Measuring the association between body mass index and all-cause mortality in the presence of missing data: analyses from the Scottish national diabetes register

Title: Measuring the association between body mass index and all-cause mortality in the presence of missing data: analyses from the Scottish national diabetes register

Running Head: Missing data in the Scottish diabetes register

Authors: Stephanie H Read[1], Steff Lewis[1], Nynke Halbesma[1], Sarah H Wild[1]

[1] Usher Institute of Population Health Sciences & Informatics, University of Edinburgh, UK.

Corresponding author/Main point of contact:
Dr Stephanie H Read,
Usher Institute of Population Health Sciences & Informatics,
University of Edinburgh,
Teviot Place,
Edinburgh, UK.
EH8 9AG
Tel: (+44) 131 651 1398
Email: Stephanie.Read@ed.ac.uk

Abstract

Incorrectly handling missing data can lead to imprecise and biased estimates. We describe the effect of applying different approaches to handling missing data in an analysis of the association between body mass index and all-cause mortality in people with type 2 diabetes. Data from the Scottish diabetes register linked to hospital admissions data and death registrations were used. The analysis was based on people diagnosed with type 2 diabetes between 2004 and 2011 with follow-up until 2014. The association between body mass index and mortality was investigated using Cox proportional hazard models with comparison of findings using four different missing data methods; complete case analysis, two multiple imputation models and nearest neighbour imputation. There were 124,451 cases of type 2 diabetes, among which there were 17,085 deaths during 787,275 person-years of follow-up. Patients with missing data (24.8%) had higher mortality than those without (Adjusted hazard ratio: 1.36 [95% confidence interval: 1.31-1.41]). A U-shaped relationship between body mass index and mortality was observed, with the lowest hazard ratios occurring amongst moderately obese people, regardless of the chosen approach for handling missing data. Missing data may affect absolute and relative risk estimates differently and should be considered in analyses of routine data.

Key words: Diabetes mellitus, methods, obesity, research design.

Abbreviations: BMI – body mass index, MCAR – Missing completely at random, MAR – Missing at random, MNAR - Missing not at random, MICE – Multiple imputation using chained equations, MVN – Multiple imputation using multivariate normal imputation, OR – odds ratio, HR – Hazard ratio, SCI-Diabetes, Scottish Care Information – Diabetes, CPRD – Clinical Practice Research Datalink, THIN – The Health Improvement Network.

INTRODUCTION

Epidemiological studies which utilise electronic health records, in which data collection is typically clinically driven are often hindered by the presence of missing data. Despite some well-known flaws,[1-4] the exclusion of cases with incomplete data, known as complete case analysis remains the most popular approach for handling missing data.[5-7]

An alternative method for handling missing data is multiple imputation. This approach is steadily becoming more popular through improved accessibility in standard statistical software packages.[8] Multiple imputation involves the production of several plausible imputed datasets using information from the observed data, the separate analysis of each imputed dataset and finally the pooling together of estimates. There are two main imputation models that are routinely used; multiple imputation using chained equations or multivariate normal imputation.

The influence of unobserved data has yet to be explored in analyses investigating the possibility of an obesity paradox, whereby obesity confers a survival advantage over normal weight individuals among people with type 2 diabetes. Recent evidence based upon complete case analyses using datasets with missing data proportions of up to 56%, has indicated that being overweight or obese may be associated with lower mortality compared to being normal weight in people with type 2 diabetes.[9-13]

This study investigates the influence of missing data on the estimation of the association between body mass index (BMI) at diagnosis with type 2 diabetes and all-cause mortality using a contemporary population-based diabetes register in Scotland. Absolute and relative mortality estimates of the association between BMI

and mortality following complete case analysis and three imputation approaches are compared.

METHODS:

Data

Data were obtained from a 2011 extract of the Scottish Care Information – Diabetes (SCI-Diabetes) dataset, a national register of patients with diagnosed diabetes in Scotland. This database captures demographic information, including an area-based measure of deprivation, the Scottish Index of Multiple Deprivation and key diabetes-related clinical data from over 99% of general practices and all hospital diabetes clinics for adults in Scotland. The register is thought to be complete from 2004 onwards and incident type 2 diabetes cases occurring between January 2004 and June 2011 among people aged 30 years or over were included in this study. To reduce the risk of reverse causation from the effect of chronic diseases on BMI, people who died within two years of diagnosis of type 2 diabetes were excluded from all analyses.

Clinical characteristics at diagnosis, including BMI, blood pressure, lipid profiles and glycated haemoglobin were available from the SCI-Diabetes register. Measurements at diagnosis were defined as those recorded within one year prior to or two months (60 days) following diagnosis. This definition was chosen to reduce the impact of the uptake of lifestyle advice and diabetes treatment on measurements while limiting the extent of missing data. Exploratory analyses among patients with a BMI recording within one month of diagnosis and earlier or later measurements suggested that BMI measurements taken within one year prior to diagnosis correlated with the true BMI

within one month of diagnosis better than BMI values recorded within one year following diagnosis.

Death data were obtained from linkage of SCI-Diabetes to the National Records of Scotland mortality register using the community health index, a unique patient identifier. Timing of data linkage enabled follow-up until 31st May 2014.

Hospital admission data were obtained from the Scottish Morbidity Records dataset and were used to develop a Charlson Comorbidity Index which uses 19 pre-defined comorbid conditions to calculate a weighted score.[14]

All data were pseudonymised and permission for creation and analyses of the linked dataset was obtained from the Scottish multi-centre research ethics committee, the Privacy Advisory Committee of NHS National Services Scotland and the Caldicott Guardians for all Health Boards in Scotland.

Statistical Analyses

Associations between BMI at date of diagnosis, grouped into 5kg/m² categories between 20 and 45, and all-cause mortality were investigated using Cox proportional hazards regression. Patients whose BMI fell outside this range were included in the lowest and highest BMI categories as appropriate. Hazard ratios (HRs) were estimated relative to the BMI category containing the largest number of people and the category which has previously been shown to have the lowest mortality in these data (30 – 34.9).[12] Follow-up was defined as time from date of diagnosis of type 2 diabetes until date of death or study end date (31/05/2014), whichever came first. Violations to the proportional hazards assumption were investigated graphically using Kaplan-Meier plots and log-minus-log survival plots. Estimates were adjusted

for age, sex, Charlson comorbidity index (No comorbid conditions/One or more comorbid conditions), smoking status at diagnosis (Never/Former/Current) and quintiles of the Scottish index of multiple deprivation. Age-standardised mortality rates for each category of BMI were calculated using the 2013 European standard population.

Analyses were conducted in Stata, version 11.2, College Station, Texas, StataCorpLP.[15]

Missing Data

The extent of incomplete data in all variables included in the analysis model and the plausibility of the mechanisms of missingness were investigated. According to Rubin's classification system[16] there are three broad mechanisms of missingness; Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR). Data are MCAR when missingness does not depend on any observed or unobserved data and therefore the missing data should form a random subsample of the full dataset. Complete case analysis produces unbiased estimates when data are MCAR or when missingness on predictors is not dependent on the outcome. Data are MAR when the probability that data are missing is dependent on the observed data. Multiple imputation generally assumes data are MAR. Finally, data may be MNAR if the probability that data are missing is dependent on unobserved data, such as the missing values themselves or some unobserved characteristics.

To assess the plausibility of the MCAR assumption, comparisons between patients with and without observed data were made using means for normally distributed

variables, medians for non-normally distributed variables and percentages for categorical variables. Systematic differences in demographic and clinical characteristics between these groups would indicate a possible violation of the MCAR assumption made by complete case analysis. Little's test for MCAR was also applied.[17]

Cox proportional hazards models adjusted for age, sex and deprivation were used to identify differences in survival between patients with and without complete data. To assess the plausibility of the MAR assumption, predictors of missing data were identified using univariate and multivariate logistic regression analyses, whereby the outcome was a missing indicator variable.

Four methods for handling unobserved data were applied.

1. Complete case analysis -

   Patients with unobserved data on any variable to be included in the analysis model were excluded.
2. Nearest neighbour imputation -

   Missing values were replaced with BMI measurements recorded closest to date of diagnosis. Where patients had two measurements recorded within equitable timing before and after diagnosis, the measurement recorded before diagnosis was used.
3. Multiple imputation using chained equations (MICE) -

   STATA'S *ice* command was used to generate 50 imputed datasets.[18] The number of imputations was chosen based on the fraction of incomplete cases, as advocated by Bodner.[19] Full details of this method are provided in Web Appendix 1.

4. Multiple imputation using multivariate normal imputation (MVN) - STATA's *mi impute mvn* command was used to generate 50 imputed datasets. Again, further details of this method are provided in Web Appendix 1.

RESULTS

Between January 2004 and May 2011, there were 134,538 incident cases of type 2 diabetes in Scotland. People who were aged below 30 years at time of diabetes diagnosis (n=1,211) or who died within two years of follow-up (n=6,452) were excluded, leaving a final sample size of 124,451 incident cases of type 2 diabetes. A case-flow schedule is presented in Web Figure 1.

There were 17,085 deaths during 787,275 person-years and 21.7 deaths/1000 person-years in the study population. The median follow-up time was 6.1 years and the mean (SD) of BMI at diagnosis was 32.4 (6.6). Patient characteristics by categories of BMI are presented in **Table 1**. Increasing BMI was associated with lower age at diagnosis of diabetes, small proportions of people with comorbid conditions and more never smokers. There was an inverse association between BMI category and crude proportions of people that died during follow-up, though this association was confounded by age (Web Table 1).

Of the 124,451 members of the cohort, 93,622 people (75.2%) had complete data on age at diagnosis, sex, vital status, deprivation status, Charlson comorbidity status, smoking status at diagnosis and BMI at diagnosis. Ten percent of patients were without a smoking status at diagnosis recording, 17% did not have a BMI at date of diagnosis and 2% did not have a recorded deprivation status. There were 1,519

patients without a single BMI measurement who were excluded from the nearest neighbour imputation analyses.

Among patients with complete data, there were 12,575 deaths during 592,805 person-years, representing a crude mortality rate of 21.2 deaths/1000 person-years. For patients with incomplete data, there were 4510 deaths during 194,470 person-years and 23.2 deaths/1000 person-years. HRs (95% confidence intervals) adjusted for age, sex and deprivation indicated higher mortality in people with incomplete data compared to people with complete data (HR: 1.36 [1.31, 1.41]). A p-value of <0.000 was obtained from Little's test of MCAR, indicating data were unlikely to be MCAR. **Table 2** and Web Table 2 presents a comparison of characteristics for people with and without complete data. Briefly, women were more likely to have incomplete data than men (adjusted odds ratio (aOR): 1.16 [1.13, 1.16]), whilst older people were less likely to have incomplete data than younger people (aOR: 0.95 [0.94, 0.96]). Patients with comorbid conditions were more likely to have unobserved data (aOR: 1.24 [1.20, 1.27]), as were people who were deceased at study end (aOR: 1.04 [1.03, 1.05]).

To investigate whether missingness in BMI and smoking status were dependent on their values, we compared BMI and smoking status measurements recorded at date of diagnosis to earlier available measurements, where possible. According to these comparisons, BMI measurements recorded at diagnosis correlated strongly with earlier BMI measurements (Pearson's correlation coefficient >0.9). A similar pattern was observed in smoking statuses.

Absolute estimates for the association between BMI and all-cause mortality following the application of four separate methods for handling missing data are presented in

Figure 1. Mortality was highest in the lowest BMI category regardless of the approach used for handling missing data. Absolute mortality gradually increased with increasing BMI among people with a BMI above 25. Mortality estimates were marginally lower and confidence intervals were wider for almost all BMI categories based on complete case analysis than when the multiple imputation approaches were used. Estimates from MVN and MICE were very similar.

Figure 2 presents the relative association between BMI and all-cause mortality. Across all methods for handling missing data there was a U-shaped relationship between BMI and all-cause mortality, with the lowest risk of mortality at 30 to <35 following adjustment for age, sex, smoking status, deprivation status and Charlson comorbidity index. A steeper U-shaped relationship was observed when nearest neighbour observation imputation was applied.

DISCUSSION

In this study, the extent of incomplete data in the population-based SCI-Diabetes register was investigated and their influence on estimates of the association between BMI and all-cause mortality was examined.

We found that despite recent improvements, incomplete data remain a considerable barrier to research using this database, a problem that is commonly observed in databases derived from routine health care. In patients diagnosed with type 2 diabetes between 2004 and 2011, a quarter of patients had missing data in variables relevant to important outcomes such as cardiovascular disease and cancer.

The distribution of unobserved data varied by patient characteristics. People with missing data were more likely to have comorbid conditions and have attenuated

survival rates. This pattern in which data completeness is related to patient survival has been reported in other observational studies [20-22] and may reflect the perceived lack of relevance of such factors for people with poor prognosis. This finding is indicative of a MAR mechanism and therefore undermines the likely accuracy of estimates from complete case analysis, as shown by its under-estimation of absolute mortality across categories of BMI.

In this study, longitudinal patient data were compared to observe potential differences in missingness by previous variable recordings and we report no difference in smoking status at diagnosis recording according to previous smoking status. This finding contrasts with findings from studies using primary care databases, including the Clinical Practice Research Datalink (CPRD)[23] and The Health Improvement Network (THIN).[24] In both THIN and the CPRD databases, results indicated that missingness of smoking status was related to smoking status, with smokers more likely to have complete data than never smokers. These conflicting findings may be explained by differences in the populations included in these healthcare registers. While THIN and CPRD include all patients registered at selected general practices across the United Kingdom, the Scottish diabetes register only includes patients with diabetes, for whom there are several indicators as part of the Quality of Outcome Framework, a pay for performance scheme in the United Kingdom to encourage widespread and regular risk factor recording. There are subsequently fewer incentives to record risk factors including smoking status in patients without pre-existing disease.

Regardless of the approach used for handling the unobserved data, estimates indicated the presence of an obesity paradox which is consistent with several other

studies which have investigated the relationship between BMI and all-cause mortality.[9, 13, 25-27] However, the problems surrounding missing data were rarely discussed in this body of literature and the approaches chosen to handle missing data were frequently not described. However if the findings from our study are similar in other populations it appears that the patterns observed are not an artefact arising from missing data.

The alternative methods to complete case analysis had a number of strengths and weaknesses. The development of packages for multiple imputation in all major statistical software programs including SAS, R and SPSS has ensured that this method is widely accessible.[8] However, in our analyses, the imputation of a large number of unobserved values using MICE and MVN required considerable time, in terms of specifying and running the imputation model, though advances in computational power should ensure that the latter problem of running large imputation models can be overcome in time. Furthermore, multiple imputation may produce biased results when data are MNAR, a setting which cannot be ruled out. Ongoing uncertainty regarding the best approach for handling non-linear relationships and interactions in multiple imputation is a further limitation of these methods.[28-31]

From our results, the more straightforward approach of nearest neighbour imputation may also be capable of producing valid results when the incomplete variable is unlikely to have changed considerably during the measurement period. However, this single imputation approach is not recommended as it will overestimate the precision of the estimates and cannot be used in the absence of repeated measurements.[32]

Our findings of similar estimates between the two multiple imputation methods, MICE and MVN was reassuring given the different assumptions made about the distribution of the data. In particular, MVN assumes a joint normal distribution of the data, an assumption that is violated in the presence of binary or categorical variables. Previous studies have shown that the MVN model is relatively robust to departures from the normal distribution.[32-34] In a simulation study conducted in a political research setting, the authors reported that MVN and MICE models performed similarly well when continuous variables were imputed which did not exhibit a multivariate normal distribution.[34] However, when imputing categorical variables, MICE performed better. Another simulation study reported a reasonable performance of MVN when the normality assumption did not hold, particularly when the sample size was large.[33]

To handle ordinal data in MVN, it is possible to impute ordinal data in MVN as either indicator variables or continuous variables. However, recent research has indicated that the latter method is likely to distort non-linear relationships between imputed covariates variables and the outcome of interest.[35] In our study, we used the MVN model and imputed categorical and ordinal variables as a set of indicator variables and used a simple rounding approach to ascertain the variables imputed value, an approach that has previously been suggested to introduce bias.[35, 36] Despite this, estimates from this study were not dissimilar to those obtained from MICE, which is more flexible in handling non-continuous data. Moreover, Lee and Carlin presented results which indicated that comparable results can be expected from MICE and MVN approaches in the context of linear regression even when simple approaches to rounding are used.[37]

A strength of this work was the application of four methods for handling unobserved data in a real-life setting using a population-based register of patients with diabetes. Many previous studies investigating the performance of missing data methods have used simulated data which may not adequately reflect the complexity of real-world data. Linkage of SCI-Diabetes to other datasets also ensured a large number of variables were available to investigate the possible mechanisms of missingness and to include in imputation models.

A major limitation of this work was that the true values of the unobserved data were unknown and so we could not ascertain which missing data methods provided the least biased results. Nonetheless, this reflects the situation in many analyses and we have illustrated the problems associated with handling missing data in electronic healthcare records.

Despite explanatory analyses indicating missingness was associated with patient survival, it is not possible to rule out the possibility that data were MNAR. However, we have tried to make the MAR assumption more plausible by including a large number of potential predictors of missingness in the imputation models, an approach which has been recommended over MNAR-specific methods.[32, 38]

A further limitation of this work is the omission of a maximum likelihood estimation approach for handling missing data. This broad set of approaches can be used when data are MAR to identify population parameter estimates which have the highest probability of producing the sample data and have been found to provide unbiased estimates when the missing data mechanism and multivariate normality assumptions are met.[39, 40] However, while maximum likelihood estimation approaches require fewer decisions when specifying the model, a limited number of statistical software

packages offer automated programmes for specifying these maximum likelihood estimation models, necessitating the need for manual specification instead [1]. We chose not to apply these techniques due to the inaccessibility of the approach and subsequent limited uptake by epidemiologists. Finally, data on physical activity and alcohol consumption levels were not available and so we were unable to adjust our analyses for the effect of these potential confounders.

Our findings have demonstrated the importance of exploring missing data problems in electronic healthcare records and the need to consider the likely influence of differences between patients with and without missing data on both absolute and relative risks. Our findings provide reassurance of the robustness of MVN models in the presence of non-continuous data. Further work is required to assess if these findings are applicable in a wider range of settings.

According to our findings, the presence of an obesity paradox in people with type 2 diabetes does not appear to be a consequence of bias due to incorrectly handled missing data.

Table 1. Characteristics of patients diagnosed with type 2 diabetes in Scotland between 01st January 2004 and 31st December 2011 by categories of body mass index (BMI) in patients with complete data.

Characteristic	BMI category[a]													
	<20 (n=810)		20 – 24.9 (n= 8574)		25 – 29.9 (n=28 332)		30 – 34.9 (n= 29 791)		35 – 39.9 (n=15 716)		40 – 44.9 (n=6588)		≥ 45 (n=3811)	
	Mean (SD)	(%)	Mean (SD)	(%)	Mean (SD)	(%)	Mean (SD)	(%)	Mean (SD)	(%)	Mean (SD)	(%)	Mean (SD)	(%)
Age, years[b]	68.8 (14.1)		67.9 (11.4)		65.4 (10.8)		62.6 (10.3)		60.0 (9.9)		57.8 (9.5)		55.7 (8.7)	
Male sex		38.2		52.6		61.8		59.3		50.6		41.2		33.6
Deceased at study-end		36.8		23.9		15.2		11.7		9.8		9.0		7.2
With ≥ 1 comorbidities[b]		33.3		27.4		25.2		24.4		23.6		21.5		21.8
SIMD quintile[b]														
Q1 (most deprived)		28.8		21.4		21.3		22.5		25.2		27.9		30.0
Q2		22.2		21.8		21.8		22.7		23.6		23.7		24.8
Q3		16.3		19.8		20.2		20.5		21.0		20.6		20.4
Q4		18.3		19.0		19.1		19.3		17.4		15.7		15.4
Q5 (least deprived)		14.4		18.1		17.6		15.1		12.7		12.1		9.3
Smoking status[c]														
Never		32.6		40.7		41.2		40.4		43.1		45.5		50.0
Former		21.2		31.7		37.4		39.5		37.3		35.4		31.6
Current		46.2		27.6		21.4		20.2		19.6		19.1		18.4
Follow-up, years[d]	5.3 (3.9, 7.2)		6.0 (4.3, 8.1)		6.2 (4.5, 8.2)		6.2 (4.5, 8.1)		6.2 (4.5, 8.1)		6.1 (4.5, 8.2)		6.1 (4.5, 8.1)	

Characteristic	BMI category[a]													
	<20 (n=810)		20 – 24.9 (n= 8574)		25 – 29.9 (n=28 332)		30 – 34.9 (n= 29 791)		35 – 39.9 (n=15 716)		40 – 44.9 (n=6588)		≥ 45 (n=3811)	
	Mean (SD)	(%)	Mean (SD)	(%)	Mean (SD)	(%)	Mean (SD)	(%)	Mean (SD)	(%)	Mean (SD)	(%)	Mean (SD)	(%)
Systolic blood pressure, mmHg[c]	133.6 (21.0)		136.9 (19.1)		138.2 (18.1)		138.8 (17.5)		139.8 (17.6)		140.6 (17.5)		141.2 (17.5)	
Glycated haemoglobin, %[c]	8.4 (3.1)		8.4 (2.6)		8.1 (2.2)		8.0 (2.1)		8.0 (2.0)		8.0 (1.9)		8.1 (1.9)	
Total cholesterol, mmol/L[c]	5.0 (1.2)		5.1 (1.3)		5.1 (1.3)		5.1 (1.3)		5.1 (1.2)		5.2 (1.2)		5.1 (1.1)	
HDL-cholesterol, mmol/L[c]	1.6 (0.5)		1.4 (0.5)		1.2 (0.4)		1.2 (0.4)		1.2 (0.4)		1.2 (0.3)		1.2 (0.4)	

Abbreviations: BMI, Body mass index; SD, Standard deviation; SIMD, Scottish index of multiple deprivation; HDL-cholesterol, High density lipoprotein–cholesterol

[a] BMI given as kg/m^2

[b] At diagnosis

[c] Value recorded closest to date of diagnosis of diabetes within 12 months prior to or 2 months following diagnosis

[d] Median follow-up in years (Interquartile range)

Table 2. Characteristics of patients diagnosed with type 2 diabetes in Scotland between 01st January 2004 and 31st December 2011 with and without complete data

Characteristic	All patients with complete data (n=93622)		Patients with incomplete data (n=24.8)	
	Mean (SD)	%	Mean (SD)	%
Age, years[b]	62.9 (12.2)		56.5 (15.5)	
Male sex		55.5		55.1
Deceased at study-end		13.4		14.6
With ≥ 1 comorbidities[b]		24.5		22.1
SIMD quintile[b]				
Q1 (most deprived)		23.2		23.2
Q2		22.7		22.7
Q3		20.4		20.6
Q4		18.5		19.2
Q5 (least deprived)		15.3		14.4
Smoking status[c]				
Never		41.8		44.4
Former		37.0		33.3
Current		21.2		22.3
Follow-up, years[d]	6.1 (4.4, 8.2)		6.1 (4.3, 8.3)	
Systolic blood pressure, mmHg[c]	138.8 (17.9)		140.6 (18.7)	
Glycated haemoglobin, %[c]	8.1 (2.2)		8.3 (2.3)	
Total cholesterol, mmol/L[c]	5.1 (1.3)		5.2 (1.3)	
HDL-cholesterol, mmol/L[c]	1.2 (0.4)		1.2 (0.4)	

Abbreviations: BMI, Body mass index; SD, Standard deviation; SIMD, Scottish index of multiple deprivation; HDL-cholesterol, High density lipoprotein–cholesterol

[a] BMI given as kg/m^2

[b] At diagnosis

[c] Value recorded closest to date of diagnosis of diabetes within 12 months prior to or 2 months following diagnosis

[d] Median follow-up in years (Interquartile range)

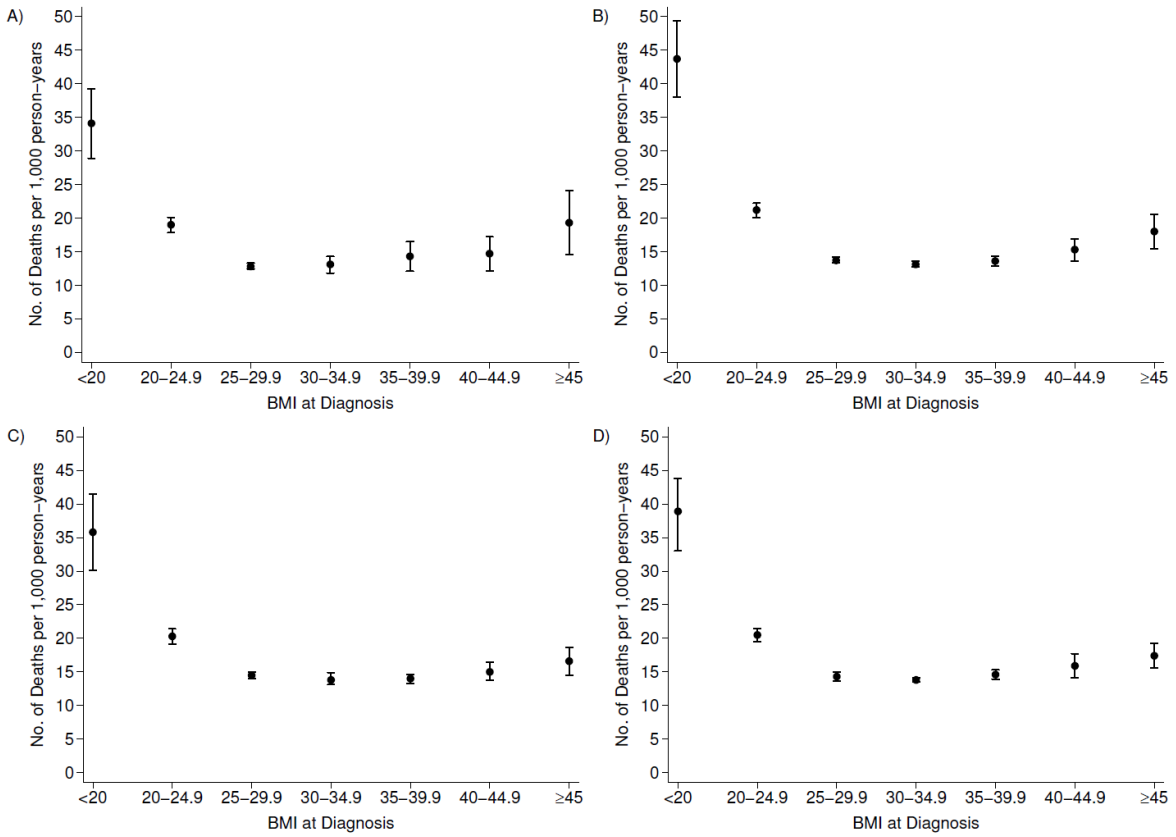

Figure 1: Age-standardised estimates (with 95% confidence intervals) of all-cause mortality rates by categories of body mass index (kg/m²) among people diagnosed with type 2 diabetes in Scotland between 01 January 2004 and 31 December 20011. Estimated from four approaches to handling the unobserved data; A) complete case analysis, B) nearest neighbour imputation, C) multiple imputation using chained equations and D) multiple imputation using multivariate normal imputation.

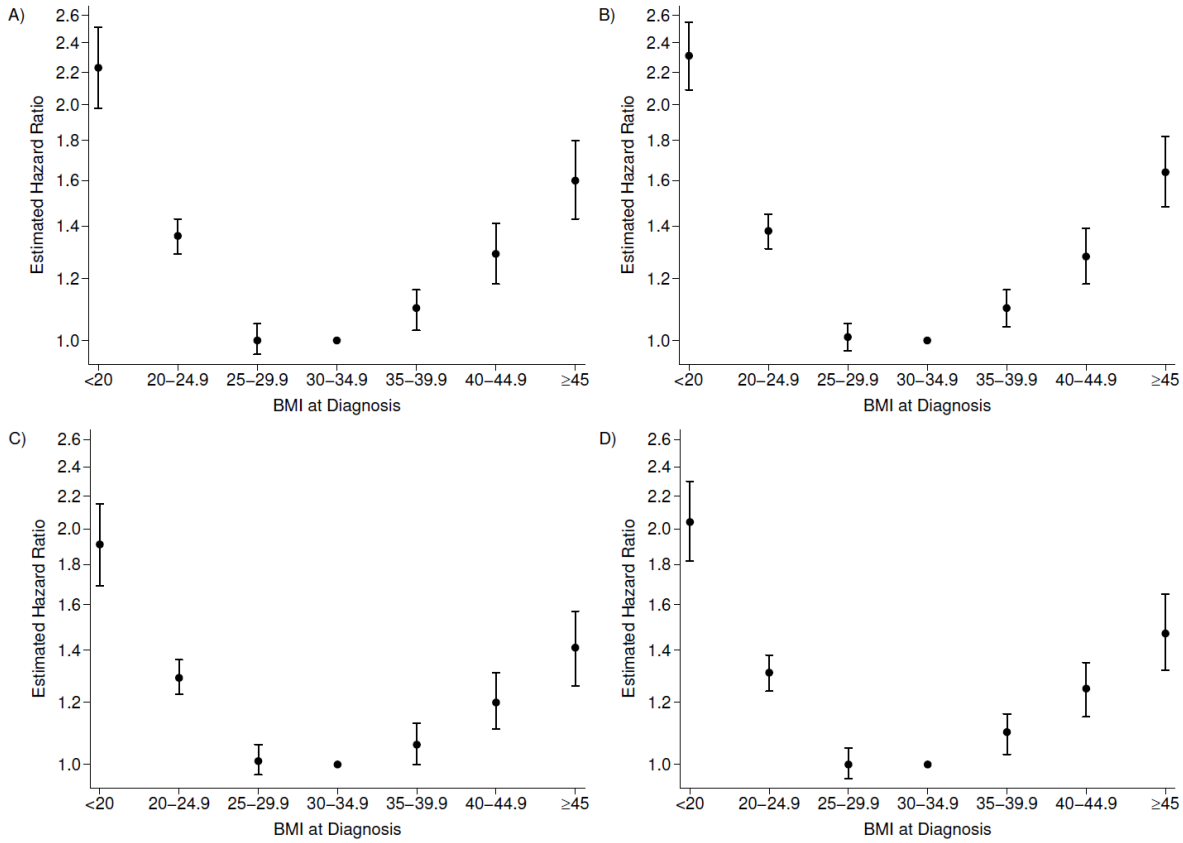

Figure 2. Hazard ratio estimates (with 95% confidence intervals) of the association between all-cause mortality and categories of body mass index (kg/m²) among people diagnosed with type 2 diabetes in Scotland between 01 January 2004 and 31 December 2011. Analyses are adjusted for age at diagnosis, sex, smoking status and Charlson comorbidity index. Estimated from four approaches to handling the unobserved data; A) complete case analysis, B) nearest neighbour imputation, C) multiple imputation using chained equations and D) multiple imputation using multivariate normal imputation.

Author Affiliations: Usher Institute of Population Health Sciences & Informatics, University of Edinburgh, Edinburgh, UK, EH8 9AG (Stephanie H Read, Steff Lewis, Nynke Halbesma & Sarah H Wild).

Funding: This project was funded by a Medical Research Council Hubs for Trials Methodology Research doctoral studentship. Data linkage of the Scottish population-based register of people with diagnosed diabetes to mortality records and data management was funded by the Scottish Government through the Scottish Diabetes Group.

Acknowledgements: Some of the data were presented as an abstract at the 2014 European Diabetes Epidemiology Group meeting in Sardinia, Italy (28/03/14 to 01/04/2014).

Contributors: The study was conceived by Stephanie H Read, Sarah H Wild and Steff Lewis; data preparation and statistical analyses were carried out by Stephanie H Read. Stephanie H Read wrote the first draft of the paper. All authors contributed to the interpretation of the findings and the paper's critical revision. All authors have approved the final version of the manuscript.

Conflict of interest: Sarah H Wild received an honorarium from Global MedEd/ Astra Zeneca in September 2014 for contributing a lecture to a series of educational videos aimed at primary care professionals and specialists in the Middle East. All other authors declare that there is no duality of interest associated with their contribution to this manuscript.

References:

1. Enders, C.K., *Applied Missing Data Analysis*. First Edition ed. Methodology in the Social Sciences, ed. T.D. Little. 2010, New York, US: Guildford Press.
2. Little, R.J.A. and D.B. Rubin, *Statistical analysis with missing data*. 1st Edition ed. 1987, New Jersey: John Wiley & Sons.
3. Rubin, D.B., *Multiple Imputation After 18+ Years.* Journal of the American Statistical Association, 1996. **91**(434): p. 473-489.
4. Schafer, J.L., *Analysis of Incomplete Multivariate Data*. 1997, London, UK: Chapman & Hall Ltd.
5. Mackinnon, A., *The use and reporting of multiple imputation in medical research – a review.* Journal of Internal Medicine, 2010. **268**(6): p. 586-593.
6. Sterne, J.A.C., I.R. White, J.B. Carlin, et al., *Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls.* BMJ, 2009. **338**: p. 2393.
7. Wood, A.M., I.R. White, and S.G. Thompson, *Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals.* Clinical Trials, 2004. **1**(4): p. 368-376.
8. Horton, N.J. and K.P. Kleinman, *Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models.* Am Stat, 2007. **61**(1): p. 79-90.
9. Carnethon, M.R., P.J. De Chavez, M.L. Biggs, et al., *Association of weight status with mortality in adults with incident diabetes.* Jama, 2012. **308**(6): p. 581-90.
10. Khalangot, M., M. Tronko, V. Kravchenko, et al., *Body mass index and the risk of total and cardiovascular mortality among patients with type 2 diabetes: a large prospective study in Ukraine.* Heart, 2009. **95**(6): p. 454-460.
11. Kokkinos, P., J. Myers, C. Faselis, et al., *BMI–Mortality Paradox and Fitness in African American and Caucasian Men With Type 2 Diabetes.* Diabetes Care, 2012. **35**(5): p. 1021-7.
12. Logue, J., J.J. Walker, G. Leese, et al., *Association between BMI measured within a year after diagnosis of type 2 diabetes and mortality.* Diabetes Care, 2013. **36**(4): p. 887-93.
13. Mulnier, H.E., H.E. Seaman, V.S. Raleigh, et al., *Mortality in people with Type 2 diabetes in the UK.* Diabetic Medicine, 2006. **23**(5): p. 516-521.
14. Charlson, M.E., P. Pompei, K.L. Ales, et al., *A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation.* Journal of Chronic Diseases, 1987. **40**(5): p. 373-383.
15. StataCorp., *Stata 11 Base Reference Manual*. 2009, College Station, TX: Stata Press.
16. Rubin, D.B., *Inference and missing data.* Biometrika, 1976. **63**(3): p. 581-592.
17. Little, R.J.A., *A Test of Missing Completely at Random for Multivariate Data with Missing Values.* Journal of the American Statistical Association, 1988. **83**(404): p. 1198-1202.
18. Royston, P., *Multiple imputation of missing values: Update of ice.* Stata Journal, 2005. **5**(4): p. 527-536.

19. Bodner, T.E., *What Improves with Increased Missing Data Imputations?* Structural Equation Modeling: A Multidisciplinary Journal, 2008. **15**(4): p. 651-675.
20. Hippisley-Cox, J., C. Coupland, Y. Vinogradova, et al., *Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study.* BMJ, 2007. **335**(7611): p. 136.
21. Hotchkiss, J.W. and A.H. Leyland, *The relationship between body size and mortality in the linked Scottish Health Surveys: cross-sectional surveys with follow-up.* Int J Obes (Lond), 2011. **35**(6): p. 838-51.
22. van Buuren, S., H.C. Boshuizen, and D.L. Knook, *Multiple imputation of missing blood pressure covariates in survival analysis.* Stat Med, 1999. **18**(6): p. 681-94.
23. Booth, H.P., A.T. Prevost, and M.C. Gulliford, *Validity of smoking prevalence estimates from primary care electronic health records compared with national population survey data for England, 2007 to 2011.* Pharmacoepidemiol Drug Saf, 2013. **22**(12): p. 1357-61.
24. Marston, L., J.R. Carpenter, K.R. Walters, et al., *Issues in multiple imputation of missing data for large general practice clinical databases.* Pharmacoepidemiology and Drug Safety, 2010. **19**(6): p. 618-626.
25. Church, T.S., M.J. LaMonte, C.E. Barlow, et al., *Cardiorespiratory fitness and body mass index as predictors of cardiovascular disease mortality among men with diabetes.* Arch Intern Med, 2005. **165**(18): p. 2114-20.
26. Hu, G., P. Jousilahti, N.C. Barengo, et al., *Physical activity, cardiovascular risk factors, and mortality among Finnish adults with diabetes.* Diabetes Care, 2005. **28**(4): p. 799-805.
27. Thomas, G., K. Khunti, V. Curcin, et al., *Obesity paradox in people newly diagnosed with type 2 diabetes with and without prior cardiovascular disease.* Diabetes Obes Metab, 2014. **16**(4): p. 317-25.
28. Morris, T.P., I.R. White, J.R. Carpenter, et al., *Combining fractional polynomial model building with multiple imputation.* Stat Med, 2015. **34**(25): p. 3298-317.
29. Seaman, S.R., J.W. Bartlett, and I.R. White, *Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods.* BMC Med Res Methodol, 2012. **12**: p. 46.
30. Bartlett, J.W., S.R. Seaman, I.R. White, et al., *Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model.* Statistical Methods in Medical Research, 2015. **24**(4): p. 462-487.
31. Shah, A.D., J.W. Bartlett, J. Carpenter, et al., *Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE: A CALIBER Study.* American Journal of Epidemiology, 2014. **179**(6): p. 764-774.
32. Schafer, J.L. and J.W. Graham, *Missing data: our view of the state of the art.* Psychol Methods, 2002. **7**(2): p. 147-77.
33. Demirtas, H., S.A. Freels, and R.M. Yucel, *Plausibility of multivariate normality assumption when multiply imputing non-Gaussian continuous outcomes: a simulation assessment.* Journal of Statistical Computation and Simulation, 2008. **78**(1): p. 69-84.

34. Kropko, J., B. Goodrich, A. Gelman, et al., *Multiple Imputation for Continuous and Categorical Data: Comparing Joint Multivariate Normal and Conditional Approaches.* Political Analysis, 2014. **22**(4): p. 497-519.
35. Lee, K.J., J.C. Galati, J.A. Simpson, et al., *Comparison of methods for imputing ordinal data using multivariate normal imputation: a case study of non-linear effects in a large cohort study.* Stat Med, 2012. **31**(30): p. 4164-74.
36. Bernaards, C.A., T.R. Belin, and J.L. Schafer, *Robustness of a multivariate normal approximation for imputation of incomplete binary data.* Stat Med, 2007. **26**(6): p. 1368-82.
37. Lee, K.J. and J.B. Carlin, *Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation.* Am J Epidemiol, 2010. **171**(5): p. 624-32.
38. White, I.R., P. Royston, and A.M. Wood, *Multiple imputation using chained equations: Issues and guidance for practice.* Statistics in Medicine, 2011. **30**(4): p. 377-399.
39. Peugh, J.L. and C.K. Enders, *Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement.* Review of Educational Research, 2004. **74**(4): p. 525-556.
40. Enders, C.K. and D.L. Bandalos, *The Relative Performance of Full Information Maximum Likelihood Estimation for Missing Data in Structural Equation Models.* Structural Equation Modeling: A Multidisciplinary Journal, 2001. **8**(3): p. 430-457.