	5 (Least deprived)	17.85	17.41	14.43	12.49	10.15	14.83
Median Charlson Comorbidity Score (IQR)		1 (1,2)	1 (1,2)	1 (1,2)	1 (1,2)	1 (1,2)	1 (1,2)
% Never smoker		40.60	40.32	42.07	43.30	47.68	42.16
Mean Systolic BP, mmHg (SD)[a]		137 (20.08)	139 (18.69)	140 (18.84)	141 (18.29)	141 (17.28)	110 (18.82)
Mean Diastolic BP, mmHg (SD) [a]		77 (10.69)	79 (10.49)	81 (10.83)	83 (10.90)	84 (10.80)	81 (10.91)
Mean glycated haemoglobin, % (SD) [a]		8.24 (2.55)	8.08 (2.20)	7.93 (2.03)	7.99 (1.98)	8.05 (1.90)	8.01 (2.10)
Mean serum creatinine, umol/L (SD) [a]		90.91 (28.65)	92.63 (28.01)	91.23 (27.39)	88.07 (21.74)	84.24 (23.36)	90.53 (26.78)
Mean Total cholesterol, mmol/L (SD)[a]		5.16 (1.35)	5.16 (1.32)	5.16 (1.31)	5.19 (1.24)	5.20 (1.19)	5.17 (1.30)
Mean Triglycerides, mmol/L [a]		2.04 (1.72)	2.50 (2.78)	2.63 (2.80)	2.68 (2.63)	2.50 (2.12)	2.54 (2.64)
Mean HDL-cholesterol, mmol/L (SD) [a]		1.42 (0.49)	1.26 (0.39)	1.23 (0.39)	1.18 (0.37)	1.17 (0.34)	1.18 (0.35)

[a] Measurement recorded closest to date of diagnosis

Table A. 6.4 Estimates of relative hazard of all-cause mortality and age-standardised mortality by categories of BMI using MICE (including Nelson Aaelen estimate of cumulative hazard rather than follow-up time)

	BMI category (kg/m²)				
	< 25	25 to < 30	30 to < 35	35 to < 40	≥ 40
Average N (%)	7,183 (10.81)	20,423 (30.86)	20,597 (31.10)	10,857 (16.21)	7,312 (11.03)
Average increase in N from CCA (%)[a]	1,528 (11.04)	4,140 (30.43)	4,229 (31.02)	2,354 (16.36)	1,527 (11.14)
Multivariable adjusted HR[bc]	1.40 (1.30, 1.50)	1.00	0.98 (0.90, 1.06)	0.99 (0.91, 1.08)	1.21 (1.08, 1.34)
Standard Error	0.049		0.039	0.046	0.07
No. deaths	1,452	2,749	2,013	857	510
Person years	39,708	116,678	116,984	61,423	40,854
Crude mortality rate [d]	36.6	23.6	17.2	14.0	12.5
Age-standardised mortality rate [c d]	19.5 (17.7, 20.9)	13.2 (12.5, 13.9)	12.2 (11.1, 13.3)	12.2 (11.2, 13.2)	14.2 (12.6, 15.9)

[a] Percentage of missing data, [b] Adjusted for age, sex and smoking status, deprivation status and Charlson comorbidity index [c] with 95% CI, [d] deaths per 1,000 person-years. CCA: CCA

Table A. 6.5. Estimates of relative hazard of all-cause mortality by categories of BMI (including BMI as a continuous variable)

BMI category (kg/m²)	Multivariable adjusted HR (95% Confidence intervals)	Standard Error
< 25	1.23 (1.04, 1.46)	0.104
25 to < 30	1.00	
30 to < 35	1.07 (0.82, 1.12)	0.074
35 to < 40	1.07 (0.89, 1.30)	0.105
≥ 40	1.34 (1.04, 1.71)	0.167

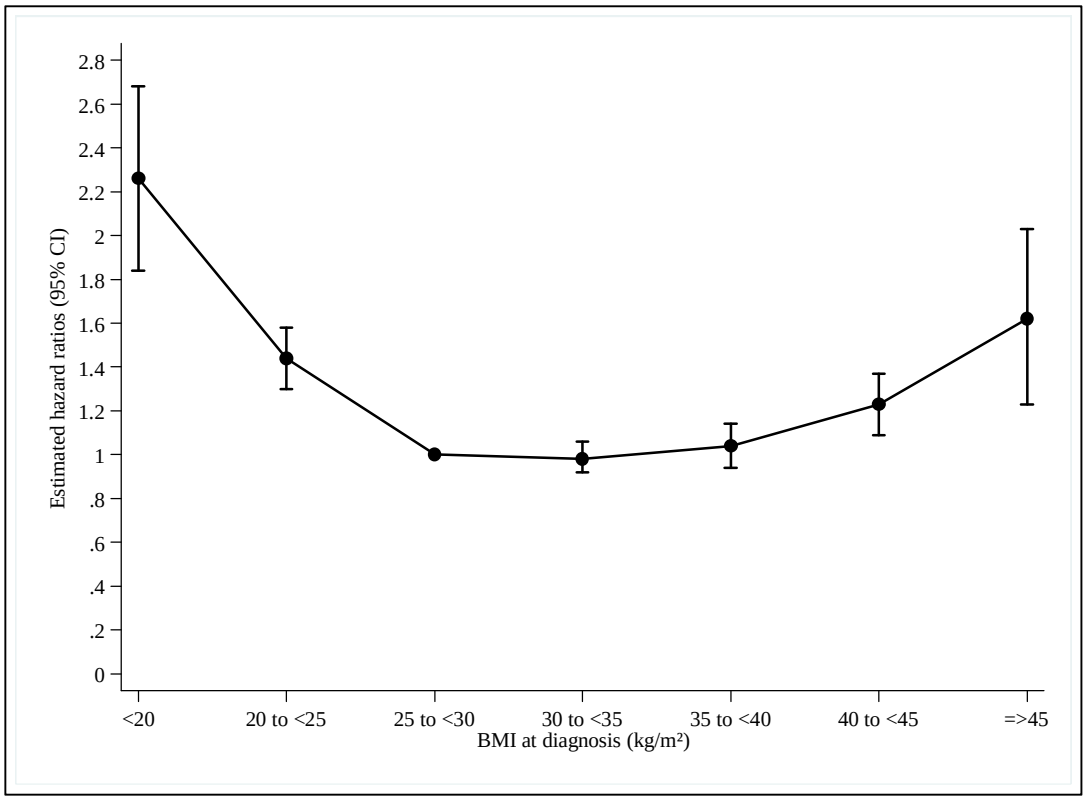

Figure A6.1. Relationship between BMI and all-cause mortality using CCA following re-categorisation of BMI at diagnosis variable. Estimates adjusted for age, sex and smoking status, deprivation status and Charlson comorbidity index

A7. APPENDICES FOR CHAPTER 7

Table A. 7.1 Characteristics of patients with unobserved BMI at diagnosis by number of BMI measurements per person

	Number of BMI measurements*					
	0	1	2	3-10	>10	TOTAL
N (%)	1,395 (10.13)	1,301 (9.44)	1,711 (12.42)	8,568 (62.20)	800 (5.81)	13,775 (100.00)
Mean age at diagnosis, yrs (SD)	64.31 (14.14)	61.93 (14.10)	61.83 (13.83)	61.87 (12.68)	59.84 (11.83)	62.00 (13.10)
% Male	51.83	54.57	53.59	53.14	48.75	52.94
% Alive	80.86	83.55	84.34	85.27	88.75	84.75
Median no. of years follow-up (IQR)	4.61 (4.05, 5.37)	4.88 (4.16, 6.08)	5.20 (4.37, 6.29)	6.29 (5.14, 7.25)	7.26 (6.37, 7.81)	5.91 (4.68, 7.07)
% With additional comorbidities	37.20	34.59	32.96	33.44	34.50	33.93
% Never smoker	47.31	45.89	47.81	44.00	43.50	44.96
Mean BMI (Median of within-patient recordings), kg/m^2 (SD)	-	31.04 (6.50)	31.02 (6.49)	31.26 (6.29)	33.43 (7.47)	31.34 (6.45)
Mean Systolic BP, mmHg (SD) [a]	140 (20.32)	139 (19.00)	139 (19.21)	140 (19.39)	141 (20.58)	140 (19.50)
Mean Diastolic BP, mmHg (SD) [a]	80 (11.52)	80 (11.10)	80 (10.86)	81 (10.91)	81 (11.44)	81 (11.01)
Mean HbA1c, % (SD) [a]	8.07 (2.13)	7.95 (2.08)	7.84 (2.01)	7.84 (1.99)	8.05 (2.02)	7.89 (2.02)
Mean serum creatinine, umol/L (SD) [a]	89.59 (24.48)	89.74 (33.42)	88.80 (25.17)	91.40 (28.50)	92.03 (28.74)	90.78 (28.74)
Mean Total cholesterol, mmol/L (SD) [a]	5.10 (1.29)	5.07 (1.31)	5.11 (1.31)	5.18 (1.31)	5.25 (1.30)	5.16 (1.31)
Median Triglycerides, mmol/L,	1.59	1.93	2.01	2.00	2.08	2.00

	Number of BMI measurements*					
(IQR) [a]	(1.37, 1.80)	(1.49, 3.00)	(1.41, 2.92)	(1.42, 2.90)	(1.44, 2.90)	(1.40, 2.90)
Mean HDL-cholesterol, mmol/L (SD) [a]	1.23 (0.37)	1.22 (0.40)	1.25 (0.39)	1.27 (0.42)	1.26 (0.46)	1.26 (0.41)

* Grouping was changed, owing to only 39 patients having more than 20 measurements

[a] Closest recording to date of diagnosis

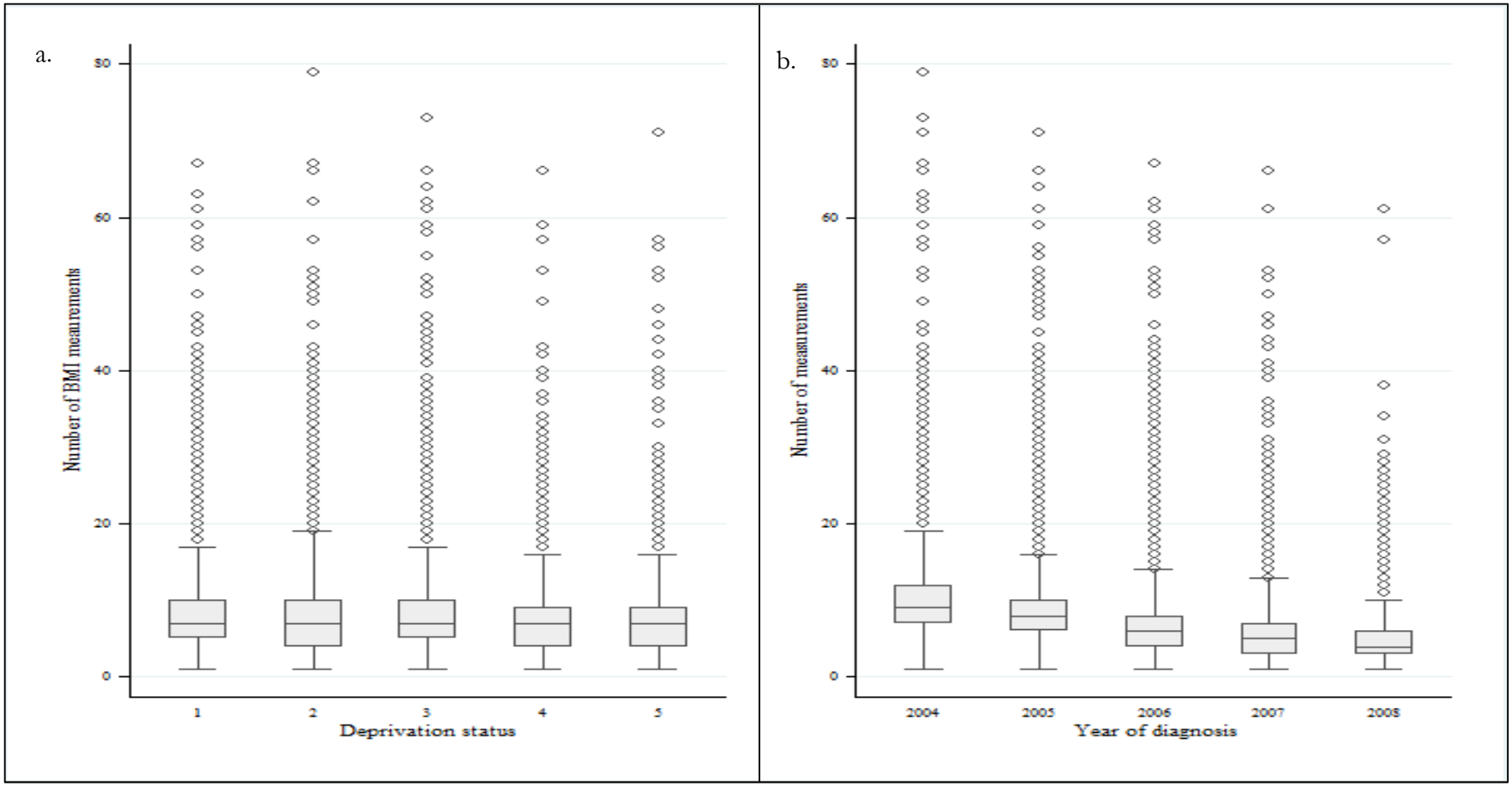

Figure A.7.1 Box and whiskers plots of number of measurements by a. Deprivation status (1=most deprived) and year of diagnosis for patients without a BMI at diagnosis.

Table A.7.2 Agreement between categories of closest BMI measurements before diagnosis and BMI at diagnosis

		BMI at date of diagnosis (kg/m²)					Diagonal Total (% of 38,820)
		< 25	25 to < 30	30 to < 35	35 to < 40	≥ 40	
Before diagnosis BMI (kg/m²)	< 25	2,639 (68.74)	1,221 (10.39)	181 (1.48)	49 (0.75)	15 (0.33)	
	25 to < 30	1,098 (28.60)	8,589 (73.10)	2,699 (22.07)	271 (4.17)	66 (1.46)	15 (0.04)
	30 to < 35	80 (2.08)	1,819 (15.48)	8,070 (66.04)	1,893 (29.11)	273 (6.06)	115 (0.30)
	35 to < 40	10 (0.26)	103 (0.88)	1,200 (9.81)	3,749 (57.66)	1,059 (23.49)	725 (1.88)
	≥ 40	12 (0.31)	17 (0.14)	72 (0.59)	540 (8.31)	3,095 (68.66)	6,872 (17.70)
Diagonal Total (% of 38,820)			12 (0.03)	27 (0.07)	255 (0.66)	4,657 (12.00)	26,142 (67.40)

Table A.7.3 Agreement between categories of closest BMI measurements after diagnosis and BMI at diagnosis

		BMI at date of diagnosis, kg/m² (column percentage)					Diagonal Total (% of 52,591)
		< 25	25 to < 30	30 to < 35	35 to < 40	≥ 40	
After diagnosis BMI (kg/m²)	< 25	4,163 (83.14)	1,426 (9.83)	46 (0.32)	11 (0.15)	3 (0.06)	
	25 to < 30	785 (15.68)	11,764 (81.10)	2,647 (18.25)	63 (0.83)	11 (0.21)	3 (0.01)
	30 to < 35	36 (0.72)	1,270 (8.76)	10,914 (75.25)	1,861 (24.62)	53 (1.03)	22 (0.05)
	35 to < 40	12 (0.24)	35 (0.24)	862 (5.94)	5,213 (69.01)	911 (17.62)	162 (0.35)
	≥ 40	11 (0.22)	10 (0.07)	35 (0.24)	407 (5.39)	4,191 (81.08)	6,845 (14.64)
Diagonal Total (% of 52,591)			11 (0.02)	22 (0.05)	106 (0.22)	3,324 (7.11)	36,245 (77.55)

Table A.7.4. Estimates of relative hazard of all-cause mortality and age-standardised mortality by categories of BMI using multiple imputation (+ before and after measurements).

	BMI category (kg/m²)				
	< 25	25 to < 30 (REF)	30 to < 35	35 to < 40	≥ 40
N	6,988 (10.76)	20,036 (30.84)	20,204 (31.10)	10,535 (16.21)	7,208 (11.09)
Increase in N from CCA (%)[a]	1,337 (10.80)	3,749 (32.28)	3,838 (31.00)	2,033 (16.42)	1,423 (10.49)
Multivariable adjusted HR (95% CI)[bc]	1.40 (1.31, 1.51)	1.00	0.98 (0.91, 1.05)	1.00 (0.91, 1.09)	1.20 (1.08, 1.33)
Standard Error	0.052		0.034	0.044	0.064
No. deaths	1,419	2,587	1,997	825	487
Person years	38,875	114,398	115,070	59,885	40,809
Crude mortality rate [d]	36.5	22.6	17.4	13.8	11.9
Age-standardised mortality rate[c d]	19.3 (17.7, 20.9)	12.7 (12.1, 13.4)	12.3 (11.6, 12.9)	12.2 (11.2, 13.2)	13.7 (12.0, 15.3)

[a] Percentage of missing data, [b] Adjusted for age, sex and smoking status, deprivation status and Charlson comorbidity index [c] with 95% CI, [d] deaths per 1,000 person-years. CCA: Complete case analysis

Table A.7.5. Extent of incompleteness in BMI at different timepoints relative to diagnosis timeframe

Variable	Observed (% of 64,971)	Unobserved (% of 64,971)
BMI t-2	19,457 (29.95)	45,514 (70.05)
BMI t-1	14,966 (23.03)	50,005 (76.97)
BMI t (at diagnosis)	52,591 (80.94)	12,380 (19.05)
BMI t+1	53,107 (81.74)	11,864 (18.26)
BMI t+2	40,028 (61.61)	24,943 (38.39)

A8. APPENDICES FOR CHAPTER 8

Table A.8.1: Predictors of missing data in all variables to be included in final analysis Model I

Variable		Complete case (79.71%)	Missing data (20.29%)	Univariate models OR (95% CI)	Multivariate model OR (95% CI)
% Age at diagnosis, years	< 50	69.23	30.77	1 (Ref)	1 (Ref)

Variable		Complete case (79.71%)	Missing data (20.29%)	Univariate models OR (95% CI)	Multivariate model OR (95% CI)
	50 – 59	71.36	28.64	0.90 (0.86, 0.95)	1.02 (0.94, 1.10)
	60 – 69	73.02	26.98	0.84 (0.80, 0.88)	0.97 (0.90, 1.05)
	70 – 79	71.76	28.24	0.92 (0.87, 0.96)	0.99 (0.91, 1.08)
	≥80	62.33	37.67	1.58 (1.48, 1.68)	1.45 (1.31, 1.61)
% Sex	Female	67.46	32.54	1 (Ref)	1 (Ref)
	Male	69.50	30.50	0.90 (0.88, 0.93)	0.89 (0.84, 0.93)
% Vital status	Alive	71.35	28.65	1 (Ref)	1 (Ref)
	Deceased	51.02	48.98	2.26 (2.18, 2.35)	1.25 (1.13, 1.39)
% SIMD	1 (Most deprived)	70.49	29.51	0.95 (0.91, 1.00)	0.99 (0.91, 1.07)
	2	69.63	30.37	1.01 (0.96, 1.06)	1.06 (0.98, 1.15)
	3	68.15	31.85	1.07 (1.02, 1.13)	1.08 (1.00, 1.17)
	4	66.61	33.39	1.15 (1.10, 1.22)	1.13 (1.04, 1.23)
	5 (Least deprived)	69.83	30.17	1 (Ref)	1 (Ref)
% Year of diagnosis	2004	67.05	32.96	1 (Ref)	1 (Ref)
	2005	71.56	28.44	0.81 (0.78, 0.85)	0.87 (0.81, 0.94)
	2006	73.31	26.69	0.74 (0.71, 0.78)	0.71 (0.65, 0.77)
	2007	72.42	27.58	0.77 (0.74, 0.81)	0.79 (0.71, 0.87)
	2008	69.58	30.42	0.88 (0.83, 0.95)	0.86 (0.75, 0.98)
Median Follow-up time, years (IQR)		5.67 (4.63, 6.79)	5.71 (4.52, 6.95)	0.95 (0.94, 0.96)	0.94 (0.91, 0.96)
Level of comorbidity (derived from CCI), %	0-1 No additional	71.27	28.73	1 (Ref)	1 (Ref)
	> 1 Additional	63.25	36.75	1.48 (1.43, 1.53)	1.20 (1.14, 1.26)

Variable		Complete case (79.71%)	Missing data (20.29%)	Univariate models OR (95% CI)	Multivariate model OR (95% CI)
% Smoking status	Current	84.96	15.04	1.06 (1.00, 1.13)	1.00 (0.95, 1.05)
	Former	86.89	13.11	0.91 (0.86, 0.96)	0.93 (0.88, 0.99)
	Never	85.72	14.28	1 (Ref)	1 (Ref)
Mean SBP, mmHg (SD) [a]		139 (18.59)	140 (19.32)	1.001 (1.000, 1,002)	-
Mean DBP, mmHg (SD) [a]		81 (10.87)	80.67 (11.02)	0.998 (0.996, 0.999)	-
Mean HbA_{1c}, % (SD) [a]		8.04 (2.12)	7.92 (2.04)	0.97 (0.96, 0.98)	-
Mean serum creatinine, umol/L (SD) [a]		90.28 (25.87)	90.81 (28.01)	1.005 (1.002, 1.007)	-
Mean total cholesterol, mmol/L (SD) [a]		5.18 (1.29)	5.16 (1.30)	0.98 (0.97, 1.00)	1.02 (1.00, 1.04)
Median triglycerides, mmol/L (SD) [a]		2.01 (1.42, 2.90)	2.00 (1.40, 2.90)	0.99 (0.98, 1.00)	-
Mean HDL-cholesterol, mmol/L (SD) [a]		1.24 (0.39)	1.25 (0.41)	1.09 (1.04, 1.14)	-
Health Board Area					
Ayrshire & Arran		65.64	34.36	1.27 (1.19, 1.34)	1.28 (1.17, 1.40)
Borders		71.39	28.61	0.98 (0.88, 1.09)	1.07 (0.91, 1.26)
Dumfries & Galloway		69.88	30.12	1.06 (0.97, 1.17)	1.18 (1.03, 1.36)
Fife		70.20	29.80	1.05 (0.98, 1.12)	1.14 (1.03, 1.26)
Forth Valley		71.76	28.24	0.95 (0.89, 1.03)	1.16 (1.04, 1.29)
Grampian		66.54	33.46	1.21 (1.15, 1.28)	0.85 (0.77, 0.94)
Greater Glasgow		70.92	29.08	1 (Ref)	1 (Ref)
Highlands		55.66	44.34	1.95 (1.82, 2.08)	1.14 (1.01, 1.29)
Lanarkshire		70.93	29.07	1.00 (0.95, 1.06)	1.03 (0.95, 1.12)
Lothian		66.70	33.30	1.22 (1.15, 1.28)	1.18 (1.09, 1.28)

Variable	Complete case (79.71%)	Missing data (20.29%)	Univariate models OR (95% CI)	Multivariate model OR (95% CI)
Orkney	49.68	50.32	2.56 (2.06, 3.19)	1.15 (0.76, 1.74)
Shetland	60.64	39.36	1.63 (1.29, 2.06)	1.29 (0.88, 1.88)
Tayside	71.89	28.11	0.96 (0.90, 1.02)	0.93 (0.84, 1.03)
Western Isles	68.18	31.82	1.21 (0.97, 1.51)	0.76 (0.51, 1.15)

[a] Recorded closest to date of diagnosis

Table A. 8.2: Biochemical predictors at date of diagnosis (defined as a measurement recorded 10 months prior to or 2 months after diagnosis) of missingness in variables included in Model I

Variable	Complete case (n=52,980)	Missing data (n=20,910)	Univariate models OR (95% CI)	Multivariate model* OR (95% CI)
Mean Systolic BP, mmHg (SD) [a] (n=62,674)	140 (18.56)	141 (19.02)	1.002 (1.002, 1.004)	1.003 (1.002, 1.004)
Mean Diastolic BP, mmHg (SD) [a] (n=62,676)	81 (10.86)	81 (10.90)	1.000 (0.998, 1.002)	1.006 (1.004, 1.007)
Mean Glycated haemoglobin, % (SD) [a] (n=52,510)	8.23 (2.21)	8.34 (2.23)	1.02 (1.01, 1.03)	1.03 (1.02, 1.04)
Mean serum creatinine, umol/L (SD) [a] (n=59,258)	90.46 (25.98)	91.70 (27.44)	1.002 (1.001, 1.003)	1.001 (1.000, 1.002)
Mean total cholesterol, mmol/L (SD) [a] (n=61,761)	5.19 (1.30)	5.26 (1.32)	1.03 (1.02, 1.05)	1.01 (1.00, 1.072
Median triglycerides, mmol/L (IQR) [a] (n=17,994)	2.11 (1.50, 3.02)	2.05 (1.47, 2.92)	1.00 (0.98, 1.01)	1.01 (0.99, 1.01)
Mean HDL cholesterol, mmol/L (SD) [a] (n=41,622)	1.24 (0.38)	1.24 (0.40)	1.07 (1.01, 1.14)	1.03 (0.97, 1.10)

* Adjusted for Age, sex, vital status, deprivation status and charlson comorbidity index

Table A.8.3: Survival status by presence of missing data in any variable to be included in final analysis model I.

	Complete case	Missing data
Crude HR (95% CI)	1	1.34 (1.28, 1.40)
Age and sex adjusted HR (95% CI)	1	1.27 (1.22, 1.33)

Deaths	5,484	2,682*
Person-years	300,593	104,370
Crude mortality, deaths per 1,000 person-years	18.2	25.7
Age-standardised mortality, per 1,000 person years (95% CI)	12.6 (12.2, 13.0)	16.1 (15.4, 16.7)

* 2,485 patients could not be included in absolute mortality calculations due to incomplete age/date of diagnosis data. Amongst these patients there were 2,116 deaths. The total person-years for these patients could not be calculated.

Table A.8.4: Predictors of missing data in all variables to be included in final analysis Model II

Variable	Complete case (n=37.49%)	Missing data (n=62.51%)	Univariate models OR (95% CI)	Multivariate model OR (95% CI)

Variable		Complete case (n=37.49%)	Missing data (n=62.51%)	Univariate models OR (95% CI)	Multivariate model OR (95% CI)
Age at diagnosis, years %	< 50	36.22	63.78	1 (Ref)	1 (Ref)
	50 – 59	38.45	61.55	0.91 (0.87, 0.95)	0.94 (0.89, 1.00)
	60 – 69	39.00	61.00	0.89 (0.85, 0.93)	0.91 (0.86, 0.97)
	70 – 79	37.24	62.76	0.96 (0.91, 1.01)	0.95 (0.89, 1.01)
	≥80	32.21	67.79	1.20 (1.12, 1.28)	1.03 (0.94, 1.13)
Sex, %	Female	32.52	64.48	1 (Ref)	1 (Ref)
	Male	36.90	63.10	1.06 (1.03, 1.09)	1.04 (1.01, 1.08)
Vital status, %	Alive	38.00	62.00	1 (Ref)	1 (Ref)
	Deceased	25.38	74.62	1.80 (1.72, 1.89)	1.46 (1.33, 1.61)
SIMD, %	1 (Most deprived)	36.83	63.17	1.09 (1.04, 1.14)	1.10 (1.03, 1.17)
	2	36.91	63.09	1.08 (1.03, 1.14)	1.14 (1.07, 1.21)
	3	35.31	64.69	1.16 (1.10, 1.22)	1.14 (1.07, 1.21)
	4	35.09	64.91	1.17 (1.11, 1.24)	1.09 (1.02, 1.16)
	5 (Least deprived)	38.80	61.20	1 (Ref)	1 (Ref)
Year of diagnosis, %	2004	30.05	69.95	1 (Ref)	1 (Ref)
	2005	38.15	61.85	0.70 (0.67, 0.73)	0.75 (0.71, 0.80)
	2006	40.62	59.38	0.63 (0.50, 0.66)	0.71 (0.65, 0.77)
	2007	40.89	59.11	0.62 (0.59, 0.65)	0.67 (0.60, 0.75)
	2008	41.25	58.75	0.61 (0.57, 0.65)	0.42 (0.36, 0.49)
Median Follow-up time, years (IQR)		5.56 (4.57, 6.64)	5.76 (4.61, 6.95)	1.09 (1.07, 1.10)	1.12 (1.09, 1.16)
Level of comorbidity (derived from CCI), %	0-1 No additional	71.27	28.73	1 (Ref)	1 (Ref)
	> 1 Additional	63.25	36.75	1.23 (1.19, 1.27)	1.08 (1.03, 1.12)

Variable		Complete case (n=37.49%)	Missing data (n=62.51%)	Univariate models OR (95% CI)	Multivariate model OR (95% CI)
Smoking status, %[a]	Current	84.96	15.04	1.00 (0.96, 1.04)	0.99 (0.94, 1.04)
	Former	86.89	13.11	0.94 (0.91, 0.98)	0.96 (0.92, 1.00)
	Never	85.72	14.28	1 (Ref)	1 (Ref)
Mean BMI, kg/m² (SD)		32.03 (6.40)	32.07 (6.55)	1.001 (0.998, 1.004)	-
Mean Systolic BP, mmHg (SD) [a]		140 (18.61)	139 (18.92)	0.999 (0.997, 1.000)	-
Mean Diastolic BP, mmHg (SD) [a]		80 (10.86)	81 (10.94)	0.999 (0.997, 1.000)	-
Mean Glycated haemoglobin, % (SD) [a]		8.20 (2.18)	7.88 (2.04)	0.93 (0.92, 0.94)	0.92 (0.91, 0.93)
Mean serum creatinine, umol/L (SD) [a]		90.12 (23.89)	90.62 (27.98)	1.000 (1.001, 0.999)	-
Mean total cholesterol, mmol/L (SD) [a]		5.18 (1.30)	5.16 (1.29)	0.99 (0.98, 1.00)	-
Median triglycerides, mmol/L (IQR) [a]		2.03 (1.45, 2.95)	2.00 (1.40, 2.90)	0.99 (0.98, 1.00)	-
Mean HDL cholesterol, mmol/L (SD) [a]		1.24 (0.39)	1.25 (0.41)	1.08 (1.04, 1.12)	1.04 (0.99, 1.09)
Health Board Area					
Ayrshire & Arran		45.98	54.02	0.55 (0.52, 0.59)	0.68 (0.63, 0.74)
Borders		10.66	89.34	3.94 (3.34, 4.66)	3.08 (2.54, 3.73)
Dumfries & Galloway		29.62	70.38	1.12 (1.01, 1.23)	1.63 (1.45, 1.83)
Fife		34.50	65.50	0.89 (0.84, 0.95)	1.45 (1.34, 1.58)
Forth Valley		23.55	76.45	1.53 (1.41, 1.65)	1.69 (1.52, 1.87)
Grampian		38.30	61.70	0.76 (0.72, 0.80)	1.12 (1.05, 1.20)
Greater Glasgow		32.01	67.99	1 (Ref)	1 (Ref)
Highlands		30.63	69.37	1.07 (0.99, 1.15)	1.06 (0.96, 1.18)
Lanarkshire		46.38	53.62	0.54 (0.52, 0.57)	0.73 (0.68, 0.79)
Lothian		33.01	66.99	0.96 (0.91, 1.01)	0.95 (0.88, 1.02)

Variable	Complete case (n=37.49%)	Missing data (n=62.51%)	Univariate models OR (95% CI)	Multivariate model OR (95% CI)
Orkney	25.80	74.20	1.35 (1.05, 1.75)	0.92 (0.65, 1.30)
Shetland	35.46	64.54	0.86 (0.67, 1.10)	0.94 (0.70, 1.26)
Tayside	48.94	51.06	0.49 (0.46, 0.52)	0.77 (0.70, 0.83)
Western Isles	37.27	62.73	0.79 (0.63, 0.99)	1.08 (0.82, 1.41)

[a] Recorded closest to date of diagnosis

Table A.8.5: Biochemical predictors at date of diagnosis of missingness in variables included in Model II

Variable	Complete case (37.49%)	Missing data (62.51%)	Univariate models OR (95% CI)	Multivariate model* OR (95% CI)
Mean Systolic BP, mmHg (SD) [a](n=62,674)	140 (18.58)	141 (19.02)	1.002 (1.002, 1.004)	1.000 (0.999, 1.001)
Mean Diastolic BP, mmHg (SD) [a] (n=62,676)	81 (10.84)	81 (10.88)	0.999 (0.998, 1.001)	1.002 (1.000, 1.003)
Mean Glycated haemoglobin, % (SD) [a] (n=49,992)	8.20 (2.18)	8.32 (2.24)	1.02 (1.01, 1.03)	1.02 (1.01, 1.03)
Mean serum creatinine, umol/L (SD) [a] (n=56,254)	90.28 (24.68)	91.09 (27.59)	1.001 (1.000, 1.002)	1.001 (1.000, 1.002)
Mean total cholesterol, mmol/L (SD) [a] (n=58,829)	5.19 (1.30)	5.23 (1.30)	1.02 (1.01, 1.04)	1.04 (1.02, 1.05)
Median triglycerides, mmol/L (IQR) [a] (n=17,153)	2.10 (1.50, 3.01)	2.10 (1.48, 3.00)	1.01 (1.00, 1.02)	1.01 (1.00, 1.02)
Mean HDL cholesterol, mmol/L (SD) [a] (n=39,785)	1.24 (0.39)	1.24 (0.38)	1.04 (0.98, 1.09)	1.00 (0.95, 1.06)

* Adjusted for Age, sex, vital status, deprivation status and charlson comorbidity index

Table A.8.6: Survival status by presence of missing data in any variable to be included in final analysis model.

	Complete case	Missing data

Crude HR (95% CI)	1	1.21 (1.16, 1.27)
Age and sex adjusted HR (95% CI)	1	1.17 (1.12, 1.23)
Deaths	2,700	5,466
Person-years	154,589	250,373
Crude mortality, deaths per 1,000 person-years	17.5	21.8
Age-standardised mortality, per 1,000 person years (95% CI)	12.0 (11.5, 12.5)	14.6 (14.2, 15.0)

* 2,485 patients could not be included in absolute mortality calculations due to incomplete age/date of diagnosis data. Amongst these patients there were 2,116 deaths. The total person-years for these patients could not be calculated.

Table A. 8.7: Predictors of missing data in BMI at diagnosis variable

Variable		Observed BMI (73.80%)	Missing BMI (26.20%)	Univariate models OR (95% CI)	Multivariate model OR (95% CI)
Age at diagnosis, years, %	< 50	75.63	24.37	1 (Ref)	1 (Ref)
	50 – 59	76.67	23.33	0.94 (0.90, 1.00)	1.02 (0.94, 1.10)
	60 – 69	78.02	21.98	0.87 (0.83, 0.92)	0.97 (0.90, 1.04)
	70 – 79	77.26	22.74	0.91 (0.86, 0.97)	1.01 (0.93, 1.09)
	≥80	67.22	32.78	1.51 (1.41, 1.63)	1.42 (1.28, 1.58)
Sex, %	Female	72.93	27.07	1 (Ref)	1 (Ref)
	Male	74.61	25.39	0.92 (0.89, 0.95)	0.89 (0.84, 0.93)
Vital status, %	Alive	76.83	23.17	1 (Ref)	1 (Ref)
	Deceased	55.10	44.90	2.70 (2.59, 2.82)	1.31 (1.17, 1.47)
SIMD, %	1 (Most deprived)	74.72	25.28	1.03 (0.97, 1.09)	0.99 (0.91, 1.07)
	2	74.08	25.92	1.07 (1.01, 1.13)	1.03 (0.94, 1.11)
	3	73.40	26.60	1.10 (1.04, 1.17)	1.07 (0.98, 1.16)
	4	71.98	28.02	1.19 (1.12, 1.26)	1.13 (1.03, 1.23)
	5 (Least deprived)	75.27	24.73	1 (Ref)	1 (Ref)
Year of diagnosis, %	2004	72.78	27.22	1 (Ref)	1 (Ref)
	2005	77.21	22.79	0.79 (0.75, 0.83)	0.86 (0.80, 0.93)
	2006	78.61	21.39	0.73 (0.69, 0.76)	0.73 (0.66, 0.81)
	2007	77.42	22.58	0.78 (0.74, 0.82)	0.83 (0.73, 0.94)
	2008	74.79	25.21	0.90 (0.84, 0.97)	0.98 (0.82, 1.16)
Median Follow-up time, years (IQR)		5.68 (4.63, 6.80)	5.68 (4.49, 6.98)	1.00 (0.99, 1.01)	0.96 (0.93, 1.00)
Comorbidity (derived from	None	76.87	23.13	1 (Ref)	1 (Ref)

Variable		Observed BMI (73.80%)	Missing BMI (26.20%)	Univariate models OR (95% CI)	Multivariate model OR (95% CI)
CCI) %	One or more	67.93	32.07	1.57 (1.52, 1.62)	1.18 (1.12, 1.24)
Smoking status, %	Current	85.61	14.39	1.05 (1.00, 1.13)	1.08 (1.01, 1.15)
	Former	87.38	12.62	0.91 (0.86, 0.96)	0.91 (0.86, 0.97)
	Never	86.32	13.68	1 (Ref)	1 (Ref)
Mean BMI, kg/m² (SD) [b]		32.03 (6.48)	31.37 (6.67)		
Mean Systolic BP, mmHg (SD) [a]		140 (15.54)	141 (19.24)	1.004 (1.003, 1.005)	-
Mean Diastolic BP, mmHg (SD) [a]		81 (10.87)	81 (10.86)	1.000 (0.998, 1.002)	-
Mean Glycated haemoglobin, % (SD) [a]		8.24 (2.21)	8.34 (2.23)	1.02 (1.01, 1.03)	1.00 (0.99, 1.01)
Mean serum creatinine, umol/L (SD) [a]		90.52 (26.08)	91.83 (27.43)	1.002 (1.001, 1.003)	-
Mean total cholesterol, mmol/L (SD) [a]		5.20 (1.30)	5.27 (1.32)	1.04 (1.02, 1.06)	1.001 (1.000, 1.003)
Median triglycerides, mmol/L (IQR) [a]		2.01 (1.42, 2.91)	2.00 (1.38, 2.88)	1.00 (0.99, 1.02)	-
Mean HDL cholesterol, mmol/L (SD) [a]		1.24 (0.39)	1.25 (0.39)	1.09 (1.02, 1.17)	1.02 (1.00, 1.04)
Health Board Area					
Ayrshire & Arran		70.99	29.01	1.28 (1.20, 1.36)	1.27 (1.16, 1.39)
Borders		75.89	24.11	1.00 (0.88, 1.13)	1.03 (0.87, 0.97)
Dumfries & Galloway		75.33	24.67	1.03 (0.93, 1.14)	1.20 (1.04, 1.38)
Fife		73.70	26.30	1.12 (1.04, 1.20)	1.11 (1.01, 1.23)
Forth Valley		75.02	24.98	1.04 (0.96, 1.13)	1.14 (1.02, 1.27)
Grampian		73.99	26.01	1.10 (1.04, 1.17)	0.85 (0.77, 0.94)
Greater Glasgow		75.81	24.19	1 (Ref)	1 (Ref)
Highlands		66.02	33.98	1.61 (1.50, 1.74)	1.16 (1.02, 1.27)
Lanarkshire		73.71	26.29	1.12 (1.05, 1.18)	1.01 (0.92, 1.10)

Variable	Observed BMI (73.80%)	Missing BMI (26.20%)	Univariate models OR (95% CI)	Multivariate model OR (95% CI)
Lothian	73.46	26.54	1.13 (1.07, 1.20)	1.15 (1.06, 1.25)
Orkney	63.06	36.94	1.84 (1.46, 2.31)	1.20 (0.80, 1.80)
Shetland	73.40	26.60	1.14 (0.87, 1.48)	1.17 (0.79, 1.73)
Tayside	75.74	24.26	1.00 (0.94, 1.08)	0.90 (0.81, 1.00)
Western Isles	74.24	25.76	1.09 (0.85, 1.39)	0.72 (0.47, 1.10)

[a] Recorded at date of diagnosis
[b] Closest measurement recorded outside date of diagnosis time-frame

Table A. 8.8: Survival status by presence of missing data in BMI at diagnosis variable

	Complete smoking status	Missing smoking status
Crude HR (95% CI)	1	1.44 (1.37, 1.51)
Age and sex adjusted HR (95% CI)	1	1.37 (1.31, 1.44)
Deaths	5,645	2,521
Person-years	309,172	95,792
Crude mortality, deaths per 1,000 person-years	18.3	26.3
Age-standardised mortality, per 1,000 person years (95% CI)	12.5 (12.2, 12.9)	16.9 (16.1, 17.6)

* 2,485 patients could not be included in absolute mortality calculations due to incomplete age/date of diagnosis data. Amongst these patients there were 2,116 deaths. The total person-years for these patients could not be calculated.

Table A.8.9: Predictors of missing data in smoking status at diagnosis variable

Variable		Observed smoking status (88.26%)	Missing smoking status (11.74%)	Univariate models OR (95% CI)	Multivariate model OR (95% CI)
Age at diagnosis, years, %	< 50	90.09	9.91	1 (Ref)	1 (Ref)
	50 – 59	91.28	8.72	0.85 (0.72, 0.81)	0.82 (0.77, 0.87)
	60 – 69	91.37	8.63	0.77 (0.72, 0.81)	0.70 (0.66, 0.74)
	70 – 79	91.85	8.15	0.81 (0.76, 0.85)	0.70 (0.66, 0.75)
	≥80	89.71	10.29	1.26 (1.17, 1.35)	0.98 (0.90, 1.07)
Sex, %	Female	88.56	11.44	1 (Ref)	1 (Ref)
	Male	88.09	11.91	0.95 (0.91, 0.98)	0.97 (0.93, 1.01)
Vital status, %	Alive	90.83	9.17	1 (Ref)	1 (Ref)
	Deceased	72.39	27.61	2.24 (2.16, 2.37)	1.18 (1.11, 1.25)
SIMD, %	1 (Most deprived)	88.51	11.49	0.92 (0.87, 0.98)	0.90 (0.84, 0.96)
	2	88.61	11.39	0.95 (0.90, 1.01)	0.92 (0.86, 0.98)
	3	88.27	11.73	1.04 (0.98, 1.11)	0.98 (0.91, 1.05)
	4	86.95	13.05	1.15 (1.08, 1.22)	1.11 (1.04, 1.20)
	5 (Least deprived)	89.59	10.41	1 (Ref)	1 (Ref)
Year of diagnosis, %	2004	90.33	9.67	1 (Ref)	1 (Ref)
	2005	92.16	7.84	0.77 (0.74, 0.82)	0.75 (0.71, 0.80)
	2006	92.06	7.94	0.75 (0.72, 0.80)	0.67 (0.64, 0.71)
	2007	90.94	9.06	0.71 (0.68, 0.75)	0.61 (0.58, 0.65)
	2008	87.43	12.57	0.77 (0.71, 0.84)	0.55 (0.50, 0.61)
Median Follow-up time, years (IQR)		5.68 (4.61, 6.83)	5.58 (4.41, 6.92)	1.00 (0.99. 1.01)	1.01 (0.98, 1.03)
Level of comorbidity	0-1 No additional	91.03	8.97	1 (Ref)	1 (Ref)

Variable	Observed smoking status (88.26%)	Missing smoking status (11.74%)	Univariate models OR (95% CI)	Multivariate model OR (95% CI)
(derived from CCI) % > 1 Additional	82.94	17.06	1.51 (1.45, 1.56)	1.12 (1.06, 1.17)
Mean BMI, kg/m² (SD)[b]	31.89 (6.49)	32.27 (6.99) *n = 3,798	1.01 (1.00, 1.02)	-
Mean SBP, mmHg (SD) [a]	139 (18.79)	139 (19.42)	0.999 (0.998, 1.000)	-
Mean DBP, mmHg (SD) [a]	80 (10.94)	80 (11.24)	0.999 (0.997, 1.000)	-
Mean HbA_{1c}, % (SD) [a]	8.03 (2.12)	7.88 (2.03)	0.965 (0.956, 0.975)	0.97 (0.96, 0.98)
Mean serum creatinine, umol/L (SD) [a]	90.91 (27.21)	92.09 (30.53)	1.001 (1.000, 1.002)	-
Mean total cholesterol, mmol/L (SD) [a]	5.16 (1.30)	5.11 (1.30)	0.97 (0.96, 0.99)	0.97 (0.96, 0.99)
Median triglycerides, mmol/L (IQR) [a]	2.00 (1.41, 2.90)	2.05 (1.40, 3.03)	1.00 (0.99, 1.01)	-
Mean HDL cholesterol, mmol/L (SD) [a]	1.24 (0.40)	1.25 (0.41)	1.06 (1.00, 1.12)	-
Health Board Area				
Ayrshire & Arran	85.67	14.33	1.23 (1.15, 1.31)	1.09 (1.00, 1.18)
Borders	91.39	8.61	0.88 (0.77, 1.01)	0.88 (0.75, 1.04)
Dumfries & Galloway	92.80	7.20	0.94 (0.84, 1.05)	1.17 (1.03, 1.33)
Fife	90.93	9.07	0.99 (0..91, 1.07)	1.06 (0.96, 1.16)
Forth Valley	92.01	7.99	0.81 (0.74, 0.89)	0.80 (0.72, 0.89)
Grampian	85.92	14.08	1.44 (1.35, 1.53)	1.58 (1.46, 1.70)
Greater Glasgow	90.48	9.52	1 (Ref)	1.00 (Ref)
Highlands	78.35	21.65	2.32 (2.16, 2.50)	2.63 (2.41. 2.87)
Lanarkshire	88.14	11.86	1.00 (0.94, 1.07)	0.93 (0.86, 1.00)
Lothian	87.09	12.91	1.21 (1.14, 1.29)	1.35 (1.26, 1.45)
Orkney	73.25	26.75	3.15 (2.53, 3.93)	3.89 (3.06, 4.94)

Variable	Observed smoking status (88.26%)	Missing smoking status (11.74%)	Univariate models OR (95% CI)	Multivariate model OR (95% CI)
Shetland	82.62	17.38	1.84 (1.43, 2.37)	2.20 (1.68, 2.89)
Tayside	90.31	9.69	0.98 (0.91, 1.06)	1.21 (1.11, 1.32)
Western Isles	88.18	11.82	1.34 (1.04, 1.72)	1.44 (1.07, 1.93)

[a] Recorded at date of diagnosis
b

Table A.8.10: Survival status by presence of missing data in smoking status at diagnosis variable

	Complete smoking status	Missing smoking status
Crude HR (95% CI)	1	1.25 (1.20, 1.30)
Age and sex adjusted HR (95% CI)	1	1.22 (1.16, 1.27)
Deaths	7,335	831
Person-years	369,282	35,681
Crude mortality, deaths per 1,000 person-years	23.3	23.3
Age-standardised mortality, per 1,000 person years (95% CI)	13.5 (13.1, 13.8)	15.6 (14.5, 16.8)

Table A.8.11: Predictors of missing data in systolic blood pressure at diagnosis variable

Variable		Observed systolic BP (84.92%)	Missing systolic BP (15.18%)	Univariate models OR (95% CI)	Multivariate model OR (95% CI)
Age at diagnosis, years, %	< 50	83.24	16.76	1 (Ref)	1 (Ref)
	50 – 59	87.25	12.75	0.73 (0.68, 0.77)	0.70 (0.62, 0.79)
	60 – 69	89.14	10.86	0.60 (0.57, 0.64)	0.61 (0.54, 0.68)
	70 – 79	90.11	9.89	0.55 (0.51, 0.58)	0.55 (0.48, 0.63)
	≥80	87.11	12.19	0.69 (0.63, 0.76)	0.51 (0.41, 0.63)
Sex, %	Female	84.92	15.08	1 (Ref)	1 (Ref)
	Male	84.82	15.18	1.01 (0.97, 1.05)	0.95 (0.87, 1.04)
Vital status, %	Alive	87.42	12.58	1 (Ref)	1 (Ref)
	Deceased	68.76	31.24	3.16 (3.01, 3.31)	1.18 (0.95, 1.46)
SIMD, %	1 (Most deprived)	84.43	15.57	1.20 (1.12, 1.29)	1.17 (1.00, 1.36)
	2	84.95	15.05	1.16 (1.08, 1.24)	1.11 (0.96, 1.30)
	3	84.74	15.26	1.17 (1.09, 1.26)	1.21 (1.04, 1.42)
	4	84.29	15.71	1.22 (1.13, 1.31)	1.26 (1.08, 1.47)
	5 (Least deprived)	86.71	13.29	1 (Ref)	1 (Ref)
Year of diagnosis, %	2004	85.90	14.20	1 (Ref)	1 (Ref)
	2005	89.41	10.59	0.72 (0.67, 0.76)	0.68 (0.59, 0.79)
	2006	88.73	11.27	0.77 (0.72, 0.82)	0.81 (0.67, 0.97)
	2007	87.85	12.15	0.84 (0.78, 0.89)	1.19 (0.93, 1.52)
	2008	84.70	15.30	1.09 (1.00, 1.19)	1.34 (0.98, 1.84)
Median Follow-up time, years (IQR)		5.68 (4.61, 6.81)	5.64 (4.47, 7.00)	1.01 (0.99, 1.02)	0.93 (0.86, 1.00)
Level of comorbidity (derived from CCI) %	0-1 No additional	87.29	12.71	1 (Ref)	1 (Ref)
	> 1 Additional	80.10	19.90	1.71 (1.64, 1.78)	1.02 (0.93, 1.12)

Variable		Observed systolic BP (84.92%)	Missing systolic BP (15.18%)	Univariate models OR (95% CI)	Multivariate model OR (95% CI)
Smoking status, %	Current	94.97	5.03	1.27 (1.15, 1.41)	1.20 (1.07, 1.33)
	Former	96.33	3.67	0.92 (0.83, 1.01)	1.00 (0.90, 1.11)
	Never	96.01	3.99	1 (Ref)	1 (Ref)
Mean BMI, kg/m² (SD) [a]		32.04 (6.45)	32.34 (7.40)	1.01 (0.99, 1.02)	-
Mean Systolic BP , mmHg (SD)[b]		137 (17.85)	136 (19.46)	1.004 (1.003, 1.005)	
Mean Diastolic BP, mmHg (SD) [a]		81 (10.90)	80 (11.02)	0.994 (0.992, 0.997)	-
Mean glycated haemoglobin, % (SD) [a]		8.02 (2.11)	7.94 (2.06)	0.98 (0.97, 0.99)	-
Mean serum creatinine, umol/L (SD) [a]		90.68 (26.47)	88.21 (26.38)	0.995 (0.994, 0.996)	-
Mean total cholesterol, mmol/L (SD) [a]		5.18 (1.29)	5.12 (1.30)	0.97 (0.95, 0.99)	1.02 (0.99, 1.06)
Median triglycerides, mmol/L (IQR) [a]		2.00 (1.41, 2.90)	2.01 (1.36, 2.95)	1.00 (0.99, 1.02)	-
Mean HDL cholesterol, mmol/L (SD) [a]		1.24 (0.40)	1.24 (0.38)	0.98 (0.92, 1.05)	-
Health Board Area					
Ayrshire & Arran		81.58	18.42	1.44 (1.33, 1.55)	1.78 (1.53, 1.06)
Borders		86.62	13.38	0.98 (0.84, 1.15)	0.96 (0.69, 1.35)
Dumfries & Galloway		87.16	12.84	0.94 (0.82, 1.07)	1.38 (1.08, 1.77)
Fife		84.89	15.11	1.13 (1.04, 1.24)	1.16 (0.97, 1.40)
Forth Valley		84.17	15.83	1.20 (1.09, 1.32)	1.83 (1.54, 2.18)
Grampian		87.49	12.51	0.91 (0.84, 0.99)	0.79 (0.65, 0.95)
Greater Glasgow		86.41	13.59	1 (Ref)	1 (Ref)
Highlands		81.19	18.81	1.47 (1.35, 1.61)	1.39 (1.13, 1.72)
Lanarkshire		82.74	17.26	1.33 (1.24, 1.42)	1.23 (1.06, 1.43)
Lothian		84.70	15.30	1.15 (1.07, 1.23)	1.33 (1.14, 1.54)

Variable	Observed systolic BP (84.92%)	Missing systolic BP (15.18%)	Univariate models OR (95% CI)	Multivariate model OR (95% CI)
Orkney	75.48	24.52	2.07 (1.59, 2.68)	1.31 (0.64, 2.70)
Shetland	85.46	14.54	1.08 (0.77, 1.51)	3.18 (1.95, 5.19)
Tayside	86.10	13.90	1.03 (0.94, 1.12)	0.88 (0.72, 1.08)
Western Isles	81.82	18.18	1.41 (1.06, 1.88)	1.34 (0.70, 2.55)

[a] Recorded at date of diagnosis
[b] Closest measurement recorded outside date of diagnosis time-frame

Table A.8.12: Survival status by presence of missing data in systolic blood pressure at diagnosis variable

	Complete systolic blood pressure	Missing systolic blood pressure
Crude HR (95% CI)	1	1.13 (1.06, 1.20)
Age and sex adjusted HR (95% CI)	1	1.29 (1.22, 1.38)
Deaths	7,034	1,643
Person-years	350,954	48,567
Crude mortality, deaths per 1,000 person-years	20.0	23.0
Age-standardised mortality, per 1,000 person years (95% CI)	13.2 (12.9, 13.6)	17.0 (15.9, 18.0)

Table A. 8.13: Predictors of missing data in glycated haemoglobin at diagnosis variable

Variable		Observed glycated haemoglobin (67.66%)	Missing glycated haemoglobin (32.34%)	Univariate models OR (95% CI)	Multivariate model OR (95% CI)
Age at diagnosis, years, %	< 50	70.65	29.35	1 (Ref)	1 (Ref)
	50 – 59	70.32	29.68	1.02 (0.97, 1.07)	1.07 (1.00, 1.15)
	60 – 69	70.64	29.36	1.00 (0.95, 1.05)	1.11 (1.03, 1.19)
	70 – 79	69.80	30.20	1.04 (0.99, 1.10)	1.18 (1.10, 1.27)
	≥80	65.07	34.93	1.29 (1.21, 1.38)	1.31 (1.18, 1.45)
Sex, %	Female	67.41	32.59	1 (Ref)	1 (Ref)
	Male	67.94	32.06	0.98 (0.95, 1.01)	1.00 (0.95, 1.04)
Vital status, %	Alive	70.03	29.97	1 (Ref)	1 (Ref)
	Deceased	52.98	47.02	2.07 (1.99, 2.16)	1.16 (1.05, 1.29)
SIMD, %	1 (Most deprived)	68.96	31.04	0.96 (0.91, 1.01)	1.01 (0.94, 1.09)
	2	68.57	31.43	0.98 (0.93, 1.03)	1.00 (0.93, 1.08)
	3	66.64	33.36	1.07 (1.01, 1.13)	1.05 (0.97, 1.13)
	4	65.96	34.04	1.10 (1.04, 1.16)	1.03 (0.96, 1.11)
	5 (Least deprived)	68.11	31.89	1 (Ref)	1 (Ref)
Year of diagnosis, %	2004	65.44	34.56	1 (Ref)	1 (Ref)
	2005	70.57	29.43	0.79 (0.75, 0.83)	0.89 (0.83, 0.96)
	2006	71.84	28.16	0.74 (0.71, 0.78)	0.94 (0.85, 1.03)
	2007	82.11	27.89	0.73 (0.70, 0.77)	0.97 (0.86, 1.10)
	2008	71.37	28.63	0.76 (0.71, 0.82)	0.94 (0.80,1.12)
Median Follow-up time, years (IQR)		5.64 (4.59, 6.75)	5.79 (4.61, 7.00)	1.06 (1.05, 1.07)	1.07 (1.03, 1.11)
Level of comorbidity	0-1 No additional	70.26	29.74	1 (Ref)	1 (Ref)

Variable		Observed glycated haemoglobin (67.66%)	Missing glycated haemoglobin (32.34%)	Univariate models OR (95% CI)	Multivariate model OR (95% CI)
(derived from CCI) %	> 1 Additional	62.68	37.32	1.41 (1.36, 1.45)	1.00 (0.95, 1.05)
Smoking status, %	Current	76.44	23.56	1.01 (0.96, 1.05)	1.04 (0.98, 1.10)
	Former	75.68	24.32	1.05 (1.00, 1.10)	1.03 (0.98, 1.09)
	Smoker	76.54	23.46	1 (Ref)	1 (Ref)
Mean BMI, kg/m² (SD)		32.00 (6.47)	32.22 (6.49)	1.005 (1.002, 1.008)	-
Mean SBP, mmHg (SD) [a]		139 (18.66)	140 (19.13)	1.002 (1.001, 1.003)	-
Mean DBP, mmHg (SD) [a]		81 (10.85)	81 (11.06)	1.001 (0.999, 1.002)	-
Mean HbA1c, % (SD) [b]		7.35 (1.53)	7.33 (1.59)	0.99 (0.98, 1.00)	
Mean serum creatinine, umol/L (SD) [a]		90.01 (25.80)	91.49 (28.08)	1.002 (1.001, 1.003)	-
Mean total cholesterol, mmol/L (SD) [a]		5.19 (1.30)	5.12 (1.28)	0.96 (0.95, 0.97)	0.98 (0.96, 1.00)
Median triglycerides, mmol/L (IQR) [a]		2.10 (1.50, 3.02)	2.09 (1.48, 2.97)	0.99 (0.98, 1.00)	-
Mean HDL cholesterol, mmol/L (SD) [a]		1.24 (0.40)	1.25 (0.40)	1.05 (1.00, 1.10)	1.04 (0.98, 1.09)
Health Board Area					
Ayrshire & Arran		71.08	28.92	0.97 (0.91, 1.04)	0.78 (0.71, 0.85)
Borders		52.32	47.68	2.18 (1.96, 2.43)	3.48 (2.94, 4.10)
Dumfries & Galloway		65.43	34.57	1.26 (1.15, 1.39)	1.43 (1.26, 1.62)
Fife		56.39	43.61	1.85 (1.74, 1.97)	2.18 (2.01, 2.37)
Forth Valley		64.93	35.07	1.29 (1.20, 1.39)	1.48 (1.33, 1.64)
Grampian		60.13	39.87	1.59 (1.50, 1.68)	1.90 (1.76, 2.05)
Greater Glasgow		70.53	29.47	1 (Ref)	1 (Ref)
Highlands		66.27	33.73	1.22 (1.13, 1.31)	1.07 (0.96, 1.20)
Lanarkshire		70.96	29.04	0.98 (0.93, 1.04)	0.78 (0.72, 0.85)

Variable	Observed glycated haemoglobin (67.66%)	Missing glycated haemoglobin (32.34%)	Univariate models OR (95% CI)	Multivariate model OR (95% CI)
Lothian	71.98	28.02	0.93 (0.88, 0.98)	0.84 (0.77, 0.92)
Orkney	61.78	38.22	1.48 (1.18, 1.86)	1.20 (0.81, 1.77)
Shetland	75.18	24.82	0.79 (0.60, 1.04)	0.93 (0.65, 1.33)
Tayside	70.47	29.53	1.00 (0.94, 1.07)	1.18 (1.08, 1.28)
Western Isles	67.88	32.12	1.13 (0.90, 1.43)	0.99 (0.72, 1.36)

[a] Recorded at date of diagnosis
[b] Closest measurement recorded outside date of diagnosis time-frame

Table A. 8.14: Survival status by presence of missing data in glycated haemoglobin at diagnosis variable

	Complete glycated haemoglobin	Missing glycated haemoglobin
Crude HR (95% CI)	1	1.13 (1.08, 1.18)
Age and sex adjusted HR (95% CI)	1	1.09 (1.04, 1.14)
Deaths	5,426	2,727
Person-years	278,144	121,379
Crude mortality, deaths per 1,000 person-years	19.5	22.5
Age-standardised mortality, per 1,000 person years (95% CI)	13.2 (12.8, 13.6)	14.8 (14.1, 15.4)

Table A. 8.15: Predictors of missing data in total cholesterol at diagnosis variable

Variable		Observed total cholesterol (82.35%)	Missing total cholesterol (32.34%)	Univariate models OR (95% CI)	Multivariate model OR (95% CI)
Age at diagnosis, years, %	< 50	77.51	22.49	1 (Ref)	1 (Ref)
	50 – 59	82.13	17.87	0.75 (0.71, 0.79)	0.70 (0.64, 0.76)
	60 – 69	84.63	15.37	0.63 (0.59, 0.66)	0.57 (0.52, 0.63)
	70 – 79	84.87	15.13	0.61 (0.58, 0.65)	0.54 (0.49, 0.60)
	≥80	78.57	21.43	0.94 (0.87, 1.02)	0.73 (0.63, 0.84)
Sex, %	Female	78.69	21.31	1 (Ref)	1 (Ref)
	Male	80.47	19.53	0.90 (0.86, 0.93)	0.86 (0.81, 0.92)
Vital status, %	Alive	82.43	17.57	1 (Ref)	1 (Ref)
	Deceased	62.23	37.77	2.85 (2.72, 2.98)	1.54 (1.33, 1.89)
SIMD, %	1 (Most deprived)	79.24	20.76	1.11 (1.04, 1.17)	1.10 (0.99, 1.23)
	2	79.94	20.06	1.06 (1.00, 1.13)	1.01 (0.91, 1.13)
	3	79.83	20.17	1.07 (1.00, 1.13)	0.97 (0.87, 1.08)
	4	79.04	20.96	1.12 (1.05, 1.19)	1.02 (0.91, 1.14)
	5 (Least deprived)	80.84	19.16	1 (Ref)	1 (Ref)
Year of diagnosis, %	2004	75.93	24.07	1 (Ref)	1 (Ref)
	2005	83.18	16.82	0.64 (0.60, 0.67)	0.69 (0.63, 0.76)
	2006	84.98	15.02	0.56 (0.53, 0.59)	0.63 (0.55, 0.720
	2007	85.69	14.31	0.53 (0.50, 0.56)	0.58 (0.49, 0.70)
	2008	83.15	16.85	0.64 (0.59, 0.70)	0.32 (0.14, 0.42)
Median Follow-up time, years (IQR)		5.64 (4.59, 6.75)	5.93 (4.65, 7.15)	1.12 (1.10, 1.13)	1.07 (1.02, 1.13)
Level of comorbidity	0-1 No additional	81.85	18.15	1 (Ref)	1 (Ref)

Variable		Observed total cholesterol (82.35%)	Missing total cholesterol (32.34%)	Univariate models OR (95% CI)	Multivariate model OR (95% CI)
(derived from CCI) %	> 1 Additional	75.35	24.65	1.48 (1.42, 1.53)	0.98 (0.89, 1.06)
Smoking status, %	Current	88.48	11.52	1.04 (0.97, 1.11)	1.03 (0.95, 1.12)
	Former	90.55	9.45	0.83 (0.78, 0.89)	0.92 (0.86, 0.99)
	Smoker	88.87	11.13	1 (Ref)	1 (Ref)
Mean BMI, kg/m² (SD)		32.05 (6.42)	32.07 (7.02)	1.000 (0.996, 1.004)	-
Mean SBP, mmHg (SD) [a]		140 (18.66)	139 (19.51)	0.998 (0.997, 0.999)	-
Mean DBP, mmHg (SD) [a]		80.75 (10.89)	80.66 (11.05)	0.999 (0.997, 1.001	-
Mean HbA1c, % (SD) [a]		8.05 (2.12)	7.77 (1.98)	0.93 (0.92, 0.94)	0.95 (0.94, 0.97)
Mean serum creatinine, umol/L (SD) [a]		90.53 (25.62)	89.88 (30.70)	0.999 (0.998, 1.000)	-
Mean Total cholesterol, mmol/L (SD) [b]		4.69 (1.13)	5.03 (1.25)	1.27 (1.25, 1.29)	
Median triglycerides, mmol/L (IQR) [a]		2.10 (1.50, 3.00)	2.60 (1.60, 3.98)	0.99 (0.98, 1.00)	-
Mean HDL cholesterol, mmol/L (SD) [a]		1.24 (0.40)	1.26 (0.42)	1.12 (1.06, 1.18)	1.12 (1.04, 1.21)
Health Board Area					
Ayrshire & Arran		81.07	18.93	0.98 (0.92, 1.06)	0.83 (0.73, 0.95)
Borders		78.87	21.13	1.13 (0.99, 1.28)	1.41 (1.06, 1.89)
Dumfries & Galloway		78.86	21.14	1.13 (1.01, 1.26)	1.75 (1.47, 2.09)
Fife		76.79	23.21	1.27 (1.18, 1.37)	1.76 (1.56, 1.99)
Forth Valley		81.67	18.33	0.95 (0.87, 1.03)	1.10 (0.93, 1.30)
Grampian		78.52	21.48	1.15 (1.08, 1.23)	1.74 (1.55, 1.95)
Greater Glasgow		80.83	19.17	1 (Ref)	1 (Ref)
Highlands		74.45	25.55	1.45 (1.33, 1.57)	1.63 (1.39, 1.90)
Lanarkshire		79.48	20.52	1.09 (1.02, 1.16)	0.92 (0.82, 1.04)

Variable	Observed total cholesterol (82.35%)	Missing total cholesterol (32.34%)	Univariate models OR (95% CI)	Multivariate model OR (95% CI)
Lothian	78.90	21.10	1.13 (1.06, 1.20)	1.34 (1.19, 1.50)
Orkney	70.38	29.62	1.77 (1.39, 2.27)	1.37 (0.77, 2.46)
Shetland	85.11	14.89	0.74 (0.53, 1.03)	1.16 (0.70, 1.93)
Tayside	82.92	17.08	0.87 (0.80, 0.94)	1.35 (1.19, 1.52)
Western Isles	76.06	23.94	1.33 (1.03, 1.71)	1.64 (1.08, 2.50)

[a] Recorded at date of diagnosis
[b] Closest measurement recorded outside date of diagnosis time-frame

Table A. 8.16: Survival status by presence of missing data in total cholesterol at diagnosis variable

	Complete total cholesterol	Missing total cholesterol
Crude HR (95% CI)	1	1.22 (1.16, 1.28)
Age and sex adjusted HR (95% CI)	1	1.30 (1.23, 1.37)
Deaths	6,376	1,777
Person-years	327,722	71,803
Crude mortality, deaths per 1,000 person-years	19.5	24.7
Age-standardised mortality, per 1,000 person years (95% CI)	12.9 (12.5, 13.2)	17.2 (16.3, 18.0)

Table A. 8.17: Predictors of missing HDL-cholesterol at diagnosis variable

Variable		Observed HDL-cholesterol (53.84%)	Missing HDL-cholesterol (46.16%)	Univariate models OR (95% CI)	Multivariate model OR (95% CI)
Age at diagnosis, years, %	< 50	52.63	47.37	1 (Ref)	1 (Ref)
	50 – 59	56.15	43.85	0.87 (0.83, 0.91)	0.90 (0.85, 0.95)
	60 – 69	57.26	42.74	0.83 (0.79, 0.87)	0.87 (0.83, 0.92)
	70 – 79	56.67	43.33	0.85 (0.81, 0.89)	0.89 (0.84, 0.94)
	≥80	52.52	47.48	1.00 (0.94, 1.07)	0.92 (0.84, 1.00)
Sex, %	Female	52.90	47.10	1 (Ref)	1 (Ref)
	Male	54.67	45.33	0.93 (0.90, 0.96)	0.96 (0.92, 0.99)
Vital status, %	Alive	56.13	43.87	1 (Ref)	1 (Ref)
	Deceased	39.68	60.32	1.95 (1.86, 2.03)	1.41 (1.29, 1.55)
SIMD, %	1 (Most deprived)	51.96	48.04	1.25 (1.19, 1.32)	1.22 (1.14, 1.29)
	2	53.05	46.95	1.20 (1.14, 1.26)	1.27 (1.20, 1.35)
	3	54.20	45.80	1.15 (1.09, 1.20)	1.19 (1.12, 1.27)
	4	54.29	45.71	1.14 (1.08, 1.20)	1.15 (1.08, 1.23)
	5 (Least deprived)	57.54	42.46	1 (Ref)	1 (Ref)
Year of diagnosis, %	2004	43.16	56.84	1 (Ref)	1 (Ref)
	2005	55.22	44.78	0.62 (0.59, 0.64)	0.68 (0.64, 0.72)
	2006	60.45	39.55	0.50 (0.48, 0.52)	0.60 (0.56, 0.65)
	2007	62.84	37.16	0.45 (0.43, 0.47)	0.58 (0.52, 0.65)
	2008	64.08	35.92	0.43 (0.40, 0.46)	0.51 (0.44, 0.49)
Median Follow-up time, years (IQR)		5.50 (4.52, 6.60)	5.93 (4.72, 7.08)	1.18 (1.17, 1.19)	1.11 (1.08, 1.15)
Level of comorbidity (derived from CCI) %	0-1 No additional	55.67	44.33	1 (Ref)	1 (Ref)
	> 1 Additional	50.5	49.65	1.24 (1.20, 1.28)	1.02 (0.99, 1.07)

Variable		Observed HDL-cholesterol (53.84%)	Missing HDL-cholesterol (46.16%)	Univariate models OR (95% CI)	Multivariate model OR (95% CI)
Smoking status, %	Current	60.01	39.99	1.00 (0.96, 1.05)	0.95 (0.90, 0.99)
	Former	62.28	37.72	0.91 (0.88, 0.95)	0.94 (0.90, 0.98)
	Smoker	60.07	39.93	1 (Ref)	1 (Ref)
Mean BMI, kg/m² (SD)		32.07 (6.41)	32.01 (6.59)	0.998 (0.996, 1.001)	-
Mean SBP, mmHg (SD) [a]		140 (18.64)	139 (19.01)	0.998 (0.997, 0.999)	-
Mean DBP, mmHg (SD) [a]		81 (10.89)	81 (10.93)	0.997 (0.995, 0.998)	-
Mean HbA1c, % (SD) [a]		8.05 (2.13)	7.95 (2.07)	0.98 (0.97, 0.99)	-
Mean serum creatinine, umol/L (SD) [a]		90.68 (24.46)	90.08 (28.96)	0.999 (0.998, 1.000)	-
Mean total cholesterol, mmol/L (SD) [a]		5.19 (1.30)	5.14 (1.29)	0.97 (0.96, 0.98)	0.97 (0.96, 0.98)
Median triglycerides, mmol/L (IQR) [a]		2.10 (1.50, 3.00)	2.20 (1.50, 3.28)	1.00 (0.99, 1.01)	-
Mean HDL-cholesterol , mmol/L (SD) [b]		1.25 (0.37)	1.25 (0.42)	0.99 (0.95, 1.03)	
Health Board Area					
Ayrshire & Arran		68.32	31.68	0.39 (0.37, 0.41)	0.29 (0.27, 0.31)
Borders		21.66	78.34	3.02 (2.67, 3.43)	3.43 (3.00, 3.93)
Dumfries & Galloway		44.93	55.07	1.02 (0.94, 1.12)	1.11 (1.00, 1.23)
Fife		57.40	42.60	0.62 (0.58, 0.66)	0.56 (0.52, 0.60)
Forth Valley		34.95	65.05	1.56 (1.45, 1.67)	1.71 (1.58, 1.85)
Grampian		68.12	31.88	0.39 (0.37, 0.41)	0.34 (0.32, 0.37)
Greater Glasgow		45.54	54.46	1 (Ref)	1 (Ref)
Highlands		51.08	48.92	0.80 (0.75, 0.86)	0.69 (0.63, 0.75)
Lanarkshire		62.53	37.47	0.50 (0.48, 0.53)	0.38 (0.36, 0.41)
Lothian		42.84	57.16	1.12 (1.06, 1.17)	1.18 (1.12, 1.26)

Variable	Observed HDL-cholesterol (53.84%)	Missing HDL-cholesterol (46.16%)	Univariate models OR (95% CI)	Multivariate model OR (95% CI)
Orkney	43.34	56.69	1.09 (0.87, 1.37)	0.88 (0.65, 1.19)
Shetland	62.77	37.23	0.50 (0.39, 0.63)	0.48 (0.36, 0.65)
Tayside	72.31	27.69	0.32 (0.30, 0.34)	0.26 (0.24, 0.28)
Western Isles	56.67	43.33	0.64 (0.51, 0.80)	0.53 (0.41, 0.70)

[a] Recorded at date of diagnosis
[b] Closest measurement recorded outside date of diagnosis time-frame

Table A. 8.18: Survival status by presence of missing data in HDL-cholesterol at diagnosis variable

	Complete HDL-cholesterol	Missing HDL-cholesterol
Crude HR (95% CI)	1	1.15 (1.10, 1.20)
Age and sex adjusted HR (95% CI)	1	1.17 (1.12, 1.22)
Deaths	4,062	4,091
Person-years	217,604	181,920
Crude mortality, deaths per 1,000 person-years	18.7	22.5
Age-standardised mortality, per 1,000 person years (95% CI)	12.5 (12.1, 13.0)	15.1 (14.5, 15.6)

Table A.8.19: Distribution of smoking status by imputed dataset number (only first 20 of 60 are presented.

Imputation Number	Current Smokers	Ex-Smokers	Never Smokers
1	16,721(22.63)	25,245 (34.17)	31,924 (43.20)
2	16,682 (22.58)	25,324 (34.27)	31,884 (43.15)
3	16,724 (22.63)	25,250 (34.17)	31,916 (43.19)
4	16,719 (22.63)	25,349 (34.31)	31,822 (43.07)
5	16,721 (22.63)	25,280 (34.21)	31,889 (43.16)
6	16,684 (22.58)	25,277 (34.21)	31,929 (43.21)
7	16,692 (22.59)	25,317 (34.26)	31,881 (43.15)
8	16,690 (22.59)	25,236 (34.15)	31,964 (43.26)
9	16,763 (22.69)	25,211 (34.12)	31,916 (43.19)
10	16,619(22.49)	25,367 (34.33)	31,904 (43.18)
11	16,709 (22.61)	25,301 (34.24)	31,880 (43.15)
12	16,779 (22.71)	25,207 (34.11)	31,904 (43.18)
13	16,773 (22.70)	25,245 (34.17)	31,872 (43.13)
14	16,723 (22.63)	25,267 (34.19)	31,900 (43.17)
15	16,720 (22.63)	25,325 (34.27)	31,845 (43.10)
16	16,632 (22.51)	25,314 (34.26)	31,944 (43.23)
17	16,648 (22.53)	25,297 (34.24)	31,945 (43.23)
18	16,651 (22.53)	25,249 (34.17)	31,990 (43.29)
19	16,738 (22.65)	25,182 (34.08)	31,970 (43.27)
20	16,718 (22.63)	25,257(34.18)	31,915 (43.19)