Journal of Statistical Research *https://doi.org/10.47302/jsr.2020540203*
2020, Vol. 54, No. 2, pp. 131-145 *ISSN 0256 - 422 X*

USING EXTERNAL DATA TO INCORPORATE UNMEASURED CONFOUNDERS: A PLASMODE SIMULATION STUDY COMPARING ALTERNATIVE APPROACHES TO IMPUTE BODY MASS INDEX IN A STUDY OF THE RELATIONSHIP BETWEEN OSTEOARTHRITIS AND CARDIOVASCULAR DISEASE

MOHAMMAD ATIQUZZAMAN
School of Pharmacy, University of Otago
505b-18 Frederick Street, Dunedin 9016, New Zealand
Email: atiqmzaman@otago.ac.nz

MOHAMMAD EHSANUL KARIM[⋆], JACEK KOPEC, HUBERT WONG
School of Population and Public Health (SPPH), University of British Columbia
588-1081 Burrard St., Vancouver, BC V6Z1Y6, Canada
Email: ehsan.karim@ubc.ca, jkopec@arthritisresearch.ca, hubert.wong@ubc.ca

MARY A. DE VERA
Faculty of Pharmaceutical Sciences, University of British Columbia
4626-2405 Wesbrook Mall, Vancouver, BC, V6T 1Z3, Canada
Email: mdevera@mail.ubc.ca

ASLAM H. ANIS
School of Population and Public Health (SPPH), University of British Columbia
588-1081 Burrard St., Vancouver, BC V6Z1Y6, Canada
Email: aslam.anis@ubc.ca

[⋆] Corresponding author

SUMMARY

Background: Administrative databases do not contain Body Mass Index (BMI) information. In proportion-based imputation (PBI) technique, a BMI category is assigned to an individual according to the proportions observed in external survey data. Alternatively, BMI can be imputed using Multiple Imputation (MI).

Objectives: To compare MI with PBI to impute BMI variable in osteoarthritis (OA)-cardiovascular disease (CVD) relationship.

Research Design: plasmode simulation study.

Subjects: used publicly available data from the Canadian Community Health Survey (CCHS) cycles 1.1, 2.1, and 3.1.

Measures: BMI was set missing for everyone in the 500 simulated data created from CCHS 3.1 data. Dataset compiled from CCHS cycles 1.1 and 2.1 served as the external data (BMI observed). BMI missing in copies of simulated data was imputed using MI and PBI accessing observed BMI information in external data. After imputation, distribution of BMI variable and the adjusted odds ratio (aOR) estimated from multivariable logistic regression model were compared.

Results: Compared to PBI, MI produced proportions of individuals closer to the known proportions across the BMI categories except for the overweight category. Considering the known aOR of 1.59 (1.36, 1.82), BMI imputed using MI introduced less bias in OA-CVD association compared to PBI, the aOR was 1.62 (1.39, 1.86) and 1.66 (1.41, 1.90), respectively.

Conclusions: This is the first study to compare MI with PBI in the context of imputing BMI information that is not recorded at the database level. MI was superior to imputation method based on population-level proportions in imputing BMI missing for everyone in the simulated datasets.

Keywords and phrases: BMI, osteoarthritis, cardiovascular disease, plasmode simulation, Multiple Imputation

AMS Classification: 62P10

1 Introduction

Administrative databases are increasingly being used to undertake epidemiological research. As these data are primarily collected for administrative purposes, such as physician billings or hospital utilization tracking, researchers typically do not have any say over the type of information collected. Hence, administrative databases often do not include information on all the variables necessary to answer specific research questions. Specifically, data on potential confounders is often not collected. Consequently, not being able to adjust for potential confounders is a major challenge while using administrative databases (Solomon et al., 2010). For example, being obese or overweight and the related construct of body mass index (BMI) are an important risk factor for many diseases (Government of Canada, 2020; Cardiac Health Foundation Of Canada, 2020; National Institutes of Health, 2020; World Heart Federation, 2020; British Heart Foundation, 2020). However, BMI information is usually not included in administrative databases, including those in British Columbia (BC). If researchers perform regression adjustment or matching in the absence of a known confounder in the available dataset, the assumption of conditional exchangeability will not be valid anymore (Hernán and Robins, 2006). Failing to meet this identifiability condition for causal inference may lead to a biased estimate of the association of interest. To overcome this major limitation, many researchers under real-world conditions (e.g., dealing with electronic health records) attempt to impute an important confounder before trying to estimate the association of interest (Secrest et al., 2020; Sperrin and Martin, 2020). For example, in a longitudinal study investigating the association between osteoarthritis (OA) and cardiovascular diseases (CVD) using BC administrative databases, Rahman et al. (2013) imputed BMI categories for all study participants using data from another source, namely the Canadian Community Health Survey (CCHS). The imputation was done randomly in accordance with the proportions observed among individuals in CCHS who were grouped based on the OA exposure, CVD outcome and demographic variables (Rahman et al., 2013). In a separate BC population-based study investigating the association between rheumatoid arthritis (RA) and diabetes mellitus, Schmidt (2016) used a similar imputation based on proportions of obesity among RA patients versus the general population. However, none of the studies attempted to measure the extent of potential bias or confounding resulting from BMI being imputed using the proportion-based imputation method (Rahman et al., 2013; Schmidt, 2016).

A limitation of assigning a BMI category to an individual randomly according to a population-level proportion, as described in studies mentioned above (Rahman et al., 2013; Schmidt, 2016), is that this does not account for the imprecision associated with the imputation technique employed and may introduce bias (Sterne et al., 2009; Ratitch et al., 2013). Moreover, such imputation of BMI category of individuals in study data from administrative databases using population-level proportion does not account for individual-level heterogeneity. The resulting imputed values of BMI may be unrealistic for study individuals (e.g., may contradict the characteristics recorded in the administrative data). If, instead, we impute BMI values that are generated based on individual-level covariates that are predictive of BMI, those BMI values would be much more realistic and consistent with individual's characteristics. Rubin (1987) introduced the multiple imputation technique that has the potential to overcome these limitations of proportion-based imputation because of the following two essential features. First, multiple imputation accounts for the imprecision involved with the imputation, thus lessens bias and provides valid statistical inferences (Sterne et al., 2009; Ratitch et al., 2013). Second, a multivariable multiple imputation model uses variables recorded at the individual level, and therefore, imputes the missing BMI value from a series of plausible values, had it not been missing (Liu and De, 2015).

The primary objective of this study was to investigate, through a real-world case study, whether imputing an important study variable taking information from external data sources using multiple imputation approach instead of proportion-based imputation approach would result in less biased estimates of a given relationship. Towards that goal, we have investigated the relationship between OA and increased risk of CVD as an example of a relationship in which BMI is an important confounding variable for which information is usually not available in administrative databases, but such information could be imputed from an alternative source, such as a survey database. We hypothesize that imputing BMI at the individual level using the multiple imputation (Rubin, 1987) provides more realistic BMI values and introduces less bias in the estimated model coefficients compared to coefficients from a model in which BMI categories are imputed using proportion-based imputation (Rahman et al., 2013; Schmidt, 2016).

2 Methods

2.1 Data sources

2.1.1 Study dataset

To adequately assess the effect on bias and uncertainty associated with imputation of missing information on BMI, particularly in administrative data, we needed a dataset in which BMI is at least partially recorded at the individual level. Since administrative databases do not contain information on BMI, prospectively collecting BMI information for a large sample from administrative data would be both time and cost restrictive. Instead, we used data from the CCHS, a large national health survey data representing approximately 98% of the Canadian population. After starting in 2001, the CCHS was repeated every two years until 2005 collecting information on a large number of variables from approximately 130,000 respondents. We created a study dataset from CCHS cycle 3.1 (2005) (henceforth, Data.1) (Government of Canada, 2005b,a; Statistics Canada, 2006b). Specifically, Data.1 had complete information (no missing values) on OA exposure, CVD outcome and potential confounding variables including BMI. We created a dichotomous explanatory variable of OA using responses of two CCHS questions (Statistics Canada, 2006a). The first question, administered to all respondents, was “Do you have arthritis or rheumatism?” and the second question, which was only asked of respondents who answered yes to the first question, was “What type of arthritis?”. As respondents were not asked this second question in CCHS cycles after 3.1, formed the rationale for using CCHS cycle 3.1 for our study. The dichotomous outcome variable for CVD was obtained directly from a survey question ‘Do you have heart disease?’ that was asked of all respondents. The BMI was recorded as a categorical variable with underweight, normal weight, over weight and obese categories. Responses such as ‘Don’t Know’, ‘Refusal’ or ‘Not Stated’ were excluded. Comparison between MI and PBI on a single dataset context may be susceptible to chance findings. To overcome this limitation we needed to create multiple datasets. We created 500 simulated datasets (Data.2) from the Data.1 using plasmode simulation. Plasmode simulation is a validated statistical framework that helps replicate an empirical cohort study into multiple simulated datasets (Franklin et al., 2014). The core of this approach is to simulate dataset by resampling a real data in such a way that preserves the empirical associations among observed covariate and exposure. Based on the usefulness of this approach in epidemiologic studies using administrative databases,it recently gaining more popularity in various studies with methodological interest (Franklin et al., 2017; Edelmann et al., 2020; Jagdhuber et al., 2020; Karim et al., 2018). Ethics approval for this study using publicly available CCHS data was covered by item 7.10.3 in University of British Columbia’s Policy #89: Research and Other Studies Involving Human Subjects (Board of Governors , University of British Columbia, 2012) and Article 2.2 in of the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (TCPS 2) (Government of Canada, 2018).

2.1.2 Simulated datasets using plasmode simulation

In Monte Carlo simulation studies, covariates are often generated independently or by assuming a fixed correlation structure for the sake of simplicity (Fewell et al., 2007; Sahbaee et al., 2014). From an epidemiological standpoint, such oversimplified correlation structure is often inadequate in explaining the true interrelationship that exists among covariates under consideration, which in turn affects the conclusion of the study (Franklin et al., 2014). Inspired by a real administrative data plasmode simulation technique generates multiple simulated datasets in which the number of covariates as well as the complex covariance structure are preserved as in the original dataset. Consequently, plasmode simulation technique is considered to be realistic and resembles the features of a real epidemiologic dataset (Franklin et al., 2014, 2015). At first step we fitted a multivariable logistic regression model using Data.1 ($N = 84,452$). In this model the exposure was OA, outcome was CVD, and age, sex, level of education, household income level, physical activity index, smoking status, diabetes, hypertension and BMI category were entered as the confounding variables. We then used resampling with replacement from Data.1. We did not modify the OA exposure status and any of the co-variables during sampling (Franklin et al., 2014). As such, the associations among the study variables remained unchanged in the sampled populations. We replaced the Odds Ratio (OR) with 1.60 (estimated from the Data.1 in the first step). As such, we knew the true OR of 1.60 for all the simulated datasets upfront. We applied this outcome-generating model to the OA exposure and covariate data sampled with replacement from the study dataset. Finally, we created a binary CVD outcome status for each subject using the probability of outcome obtained from the outcome-generating model. We created the simulated datasets using this plasmode simulation technique starting from the sampling of individuals (Franklin et al., 2014). After each iteration, we monitored the average of the estimated OR that was stabilized at 35th iteration and did not change further. We created 500 simulated datasets (Data.2) ($n = 75,000$).

2.2 Setting missing information for BMI

We set the BMI variable missing for everyone in copies of the 500 simulated datasets (Data.3). This Data.3 mimics administrative data in which BMI is not recorded and missing for everyone in the database.

2.2.1 External survey data

Similar to the previously published proportion-based imputation method (Rahman et al., 2013), we created a large dataset (Data.4) by compiling data from CCHS cycles 1.1 (2001) (Government of Canada, 2003a) and 2.1 (2003) (Government of Canada, 2003b). This Data.4 served as the external survey data in both multiple imputation and proportion-based imputation methods. Table 1 summarizes the different datasets that are used in this study.

Table 1: Description of various datasets used to compare multiple imputation approach with proportion-based imputation method in imputing BMI variable missing for everyone in a database

Dataset	Description	No. of participants (N)	Comments
Data.1	Study dataset created from CCHS cycle 3.1 (2005)	84,452	Complete information on all variables including BMI
Data.2	500 simulated datasets created from study dataset using plasmode simulation	75,000	Complete information on all variables including BMI
Data.3	Copies of 500 simulated datasets but setting BMI as missing for everyone in the datasets	75,000	Missing information for BMI for everyone in the dataset. Complete information on other variables
Data.4	External survey data created by compiling data from CCHS cycles 1.1 (2001) and 2.1 (2003)	149,810	Complete information on all variables including BMI

2.2.2 Multiple imputation

In multiple imputation, multiple copies, usually three to five, of complete datasets are created by imputing the missing value (Rubin, 2004). Each of the complete datasets is then analyzed separately using an appropriate statistical method. Finally, the estimates from each of the complete datasets are combined using Rubin's rules to produce a single, pooled estimate (Sterne et al., 2009). The type of distribution under which a missing value will be imputed is an important consideration in selecting the multiple imputation model (Van Buuren, 2018). Unlike Markov Chain Monte Carlo method which assumes a joint multivariate normal distribution among all variables entered into the imputation model, the fully conditional specification method uses a separate conditional distribution to impute the missing variable (IDRE Stats, 2020). In addition, fully conditional specification method is advantageous because it allows selecting the imputation model based on the type of the missing variable. In this study, we aggregated data by setting Data.3 under Data.4. In this aggregated data BMI was observed for individuals from Data.4 and missing for individuals from Data.3. We implemented multiple imputation (with a number of imputations $= 5$) by PROC MI in SAS (version 9.4) using fully conditional specification logistic regression (Liu and De, 2015). We used information on age, sex, OA and CVD in the imputation model.

2.2.3 Proportion-based imputation

We grouped individuals in Data.4 based on OA, CVD, 10-year age category and sex. Within each group we calculated the proportions of individuals in each of the four BMI categories. Finally, individuals in Data.3 were grouped based on OA, CVD, 10-year age category and sex; and then similar proportions of BMI categories were imputed.

2.3 Analytic approach

We carried out the analysis in two steps. First, we implemented multiple imputation and proportion-based imputation to impute the missing BMI variable in Data.3. We then analyzed the imputed datasets and compared results with the known values estimated from Data.2. Based on the recommendations provided by Sterne et al. (2009), in evaluating the performance of imputation methods, we compared the proportion of individuals in each of the four BMI categories, and the ORs estimated from the multivariable logistic model (Please see appendix 1 for detailed reporting). In this model, CVD outcome was regressed on OA exposure adjusting for age, sex, physical activity index, level of education, household income level, smoking status, diabetes, hypertension and BMI category. To evaluate the performance of the plasmode simulation approach adopted in this study, we first estimated the proportion of individuals in each of the four BMI categories within each of the 500 simulated datasets (Data.2). Then we calculated the average of the proportions in each BMI category and compared that with the proportions observed in Data.1. We also fitted the multivariable logistic regression model in each of the 500 simulated datasets (Data.2) separately and calculated the average of the 500 ORs. The 95% confidence interval (CI) of the averaged OR was calculated by the percentile method. We compared this average OR from Data.2 with the OR estimated from Data.1.

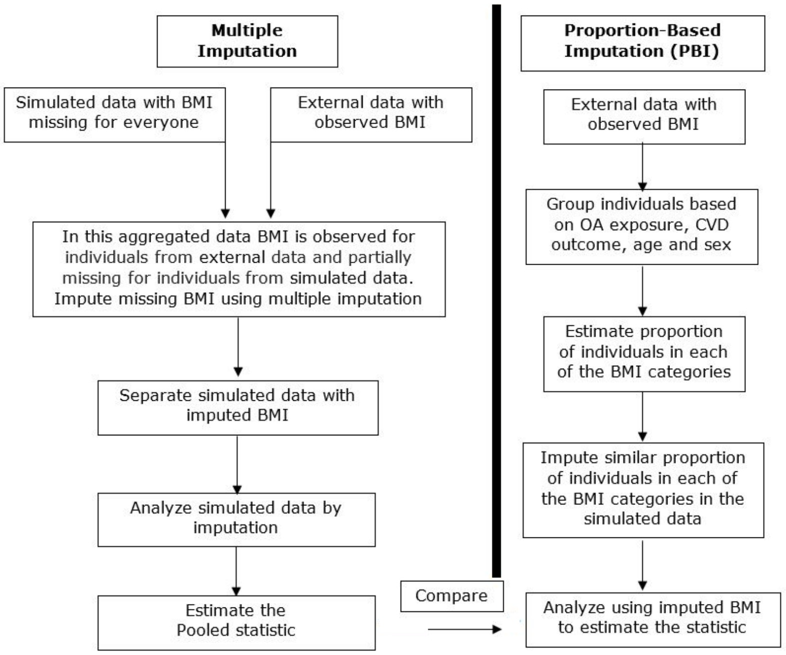
In multiple imputation method, we created five complete datasets for each of the 500 simulated datasets (Data.3) by imputing the missing BMI. We fitted the multivariable logistic regression model in each of the complete datasets separately and combined the ORs using Rubin's rule to obtain 500 pooled ORs. Finally, the average of the 500 pooled ORs was compared with the average of the 500 ORs estimated from Data.2. Rubin's rule in pooling the ORs accounted for the uncertainty associated with the imputed BMI (Rubin, 2004). This method takes the input of point estimates and standard errors from multiple imputed datasets and then generate a pooled estimate with an overall confidence interval. Rubin's rule assumes that the estimated statistics are approximately normally distributed. Similar to the method proposed by Ratitch et al. (2013), we applied log transformation to normalize the ORs. After combining, the pooled OR was back-transformed to its original log scale.

After imputing BMI using proportion-based imputation method, we estimated the average of the proportions in each BMI category observed in 500 simulated datasets (Data.3). We also analyzed the imputed datasets separately using the multivariable logistic regression model and calculated the average of the 500 ORs and 95% CI. Figure 1 presents the conceptual framework of the study to compare multiple imputation with PBI in imputing BMI variable missing for everyone in a study data using information from external database.

3 Results

The study data (Data.1) contained 84,452 survey respondents including 11,489 respondents with OA exposure and 4,963 respondents with CVD outcome. Table 2 presents the characteristics of the overall study sample by OA exposure status. The proportion of females among individuals with OA was substantially higher compared to individuals without OA, 71% versus 50%. The proportion of

Figure 1: Conceptual framework to compare multiple imputation with PBI in imputing BMI variable missing for everyone in a study data using information from an external database


Multiple Imputation
Simulated data with BMI missing for everyone
External data with observed BMI
In this aggregated data BMI is observed for individuals from external data and partially missing for individuals from simulated data. Impute missing BMI using multiple imputation
Separate simulated data with imputed BMI
Analyze simulated data by imputation
Estimate the Pooled statistic
Compare
Proportion-Based Imputation (PBI)
External data with observed BMI
Group individuals based on OA exposure, CVD outcome, age and sex
Estimate proportion of individuals in each of the BMI categories
Impute similar proportion of individuals in each of the BMI categories in the simulated data
Analyze using imputed BMI to estimate the statistic

individuals with OA was consistently higher among all groups over 50 years of age. For example, in the age group of 60-69 years, the proportion of individuals was 27% and 11% among people with OA and without OA, respectively. The prevalence of obesity was higher among individuals with OA. A substantial proportion (58%) of individuals were physically inactive among people with OA compared to non-OA individuals (48%). The prevalence of co-morbid disease conditions, such as diabetes and high blood pressure, was significantly higher among people with OA (p-value <0.0001). Individuals without OA appeared to have higher education and income level compared to people with OA.

Table 3 presents the proportion of individuals in each of the four BMI categories. Plasmode simulation appeared to produce simulated data that closely resembling the study data. The proportion of individuals in each of the BMI categories were comparable between Data.1 and Data.2; 2% versus 2% in underweight, 45% versus 46% in the normal weight, 35% versus 35% in the overweight and 18% versus 17% in the obese category. After imputing BMI categories missing for everyone in Data.3, both multiple imputation and proportion-based imputation methods underestimated the proportion of individuals in the normal weight category (Table 3). Compared to the known proportion of 46% normal weight individuals in Data.2, the proportion was only 38% when imputed using proportion-based imputation. In contrast, multiple imputation produced a less biased proportion of 41% in this category. Proportion-based imputation substantially overestimated the proportion of

Table 2: Characteristics of study sample ($n = 84,452$) for the study data (Data.1) created using data from CCHS cycle 3.1 (2005) according to osteoarthritis status

Variable		N (%) osteoarthritis	With osteoarthritis	Without	p-value
Study sample		84,452 (100%)	11,489 (13.60%)	72,963 (86.40%)	
CVD status	No	79,489 (94.12%)	9,750 (84.86%)	69,739 (95.58%)	<0.0001
	Yes	4,963 (5.86%)	1,739 (15.14%)	3,224 (4.42%)	
Sex	Female	44,455 (52.64%)	8,171 (71.12%)	36,284 (49.73%)	<0.0001
	Male	39,997 (47.36%)	3,318 (28.88%)	36,679 (50.27%)	
Age category	20-29 years	13,518 (16.00%)	150 (1.31%)	13,368 (18.32%)	<0.0001
	30-39 years	16,726 (19.81%)	413 (3.59%)	16,313 (22.36%)	
	40-49 years	16,233 (19.22%)	1,091 (9.49%)	15,142 (20.75%)	
	50-59 years	15,401 (18.23%)	2,712 (23.61%)	12,689 (17.39%)	
	60-69 years	11,483 (13.59%)	3,094 (26.93%)	8,389 (11.49%)	
	70-79 years	7,556 (8.95%)	2,644 (23.01%)	4,912 (6.73%)	
	$\geq$ 80 years	3,535 (4.19%)	1,385 (12.06%)	2,150 (2.95%)	
BMI category	Underweight	1,887 (2.23%)	231 (2.01%)	1,656 (2.27%)	<0.0001
	Normal weight	37,868 (44.84%)	4,127 (35.92%)	33,741 (46.24%)	
	Overweight	29,660 (35.12%)	4,208 (36.63%)	25,452 (34.88%)	
	Obese	15,037 (17.81%)	2,923 (25.44%)	12,114 (16.60%)	
Physical activity index	Inactive	41,890 (49.60%)	6,621 (57.63%)	35,269 (48.34%)	<0.0001
	Moderately active	22,121 (26.19%)	2,789 (24.28%)	19,332 (26.49%)	
	Active	20,441 (24.21%)	2,079 (18.10%)	18,362 (25.17%)	
Highest level of education	Less than secondary	15,730 (18.62%)	3,628 (31.59%)	12,102 (16.59%)	<0.0001
	Secondary graduate	12,973 (15.36%)	1,658 (14.43%)	11,315 (15.51%)	
	Some post- secondary	6,545 (7.75%)	794 (6.91%)	5,751 (7.88%)	
	College or university degree	49,204 (58.26%)	5,409 (47.08%)	43,795 (60.02%)	
Total household income	Less than $30,000	22,384 (26.51%)	5,112 (44.49%)	17,272 (23.67%)	<0.0001
	$30,000 to $49,999	19,166 (22.69%)	2,739 (23.84%)	16,427 (22.51%)	
	$50,000 to $79,999	21,505 (25.46%)	2,158 (18.78%)	19,347 (26.52%)	
	$80,000 or more	21,397 (25.34%)	1,480 (12.88%)	19,917 (27.29%)	
Smoking	Never smoked	26,094 (30.90%)	3,386 (29.47%)	22,708 (31.12%)	<0.0001
	Former occasional	13,196 (15.62%)	1,613 (14.04%)	11,583 (15.88%)	
	Former daily	23,653 (28.00%)	4,250 (36.99%)	19,403 (26.59%)	
	Current occasional	4,211 (4.99%)	312 (2.72%)	3,899 (5.34%)	
	Daily smoker	17,298 (20.48%)	1,928 (16.78%)	15,370 (21.07%)	
Diabetes	No	79,465 (94.10%)	10,075 (87.69%)	69,390 (95.10%)	<0.0001
	Yes	4,987 (5.90%)	1,414 (12.31%)	3,573 (4.90%)	
High blood pressure	No	68,711 (81.36%)	6,989 (60.83%)	61,722 (84.59%)	<0.0001
	Yes	15,741 (18.64%)	4,500 (39.17%)	11,241 (15.41%)	

Table 3: Comparison of proportion of individuals in each of the four BMI categories after imputing BMI for everyone in the simulated data

BMI category	Data.1 (%)	Data.2 (%)	Data.3 (%) BMI imputed using multiple imputation	Data.3 (%) BMI imputed using proportion-based method
Under weight	2.23	2.26	2.54	2.00
Normal weight	44.84	45.98	41.45	38.35
Over weight	35.12	34.93	38.66	34.77
Obese	17.81	16.83	17.34	24.87

Table 4: Comparing adjusted ORs estimated from multivariable logistic regression models

Data	Data.1	Data.2	Data.3	Data.3
Missing BMI category	None	None	BMI missing for everyone in the data	BMI missing for everyone in the data
Imputation method used	None	None	BMI imputed using multiple imputation	BMI imputed using proportion-based method
Without BMI[1]	1.62 (1.51, 1.73)	1.63 (1.19, 2.23)	1.63 (1.19, 2.23)	1.63 (1.19, 2.23)
Including BMI[1]	1.60 (1.49, 1.71)	1.59 (1.36, 1.82)	1.62 (1.39, 1.86)	1.66 (1.41, 1.90)

[1] adjusted for age, sex, physical activity index, education and income level, smoking status, diabetes and hypertension

obese individuals (25%) compared to the known proportion of 17% in Data.2. At the same time, the proportion of obese individuals was similar, 17% versus 17%, when BMI category was imputed using multiple imputation. Among the overweight individuals, although proportion-based imputation produced a proportion similar to that was observed in Data.2, multiple imputation overestimated the proportion by 3%.

Table 4 presents the adjusted ORs estimated using multivariable logistic regression. The odds of having CVD among people with OA was 1.60 times higher than that of among non-OA controls. The adjusted OR (95% CI) averaged after analyzing all the plasmode simulated datasets (Data.2) was 1.59 (1.36, 1.82), closely resembling the known OR of 1.60 estimated from Data.1. The BMI imputed using multiple imputation produced a less biased estimate of the OA-CVD association compared to the proportion-based imputation. The adjusted OR (95% CI) was 1.62 (1.39, 1.86). In contrast, BMI category imputed by proportion-based imputation resulted in an overestimate of the OA-CVD association, the adjusted OR (95% CI) was 1.66 (1.41, 1.90), much higher than that of observed in Data.2.

In a sensitivity analysis, we compared the existing multiple imputation model with a large multiple imputation model. In the later model we used information on other covariables including the level of education, household income level, physical activity index, smoking status, diabetes and hypertension, in addition to the age, sex, OA and CVD. Both the existing and large multiple imputation models produced similar proportions in each of the four BMI categories. In multivariable logistic regressions, BMI categories imputed by existing and large multiple imputation models resulted in similar point estimates of OA-CVD association. The adjusted OR (95% CI) was found to be 1.62 (1.39, 1.86) and 1.62 (1.38, 1.85) when imputed BMI from the existing and large multiple imputation models were entered into the multivariable logistic regression models, respectively.

4 Discussion

In the current work, we found that multiple imputation approach performed better than the proportion-based imputation method that has been previously used in imputing important variables in the studies based on administrative database. In a recent publication, a causal mediation analysis was used to identify the potential reasoning of the increased risk of CVD events observed in patients with OA, and based on subject-area expertise, BMI was listed as a known confounder in the causal relationship (Atiquzzaman et al., 2019; Yoshida and Desai, 2019). Although a few observational studies attempted to account for missing BMI by additionally accessing survey data, the effect of the imputed BMI in estimating an unbiased measure of exposure-outcome association remained unknown (Rahman et al., 2013; Schmidt, 2016). To the best of our knowledge, this is the first plasmode simulation-based study comparing multiple imputation with proportion-based imputation in imputing BMI category, a confounding variable in the association between OA and CVD. After imputing BMI category missing for everyone in the simulated data using information from external population-level survey data, compared to proportion-based imputation multiple imputation produced proportions closer to the known proportions across the BMI categories except overweight individuals. Also, BMI imputed using multiple imputation resulted in substantially less bias in the OA-CVD association. In contrast, the proportion-based imputation overestimated the association compared to the known point estimate.

Perhaps this can be explained by the inter-relationship among the numerous variables that are recorded at the individual level. Franklin et al. (2014) reported that the variables recorded for each individual have a unique covariance structure that can serve as a proxy for unmeasured confounders and be used to eliminate bias. It is more likely that the BMI category imputed by multiple imputation using individual-level information would fit in the existing covariate structure better than the BMI category imputed using population-level proportions. Consequently, the BMI category from multiple imputation method would perform more realistically in the multivariable outcome model, such as multivariable logistic regression employed in this analysis. The estimated ORs were not substantially different from each other after considering the overlap of the confidence intervals.

It is worth noting that, BMI variable was not measured in the dataset by design (not partially missing), making the missingness pattern for BMI variable to be missing completely at random (MCAR) by definition (Sterne et al., 2009). As such, from a theoretical perspective, in the absence of any unmeasured confounding, BMI variable imputed using any missing data imputation approach (e.g., MI approach) was not expected to have significant effect on the beta coefficient of OA exposure. However, besides problematic missingness patterns, confounding is also a major threat to estimating unbiased estimates from observational studies. In this particular work, the BMI variable itself is an important confounder established in the literature assessing the relationship between OA and CVD (Government of Canada, 2020; Cardiac Health Foundation Of Canada, 2020; National Institutes of Health, 2020; World Heart Federation, 2020; British Heart Foundation, 2020). Although we are using imputation methods (commonly used for missing data analysis) here to recover a variable, we emphasize that the purpose is not to deal with any problematic missingness patterns, but rather to produce a reasonable proxy variable that would reasonably reflect an unmeasured but known confounder variable (e.g., BMI in this context), which we can adjust in the analysis in an effort to reduce bias. There exists recent literature using similar approaches to replicate an important unmeasured variable in the epidemiologic data analysis context (Secrest et al., 2020; Sperrin and Martin, 2020). After imputing, we also looked into the contribution of the imputed BMI into the outcome models. There is a notion that the variables used to impute the BMI variable already contain the information into the outcome model. As such, we wanted to see if the imputed BMI variable contributes to the model when added along with other co-variables. The effect of adjusting for BMI was small (1% to 4%). This appeared to be reasonable because BMI was entered as a covariable into a multivariable logistic regression model. It was expected that one covariable would not make a substantial change on the estimated OR. Our findings indicated that multiple imputation was superior because the effect of adjusting for BMI imputed using proportion based imputation was in the unexpected direction compared to that observed in both study data and simulated datasets.

There are a number of strengths in our study. We used publicly available data from a Canadian population-level health survey. To the best of our knowledge, all the previous studies investigating the OA-CVD relationship were conducted using population-level health administrative data. As such, comparing MI with PBI in a study data derived from observational data would be closely related to the real-world problem addressed in this study. Plasmode simulation enabled us to create 500 large ($N = 75,000$) datasets to test the hypothesis in multiple data settings. In addition, plasmode simulation technique has the advantage of informing the true effect size set out during the data generation step that is useful in estimating the bias. One of the limitations of this study is that the plasmode simulation takes inspiration from a particular dataset to preserve realistic settings. We can be confident that the methods proposed here perform well for CCHS data under the settings we have evaluated. Future research using a variety of datasets is necessary to confirm the study findings. Another point to be noted is the cross-sectional nature of the health survey data. In longitudinal studies, individuals are prospectively followed for a long period of time. It is possible that the BMI of an individual may change over time. Imputation of BMI using survey data, either by multiple imputation or proportion-based imputation, cannot address this change in BMI over time. Under such circumstances, the imputed BMI can potentially serve as a proxy for the baseline BMI.

In conclusion, our simulation study showed that multiple imputation approach introduced less bias in the association between OA and CVD than conventional proportion-based imputation when BMI was imputed for everyone in the data using information from external survey data. A simple multiple imputation model including OA exposure, CVD outcome and demographic variables of age and sex adequately imputed BMI that performed realistically in estimating the OA-CVD association. Researchers often face the challenge of unmeasured variable. In imputing a study variable that is not recorded in a study data, multiple imputation is advantageous over the imputation method based on population-level proportions.

References

Atiquzzaman, M., Karim, M. E., Kopec, J., Wong, H., and Anis, A. H. (2019), "Role of nonsteroidal antiinflammatory drugs in the association between osteoarthritis and cardiovascular diseases: A longitudinal study," *Arthritis & Rheumatology*, 71, 1835–1843.

Board of Governors , University of British Columbia (2012), "Research Involving Human Participants," `http://universitycounsel.ubc.ca/files/2012/06/policy89.pdf`, accessed: 2020-11-25.

British Heart Foundation (2020), "Risk factors," `www.bhf.org.uk/heart-health/risk-factors`, accessed: 2020-10-13.

Cardiac Health Foundation Of Canada (2020), "Cardiac Health Foundation Of Canada," `cardiachealth.ca`, accessed: 2020-10-13.

Edelmann, D., Hummel, M., Hielscher, T., Saadati, M., and Benner, A. (2020), "Marginal variable screening for survival endpoints," *Biometrical Journal*, 62, 610–626.

Fewell, Z., Davey Smith, G., and Sterne, J. A. (2007), "The impact of residual and unmeasured confounding in epidemiologic studies: a simulation study," *American Journal of Epidemiology*, 166, 646–655.

Franklin, J. M., Eddings, W., Austin, P. C., Stuart, E. A., and Schneeweiss, S. (2017), "Comparing the performance of propensity score methods in healthcare database studies with rare outcomes," *Statistics in Medicine*, 36, 1946–1963.

Franklin, J. M., Eddings, W., Glynn, R. J., and Schneeweiss, S. (2015), "Regularized regression versus the high-dimensional propensity score for confounding adjustment in secondary database analyses," *American Journal of Epidemiology*, 182, 651–659.

Franklin, J. M., Schneeweiss, S., Polinski, J. M., and Rassen, J. A. (2014), "Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases," *Computational Statistics & Data Analysis*, 72, 219–226.

Government of Canada (2003a), "Canadian Community Health Survey (CCHS) Cycle 1.1," `http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=3359`, accessed: 2020-11-30.

— (2003b), “Canadian Community Health Survey (CCHS) Cycle 2.1,” http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=4995, accessed: 2020-11-25.

— (2005a), “Canadian Community Health Survey (CCHS),” http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SurvId=1630&InstaId=22642&SDDS=3226, accessed: 2020-11-25.

— (2005b), “Canadian Community Health Survey (CCHS) Cycle 3.1,” www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=22642, accessed: 2020-11-25.

— (2018), “Ethical Conduct for Research Involving Humans,” https://ethics.gc.ca/eng/documents/tcps2-2018-en-interactive-final.pdf, accessed: 2020-11-25.

— (2020), “Heart Disease in Canada,” www.canada.ca/en/public-health/services/publications/diseases-conditions/heart-disease-canada.html, accessed: 2020-10-13.

Hernán, M. A. and Robins, J. M. (2006), “Estimating causal effects from epidemiological data,” *Journal of Epidemiology & Community Health*, 60, 578–586.

IDRE Stats (2020), “Multiple Imputation in SAS Part 1,” https://stats.idre.ucla.edu/sas/seminars/multiple-imputation-in-sas/mi_new_1/, accessed: 2020-11-30.

Jagdhuber, R., Lang, M., Stenzl, A., Neuhaus, J., and Rahnenführer, J. (2020), “Cost-Constrained feature selection in binary classification: adaptations for greedy forward selection and genetic algorithms,” *BMC Bioinformatics*, 21, 1–21.

Karim, M. E., Pang, M., and Platt, R. W. (2018), “Can we train machine learning methods to outperform the high-dimensional propensity score algorithm?” *Epidemiology*, 29, 191–198.

Liu, Y. and De, A. (2015), “Multiple imputation by fully conditional specification for dealing with missing data in a large epidemiologic study,” *International Journal of Statistics in Medical Research*, 4, 287.

National Institutes of Health (2020), “What Are Coronary Heart Disease Risk Factors?” www.nhlbi.nih.gov/health/health-topics/topics/hd, accessed: 2020-10-13.

Rahman, M. M., Kopec, J. A., Anis, A. H., Cibere, J., and Goldsmith, C. H. (2013), “Risk of cardiovascular disease in patients with osteoarthritis: a prospective longitudinal study,” *Arthritis Care & Research*, 65, 1951–1958.

Ratitch, B., Lipkovich, I., and O’Kelly, M. (2013), “Combining analysis results from multiply imputed categorical data,” *PharmaSUG 2013-Paper SP03*, 2013, 1–19.

Rubin, D. B. (1987), *Statistical analysis with missing data*, Wiley.

— (2004), *Multiple imputation for nonresponse in surveys*, vol. 81, John Wiley & Sons.

Sahbaee, P., Segars, W. P., and Samei, E. (2014), "Patient-based estimation of organ dose for a population of 58 adult patients across 13 protocol categories," *Medical Physics*, 41, 072104.

Schmidt, T. J. (2016), "Cardiovascular disease prevention in rheumatoid arthritis: three population-based studies in British Columbia," Ph.D. thesis, University of British Columbia.

Secrest, M. H., Platt, R. W., Reynier, P., Dormuth, C. R., Benedetti, A., and Filion, K. B. (2020), "Multiple imputation for systematically missing confounders within a distributed data drug safety network: A simulation study and real-world example," *Pharmacoepidemiology and Drug Safety*, 29, 35–44.

Solomon, D. H., Rassen, J. A., Glynn, R. J., Lee, J., Levin, R., and Schneeweiss, S. (2010), "The comparative safety of analgesics in older adults with arthritis," *Archives of Internal Medicine*, 170, 1968–1978.

Sperrin, M. and Martin, G. P. (2020), "Multiple Imputation with Missing Indicators as Proxies for Unmeasured Variables: Simulation Study," *BMC Medical Research Methodology*, 20, 1–11.

Statistics Canada (2006a), "CCHS 3.1 Public Use Microdata File Data Dictionary," `https://www23.statcan.gc.ca/imdb-bmdi/pub/document/3226_D6_T9_V3-eng.pdf`, accessed: 2020-11-25.

— (2006b), "CCHS 3.1 Public Use Microdata File User Guide," `https://www23.statcan.gc.ca/imdb-bmdi/document/3226_D7_T9_V3-eng.pdf`, accessed: 2020-11-25.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., and Carpenter, J. R. (2009), "Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls," *BMJ*, 338.

Van Buuren, S. (2018), *Flexible imputation of missing data*, Chapman and Hall or CRC; Boca Raton, FL.

World Heart Federation (2020), "Cardiovascular disease risk factors," `www.world-heart-federation.org/press/fact-sheets/cardiovascular-disease-risk-factors/`, accessed: 2020-10-13.

Yoshida, K. and Desai, R. J. (2019), "Unraveling the Role of Nonsteroidal Antiinflammatory Drugs in the Link Between Osteoarthritis and Cardiovascular Disease via Causal Mediation Analysis: A Guide to Interpretation," *Arthritis & Rheumatology*, 71, 1776–1779.

Received: July 3, 2020

Accepted: December 8, 2020