Assessing Bias Associated With Missing Data from Joint Canada/U.S. Survey of Health: An Application

Sunita Ghosh[1], Punam Pahwa[2]
[1] Cross Cancer Institute, Alberta Cancer Board, Edmonton, Alberta, Canada
[2] Community Health and Epidemiology, University of Saskatchewan, Saskatoon, Canada

Abstract

Missing data are common in large scale surveys, arising mainly due to non-response in cross-sectional studies. Results are biased when missing data are ignored at the analysis stage. The present study aims to determine the bias associated when missingness is ignored and to investigate whether or not multiple imputation technique is a possible solution to address the issue of bias. The objective of the study will be achieved using the Public Use Micro Data File of Joint Canada/U.S. Survey of Health (JCUSH). JCUSH is a cross-sectional survey, which started collecting data in November 2002 and ended March 2003. The final sample contains 8,866 participants: 3,505 Canadian and 5,183 American participants. The bias will be tested using (i) available cases (only complete information), (ii) complete cases only (removing incomplete information), and (iii) multiple imputation method.

Keywords: Multiple imputation, complex survey data, JCUSH, available data, missing bias

1. Introduction

Analysis of data from large scale national surveys are complex as a researcher has to address the issue of complexity of the dataset like clustering, stratification and unequal probability of selection. The issue of missing data further complicates the situation. Survey data and other kinds of observational studies missing data are extremely common. Some of the reasons for missing data in cross-sectional survey data are incomplete response, respondents ignoring or refuse to answer a particular question. Ignoring the missing data and analyzing only complete data is easy to implement as it doesn't require any special method and can be incorporated using the appropriate statistical methods. However, the problems which arise due to ignoring the missing data is loss of information, reduction in sample size, and accounting only for complete cases can result in bias, and unrealistic estimates. The solution to this problem would be to account for missing data to reduce bias and provide valid and meaningful results.

In the recent decade or so considerable amount of work has been done in the area of missing data. In literature many methods has been proposed to deal with missing data. Some of the earlier work in this area used algorithmic and computational solution (Afifi and Elashoff 1966, Hartley and Hocking 1971). Since the algorithmic solution was computationally intense, hence general algorithm such as expectation-maximization (Dempster, Laird and Rubin 1977) and data imputation and augmentation procedure (Rubin 1987) are used. Details on these methods and for an excellent background on missing data can be found in books by Rubin (Rubin 1987), Allison (Allison 2002), Molenberghs (Molenberghs and Verbeke 2005), and Schafer (Schafer 1997).

Some of the most commonly used methods to deal with missing data analysis are last observation carried forward (LOCF), complete case, single imputation, and multiple imputations. Last observation carried forward is substituting the last measurement available whenever there is missing value. LOCF method is useful for longitudinal studies and is popular for both monotone and non-monotone missing pattern (Molenberghs, et al. 2005). Some of the limitations of LOCF methods are: If we have one measurement and all subsequent measurement missing then all the data are substituted with the same information, hence change over time cannot be studied. LOCF method increases information by treating imputed and actually observed values on the same footing. Complete case analysis includes only those observations for which complete measurements are available. The advantage of using this method is it's simplicity in nature and can be readily applied using any commercial software. The disadvantage of using this method is substantial loss of information. The result of this can impact the power and precision of the study heavily. In single imputation method the missing value is imputed with just one simulated value. The unconditional and the conditional mean are two special methods for single imputation method. Unconditional mean imputation method replaces missing values with an average value of other observed values on the same variable. The disadvantage or drawback of this method is that values imputed are unrelated to a subject's other measurement and result in bias. Another disadvantage is that it will be problematic when applying to categorical data. Conditional mean imputation method is another method where the imputation of missing measurement are based on other observed values on the same variable but based on condition. The major drawback of this method is that overestimates the precision. In the multiple imputation method the missing value is replaced by m>1 simulated version and the value of m may range from 3 to 10.

The present manuscript focuses on studying the missing data analysis to a logistic regression analysis, as outcome of interest is presence or absence of asthma. The objectives of the present analysis are to: (1) apply the multiple imputation (MI) method to the JCUSH dataset (2) compare the MI procedure with available and complete case analysis, and (3) to determine if there are any biases associated with missing data analysis.

Section 2 describes the dataset used for this particular analysis, Section 3 describes the statistical methods in detail, Section 4 focuses on the application of the methods to the dataset followed by results in Section 5. Conclusion remarks and discussion is addressed in Section 6.

2. Dataset Description

In November 2002, Health Statistics Division of Statistics Canada and the National Center of Health Statistics (NCHS) of the United States centers for Disease Control and Prevention started a collaborative project called the Joint Canada/United States Survey of Health (JCUSH). The objective behind a collaborative study was to produce a dataset to compare the Canadian and United States population. The target population was Canadian and United states population 18 years or older. In Canada, the JCUSH sample was stratified based on the provinces, and for US it divided into four geographic regions. For Canada, the ten provinces were divided into strata based on random digital dialing (RDD) frame. This RDD frame uses the elimination of non-working banks method (Norris and Paton 1991). The process starts with a list of all possible “hundred banks” sharing the same first eight digits of the ten digit telephone number. Within each RDD stratum, a bank is randomly chosen and the final two digit of the telephone number is randomly generated. This process is repeated until the required number of telephone number within each stratum is reached. For United States, a list-assisted RDD sampling frame (Lepkowski 1988) was used. The list-assisted method uses prefix area combinations of area codes and central office codes as the basis of constructing a sampling frame of banks of 100 consecutive telephone numbers. From the retained banks of 100 numbers, known as the 1+ directory-listed residential telephone numbers, a random sample of complete ten-digit telephone numbers is drawn in such a way that each number has a known and equal probability of being selected. In the final phase, the file of remaining telephone numbers (after removing the business and non-working numbers) is merged with the file of directory-listed residential numbers that were retained in the first phase. The numbers resulting from this phase were sent to the Computer-Assisted Telephone Interviewing (CATI) system. In the next phase of sampling, one person per household is selected at random from the RDD list to be interviewed. The sample was proportionally allocated within each stratum based on their population sizes to reduce bias. A more detailed description of the sampling design and sampling frame for Canada and United States can be found elsewhere[1].

The JCUSH questionnaire was administered using computer assisted telephone interview (CATI) method. Data collection was completed between November 2, 2002 and March 31, 2003. The questionnaire was administered in three languages: French and English for Canadian interviews and Spanish and English for American interviews. Interview duration was about 30 minutes.

The issue of unequal probability of selection was resolved by the weighting. The principle behind estimation in a probability sample such as the JCUSH is that each person in the sample represents himself/herself and a number of others not in the sample who have similar socio-demographic characteristics. The weighting phase is a step that calculates, for each person, his or her associated sampling weight. This weight appears on the micro data file and was used to derive meaningful estimates from the survey. The weight variable provided with the dataset is adjusted for household non-response, person-level weight, person-level non-response and post-stratification.

The overall response rate for Canada was 65.5% and 50.2% for United States. The overall response rate for Canada was calculated as: Household-level response rate (HHRR)* person-level response rate (PRR). Where

$$\mathrm{HHRR} = \frac{\#\,\text{of responding households}}{\text{all in} - \text{scope households}} \quad \text{and}$$

$$\mathrm{PRR} = \frac{\#\,\text{of responding persons}}{\text{all selected persons}}$$

[1] Joint Canada/United States Survey Of Health Public Use Microdata File User Guide, Statistics Canada and United States National Center for Health Statistics, June 2004

The overall response rate for United States was calculated as resolution rate (RR) * co-operation rate (CR). Where

$$\text{RR} = \frac{\text{\# out-of-scope} + \text{\# non-responding persons} + \text{\# responding persons}}{\text{Total of selected phone numbers}}$$ and

$$\text{CR} = \frac{\text{\# of responding persons}}{\text{\# of nonresponding persons} + \text{\# of responding persons}}$$

2. Statistical Methods

The complete case, available case and the missing data imputed by multiple imputation method were compared. Complete case analysis also known as case wise deletion or list wise deletion, excluded observation with censored values for the variable or variables of interest, thus limiting the analysis to those observation for which all values are observed. Available case analysis also known as pair wise deletion is a form of complete case analysis that limits analysis to cases with observed values for single variables that are being described or compared statistically.

Rubin (Rubin 1978) introduced the multiple imputation method. The basic principle of multiple imputation (MI) procedure is to replace the missing values with a set of M plausible values. The values which are drawn from the dataset, represents the uncertainty of the right value to impute. Using MI, the missing values are imputed and then analyzed using standard procedures available for complete cases. The basic assumption for multiple imputation is the missingness mechanism is Missing at random (MAR). When the missingness is independent of the unobserved measurement and conditional on the observed data, we refer to the missingness as MAR. The three steps or phases of multiple imputations using Rubin's terminology are best summarized by Molenberghs and Verbeke (Molenberghs, et al. 2005) is:

1. The missing data are imputed in M times to get M complete datasets.
2. These M datasets are analyzed using standard methods for complete data.
3. The results obtained from M analyses are combined into single inferences.

The reason for the popularity of the MI method is its high efficiency even for small values of M (Molenberghs, et al. 2005). Monte Carlo Markov Chain (MCMC) is used to generate pseudorandom draws from multidimensional and otherwise intractable probability distribution via Markov chains. In MCMC method, one constructs a Markov chain long enough for the distribution of the elements to stabilize to a common distribution. The steps for MCMC method assuming that data arise from multivariate normal distribution, as detailed in Molenberghs and Verbeke (Molenberghs, et al. 2005) are:

1. Starting values can be chosen by computing a vector of means and a covariance matrix from complete data.
2. These values obtained in the first step are used to estimate the prior distribution.
3. The values for missing data items are simulated by randomly selecting a value from the available distribution of values.
4. In posterior step, the posterior distribution of the mean and covariance are updated, by updating the parameters governing their distribution.
5. Based on the updated parameters, sampling from the posterior distribution of mean and covariance are done.

Imputation and the posterior steps are iterated until the distribution is stationary, and the imputation from the final iteration is used to yield a dataset with no missing values.

3. Application to JCUSH dataset

The JCUSH uses a complex survey design, which means that there is no simple formula that can be used to calculate variance estimates. Therefore, an approximate method is needed. The bootstrap method can be used to take into account the sample design information when calculating variance estimates. The bootstrap method, with the use of the Bootvar program provided with the data and discussed in the next subsection, is a method that is fairly easy to use (Rao and Wu 1988, Rao, Wu and Yue 1992).

The bootstrap method used with the JCUSH data involves the selection of simple random samples known as replicates, and the calculation of the variation in the estimates from replicate to replicate. In each replicate, the survey weight for each record is recalculated. These weights are adjusted and post-stratified according to population estimates information in the same way as the initial weights in order to obtain the final bootstrap weights.

The entire process (selecting simple random samples, recalculating and post-stratifying weights for each stratum) is repeated B times, where B is large. The JCUSH uses B=1,000 to produce 1,000 sets of bootstrap weights, which are provided with the Public Use Microdata File. To obtain a bootstrap variance estimator, the point estimate for each of the B samples must be calculated. The variance of these estimates is the bootstrap variance estimator. A program was developed and can perform all of these calculations for the user: the BOOTVAR program.

MI method was conducted using SAS software. PROC MI was used to generate the imputation. This procedure MI was used to create M (M=5) imputed dataset from the input dataset. MCMC imputation mechanism was used, and EM (Expectation-maximization) option as initial method was used. This EM option uses the means and standard deviation from available cases as the initial estimates for the EM algorithm. The final estimates obtained after applying the EM algorithm are then used to start the MCMC process. The imputed dataset was then analyzed using the GENMPD procedure. The final procedure is using the MIANALYZE procedure in SAS. This MIANALYZE procedure combines the M inferences into a single one.

For the complete case and available case analysis GENMOD procedure was used. This GENMOD procedure was used for the JCUSH dataset as it takes into account the stratification, clustering and unequal inclusion probability. In one of the previous paper by the Ghosh et al (Ghosh, Pahwa and Rennie 2008) it has been shown that this procedure is able to account for the complexity of the survey dataset just like any other design based approaches.

4. Results

The present analysis focused on adults aged 18-85 years, and this excluded pregnant or lactating females. The study aimed to study important risk factors of asthma prevalence. The average age of the respondents was 48.5 years (SD = 17.52), where average age of males being 47.63 years (SD = 16.88) and females were a bit older than males with average age being 49.26 years (SD = 17.98).

Table 1 provides the descriptive statistics of the important covariates to study asthma prevalence among adult population. The frequency (%) is provided for all potential risk factors of asthma prevalence. A total of 8566 adult participants met the inclusion criteria. Table 2 provides a summary of number of missing responses for each covariate and its corresponding percentage.

As discussed before, complete case analysis included only those respondents who had no missing information. Missing information for any of the covariate/risk factor was deleted from the dataset prior to analysis. Available case analysis included information on all covariates if the response was missing then it was coded as missing. Hence, for the complete case analysis a total of 6583 participants (About 23% of missing data were not included) were included who had complete information on all the covariates presented in Table 1. For the available case analysis, a total of 8032 (about 6.2% missing information were ignored) participants provided complete information. The ethnicity and income variables which had maximum missing values were not included in the analysis as these were not significant predictors of asthma prevalence.

Table 3 provides an overview of asthma prevalence in the JCUSH dataset. The prevalence of asthma and its 95% confidence interval for gender, smoking status and body mass index is provided. The results show higher prevalence of asthma among females, ex-smokers and obese participants.

The parameter estimates, its corresponding standard errors and odds ratio for the three models fitted are provided in Table 4. Model 1 presents the result of complete case analysis of covariates entered in the final model, model 2 provides the result for available case analysis and model 3 is for data imputed using multiple imputation procedure.

The parameter estimates for the model 2 and 3 were pretty close; however the parameter estimates were different for some variables for complete case analysis. The standard errors of model 1 (complete case analysis) and model 2 (available case analysis) were very similar. The standard errors obtained using the multiple imputation approach was very small. The smaller standard errors resulted in highly significant p-values for all the covariates of asthma prevalence. Female gender, smoking status, body mass index, and country of birth were positively associated with asthma prevalence. Education levels of the participant were negatively related with asthma prevalence. All of these associations were significant for model 3 whereas this was not true for model 1 or 2 (See Table 4).

5. Discussion

The result from current analysis suggests that there is a bias associated when missing data is ignored or not taken into account. Using the available statistical procedure in SAS, we imputed the missing values on important covariates listed in the JCUSH dataset. The differing results may indicate that there is a bias associated for partially observed data. Omitting all cases with missing data could have created selection bias in the analyzable sample with varying effect on different covariates in the model. One of our variables had more than 26% of missing data; completely ignoring that information can seriously bias the results. Complete case analysis or available case analysis approaches may have serious deficits if data are not MCAR (Kneipp and McIntosh 2001). Newgard and Haukoos

(Newgard and Haukoos 2007) in their study compared MI to complete case analysis, and the result if the study showed less bias generated by MI method with more statistical efficiency as compared to the complete case analysis.

Accounting for the missing data may account for the associated bias. The larger standard errors for complete case and available cases can be due to the sample size, as for multiple imputations we have complete information where as for other two cases we have partial information. The smaller standard errors using multiple imputations method results in higher statistical significance. The reduction in standard error for multiple imputation procedure is expected, due to the recovery of missing data (Sartori, Salvan and Thomaseth 2005).

The major concerns when there is data missing is lack of efficiency and bias resulting from differences between observed and missing data (Horton and Lipsitz 2001). Previous studies have shown that the complete case analysis is not able to account for missing data and using only partial information results in loss of efficiency and information (Arnold and Kronmal 2003, Ambler, Omar and Royston 2007, Horton and Kleinman 2007) . The reason for choosing multiple imputation procedure was due to the fact that this method is more efficient., Some of the other desirable features of MI is that introducing appropriate random error results in unbiased estimates of the parameter, repeated multiple imputation allows to get good estimates of the standard error and can be used for any kind of data (Allison 2000). The data must be missing at random (MAR), the model used to generate the imputed values must be “correct” in some sense and the model used for analysis must match up in some sense with the model used in imputation (Allison 2000) are some of the desirable properties of MI. Further details of these properties can be found in Rubin (Rubin 1987, 1996). If MAR assumption is satisfied then the methods of multiple imputations provide less bias than other approaches if the imputation is correctly specified (Horton, et al. 2001). Allison in his paper mentions that even if MAR assumption is satisfied, producing imputations that yield unbiased estimates of the desired parameters is not always easy (Allison 2000).

MI allows use of complete data methods for data analysis. MI simulates proper inferences from data; it also increases efficiency of the estimates because MI minimizes standard errors (Patrician 2002). The MAR assumption was not formally evaluated, as the information from missing data would be required to know if missing data was MCAR or MAR. The book by Schafer (Schafer 1997) modestly argues that with the use of rich multivariate data can provide protection against MAR violation and hence minimize bias. The paper by Faris et al (Faris, et al. 2002) supports the use of multiple imputations. Even when the MAR assumption is violated the multiple imputation method performs well than other ad hoc methods of handling missing data (Greenland and Finkle 1995, Schafer 1997).

Although MI is probably the most accurate and valid imputation method, it has several disadvantages. According to Rubin (Rubin 1987), the three disadvantages of multiple imputation more effort to create the multiple imputation, more time to run analysis and more space usage to store the created imputed dataset. The method itself is complex and utilizes advanced statistical modeling .

One of the major limitations of the present study was that other than multiple imputations no other method to account for missing data was used. The possibilities of bias resulting from poorly specified imputation model were not studied. The unique feature of the study was accounting for the complexity of the survey design, i.e. our analysis accounted for the stratification, clustering and unequal probability of selection when taking into account the missing data. Further studies will be helpful to compare the multiple imputation methods to other procedure available to impute missing data.

6. Conclusion

To conclude we can say that multiple imputation method provides statistically valid inferences; however as stated by Nicholas and Lipsitz (2001) that this powerful and useful tool can be dangerous if not used carefully. Despite these issues, multiple imputations is recommended for handling missing data, as completely ignoring the missing data can result in severe bias and misleading results. It is highly recommended that instead of using only complete cases for analysis; make use of the multiple imputation method, as most of the published work just deals with missing data by completely ignoring it, instead of accounting for it.

References

Afifi, A., and Elashoff, R. (1966), "Missing Observations in Multivariate Statistics I: Review of the Literature.," *journal of american statistical association*, 61, 595-604.

Allison, P. D. (2000), "Multiple Imputation for Missing Data: A Cautionary Tale.," *Sociological Methods and Research*, 28, 301-309.

Allison, P. D. (2002), *Missing Data*, SAGE University Papers.

Ambler, G., Omar, R. Z., and Royston, P. (2007), "A Comparison of Imputation Techniques for Handling Missing Predictor Values in a Risk Model with a Binary Outcome.," *Statistical Methods in Medical Research*, 16, 277-298.

Arnold, A. M., and Kronmal, R. A. (2003), "Multiple Imputation of Baseline Data in the Cardiovascular Health Study.," *American Journal of Epidemiology*, 157, 74-84.

Dempster, A. P., Laird, N., and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data Via the Em Algorithm (with Discussion)." *Journal of Royal Statistical Society, Series B*, 39, 1-38.

Faris, P. D., et al. (2002), "Multiple Imputation Versus Data Enhancement for Dealing with Missing Data in Observational Health Care Outcome Analysis," *Jouurnal of Clinical Epidemiology*, 55, 184-191.

Ghosh, S., Pahwa, P., and Rennie, D. (2008), "Comparison of Design-Based and Model-Based Methods to Estimate the Variance Using National Population Health Survey Data (1994-2003)," *Model Assisted Statistics and Application*, 3, 33-42.

Greenland, S., and Finkle, W. D. (1995), "A Critical Look at Methods for Handling Missing Covariates in Epidemiologic Regression Analyses.," *American Journal of Epidemiology*, 142, 1255-1264.

Hartley, H. O., and Hocking, R. (1971), "The Analysis of Incomplete Data.," *Biometrics*, 27, 7783-7808.

Horton, N. J., and Kleinman, K. P. (2007), "Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models.," *American Statistician*, 61, 79-90.

Horton, N. J., and Lipsitz, S. (2001), "Multiple Imputation in Practice: Comparison of Software Package for Regression Models with Missing Variables.," *The American Statistician.*, 55, 244-254.

Kneipp, S. M., and McIntosh, M. (2001), "Handling Missing Data in Nursing Research with Multiple Imputation.," *Nursing Research*, 50, 384-389.

Lepkowski, J. M. (1988), "Telephone Sampling Methods in the United States.," in *Telephone Survey Methodology*, ed. R. G. e. al, New York: John Wiley and Sons, pp. 73-98.

Molenberghs, G., and Verbeke, G. (2005), *Models for Discrete Longitudinal Data*, Springer.

Newgard, C. D., and Haukoos, J. S. (2007), "Advanced Statistics: Missing Data in Clinical Research-Part 2: Multiple Imputation.," *Academic Emergency Medicine*, 14, 669-678.

Norris, D. A., and Paton, D. G. (1991), "Canada's General Social Survey: Five Years of Experience.," *Survey Methodology*, 17, 227-240.

Patrician, P. A. (2002), "Multiple Imputation for Missing Data.," *Research in Nursing and Health*, 25, 76-84.

Rao, J. N. K., and Wu, C. F. J. (1988), "Resampling Inferences with Complex Survey Data.," *Journal of the American Statistical Association*, 83, 231-241.

Rao, J. N. K., Wu, C. F. J., and Yue, K. (1992), "Some Recent Work on Resampling Methods for Complex Surveys," *Survey Methodology*, 18, 209-217.

Rubin, D. B. (1978), "Multiple Imputations in Sample Surveys- a Phenomenological Bayesian Approach to Non-Response.," in *Imputation and Editing of Faulty or Missing Survey Data.*, Washington, D.C: US Department of Commerce., pp. 1-23.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley and Sons.

Rubin, D. B. (1996), "Multiple Imputation after 18+ Years.," *journal of american statistical association*, 91, 473-489.

Sartori, N., Salvan, A., and Thomaseth, K. (2005), "Multiple Imputation of Missing Values in Cancer Mortality Analysis with Estimated Exposure Dose.," *Computational Statistics and Data Analysis*, 49, 937-953.

Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data*, London: Chapman and Hill.

Table 1: Descriptive statistics, n(%), of potential risk factors and/or covariates of asthma prevalence

Variables	n (%)
Age (years)	
18-34 years	2084 (24.3)
35-46 years	2200 (25.7)
47-61 years	2099 (24.5)
62-85 years	2183 (25.5)
Gender	
Male	3834 (44.8)
Female	4732 (55.2)
Smoking Status	
Current smoker	1995 (23.4)
Ex-smoker	3536 (41.5)

Non-smoker	2998 (35.1)
Body mass index (BMI)	
Underweight	216 (2.6)
Normal weight	3731 (44.9)
Over weight	2797 (33.7)
Obese	1560 (18.8)
Education	
Less than high school	1344 (16.2)
High school	4200 (50.5)
Higher education	2767 (33.3)
Ethnicity	
White (Caucassian)	6655 (80.1)
Others	1652 (19.9)
Socio-economic status	
Low SES	1154 (17.0)
Low Middle SES	2399 (35.3)
High Middle SES	1731 (25.5)
High SES	1506 (22.2)
Country of birth	
Canada	2784 (33.3)
USA	4210 (50.4)
Other	1363 (16.3)

Table 2: Complete response and missing data information of covariates used in the analysis

Variables	Complete response (N=8566)	Missing	% Missing
Age	8566	0	0.000
Gender	8566	0	0.000
Smoking	8529	37	0.434
Body Mass Index	8304	262	3.155
Country of Birth	8537	209	2.448
Education	8311	255	3.068
Ethnicity	8307	259	3.118
Income	6790	1776	26.156
Asthma	8560	6	0.070

Table 3: Overall asthma prevalence

Variables	Prevalence	cil95	ciu95
Gender			
Males	9.87	8.53	11.21
Females	12.65	11.38	13.92
Smoking Status			
Current Smokers	11.06	9.09	13.03
Ex Smokers	12.67	11.19	14.15
Non Smokers	10.16	8.73	11.59
BMI			
Under Weight	13.80	6.89	20.71
Normal weight	9.65	8.36	10.94
Over weight	9.49	8.10	10.88
Obese	17.65	14.99	20.32

Table 4: Parameter estimates, standard errors and odds ratio of various risk factors of asthma prevalence using different methods to account for missingness in the data.

	Model 1			**Model 2**			**Model 3**		
	β	**S.E.**	**OR**	**β**	**S.E.**	**OR**	**β**	**S.E.**	**OR**
Intercept	-2.83	0.25	0.06	-2.97	0.22	0.05	-2.98	0.02	0.05
Age (62-85 years)									
18-34 years	0.48	0.16	1.61**	0.59	0.15	1.80***	0.64	0.01	1.90***
35-46 years	0.14	0.16	1.15	0.13	0.15	1.14	0.17	0.01	1.18***
47-61 years	0.14	0.16	1.15	0.23	0.14	1.26	0.24	0.01	1.27***
Gender (Male)									
female	0.39	0.11	1.48***	0.36	0.10	1.43***	0.33	0.00	1.39***
Smoking (Non-smoker)									
Current smoker	0.01	0.15	1.01	0.07	0.14	1.07	0.09	0.01	1.10**
Ex-Smoker	0.23	0.13	1.26	0.30	0.11	1.35**	0.31	0.01	1.37**
Body mass index (Normal weight)									
Under weight	-0.05	0.38	0.95	0.39	0.33	1.48	0.37	0.01	1.45***
Over weight	0.14	0.13	1.15	0.11	0.12	1.11	0.12	0.02	1.13***
Obese	0.69	0.14	2.00***	0.74	0.12	2.09***	0.76	0.03	2.13***
Country of birth (Other)									
Canada	0.45	0.19	1.57*	0.41	0.17	1.50*	0.35	0.04	1.42***
USA	0.54	0.19	1.72**	0.46	0.16	1.59**	0.42	0.02	1.53***
Education (Less than high school)									
High School	-0.46	0.19	0.63**	-0.33	0.17	0.72	-0.40	0.05	0.67***
University and College	-0.52	0.18	0.59**	-0.41	0.16	0.66**	-0.29	0.03	0.75***

Model 1= Complete case analysis; Model 2= Available case analysis; Model 3=Multiple imputation
Reference categories in parentheses
*** $p < 0.0001$
** $p < 0.001$
* $p < 0.05$