Ordinal Variable Imputation for Health Survey Data: A Comparison between Machine Learning and non-Machine Learning Methods

by

Sho Podolsky

Submitted in partial fulfilment of the requirements for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
July 2021

Dalhousie University is located in Mi'kma'ki, the ancestral and unceded territory of the Mi'kmaq. We are all Treaty people.

Table of Contents

List of Tables

List of Figures

Abstract

Introduction: Large amounts of data are available for analyses from survey datasets. However, missing data can potentially reduce statistical power and/or introduce bias into analyses when not addressed correctly. Data imputation methods can replace missing data with estimated values that are informed from existing data. Machine learning algorithms can improve the efficiency and accuracy of data imputation by automatically generating models that can fit to complex associations that may exist between variables in a dataset.

Methods: This thesis uses a cross-sectional simulation study of the Canadian Community Health Survey 2014 public use microdata file to induce missingness into annual total household income, an ordinal variable with 5 classes. The simulation study includes 5 imputation models from machine learning algorithms and 2 non-machine learning imputation models (ordinal logistic regression and predictive mean matching) for each simulated dataset for a total of 84 imputed datasets. The evaluation uses an ordinal-sensitive distance measure and class transition tables to compare imputation model performance.

Results: The imputation models from machine learning algorithms performed better than the non-machine learning imputation models with regards to the ordinal-sensitive distance measure (0.5-0.6 for machine learning vs 0.65-0.75 for non-machine learning, lower values indicate better performance). The class transition tables indicate that, while scoring above 80% accuracy in one class, machine learning models tend to overrepresent income classes that are easier to classify and produce imputed values that do not reflect the original class structure of the income variable. The machine learning models had very low accuracy (less than 5% in all algorithms except one) for the income class that was the most underrepresented in the imputed data. The non-machine learning models produced imputed values that reflected the original income class structure well but had poor accuracy (15-55% depending on the class) and also showed less ordinality than the imputed values from the machine learning models.

Conclusion: Machine learning algorithms provide improvements in imputation accuracy for specific groups of observations and exhibit stronger ordinality in imputed data. However, the overrepresentation of specific classes in the imputed datasets may reduce the generalizability of machine learning imputation models. While situationally suitable for variables with specific classes that hold high value, or for variables where the ordinal structure is important, future research on addressing the bias in machine learning algorithms has the potential to further improve the performance and generalizability of machine learning methods for data imputation.

Acknowledgements

I would like to thank everyone who supported me throughout the journey, my supervisor Dr. Samuel Stewart, co-supervisor Dr. Susan Kirkland, committee members Dr. Leah Cahill and Dr. George Kephart, my colleagues in the program, and my parents.

Ordinal Variable Imputation for Health Survey Data: A Comparison between Machine Learning and non-Machine Learning Methods

1 Introduction

The advent of large-scale surveys such as the Canadian Community Health Survey (CCHS) provides researchers with vast amounts of data to work with. These surveys are imperative for the understanding of health and how health is impacted by factors such as socioeconomic status, comorbidities, biological features, and environmental elements. In practice, however, advancements in data collection are not necessarily matched by advancements in data analysis. Increasing the efficiency of data analysis is an important next step to take.

The inherent appeal of large surveys is the vast number of variables collected on each participant. However, survey items usually vary in completeness of response, leading to datasets with variables that have missing observations. For example, sensitive data such as total household income or personal mental/depression status may have a lower response rate than questions about blood pressure. The optimal use of data requires researchers to deal with incomplete data and the failure to address this incompleteness may introduce bias into the analyses. Missing data not only reduce the sample size for studies, which in turn affects statistical power, but may in fact introduce bias into analyses by concealing certain populations that, if included, could change population means, variance, or observable trends in data.

Imputation is a popular method for dealing with incomplete data in which analysts replace missing observations with probable values, usually made with inferences from the available data. The process of imputation can utilize varying amounts of data depending on the imputation model. Single values or simple models (e.g. linear regression) are not the most effective or efficient in the use of available data, especially with survey datasets that have hundreds of variables (e.g. the CCHS 2014 dataset has 1129 variables in its public use micro-data file). While correctional strategies such as multiple imputation may improve certain aspects of the imputed data, the imputation model is the main determinant of the accuracy of the imputed

values (i.e. whether the estimated values in the imputed dataset truly reflect the observations that went missing).

Analyses that use imputed data assume the imputed data to be accurate, or accurate enough to provide a basis for reasonable conclusions. Imputation methods that are too simple may create imputed datasets that are lacking important associations between imputed and non-imputed data. Complex imputation models aim to incorporate multiple variables and interactions to have a higher chance of preserving features that exist in the original, unimputed dataset. However, Conventional model building techniques such as covariate selection via expert-opinion or stepwise regression may struggle to deal with large amounts of data in an efficient manner. A potential solution to this problem is to use machine learning algorithms to assist in model building.

Various types of machine learning algorithms exist, such as neural networks and decision trees, and they are powerful tools for data mining. Machine learning algorithms improve the models they produce through iterative processes of testing and/or learning. The resulting models have the potential to be complex because any information deemed useful by the algorithm is incorporated until specific parameters, pre-set by model builder, are met. The iterative processes are far more efficient than trial and error or visualizations of potential associations between a large number of variables. Also, the ability to check for all possible interactions between variables while only retaining those that are useful greatly increases the flexibility of the resulting imputation model. With the speed and volume of survey data collection increasing, and thus creating datasets with potentially more complex features, survey data imputation models can benefit greatly from the processing power of machine learning algorithms.

This thesis evaluates the performance of imputation models based on machine learning algorithms and compares them to traditional imputation models via a simulation study of the CCHS 2014 dataset. The variable of interest is total household income, an ordinal variable in the CCHS 2014 dataset.

2 Background Literature

2.1 Non-response

In the real world it is unlikely for surveys to have a response rate of 100% and further unlikely to have all respondents fully complete every single item in a survey. The term non-response describes cases in which all or some information is missing from a respondent. Two subcategories of non-response are *unit non-response* and *item non-response* (Raghunathan 2010). This study is mainly concerned with the latter.

Unit non-response is a case in which there is no information across all variables for an individual; common examples include refusals of the survey, failure to contact, and loss to follow-up. There is virtually nothing a researcher can do with these cases because any inference or imputation on a population with no information to base an inference on will lack confidence and validity. Thus the topic of unit non-response is out of the scope of this study.

Item non-response is a case in which information is missing for some variables. Unlike unit non-response the cause of item non-response, if not due to data entry errors or mistakes during data collection, is potentially open for educated inferences based on observed values and exemplary complete cases in the survey. Survey data includes responses to a wide variety of questions, some dealing with more sensitive topics than others. The reluctance of a respondent to provide a response to sensitive questions, such as household income, is likely to be influenced by the response itself. Item non-response thus introduces a threat to the integrity of survey data because the mechanism of missingness serves to suppress specific answers more often than other, more common answers. It is therefore important to try to understand the mechanism of missingness for a variable that exhibits item non-response in order to identify potential biases in the pattern of responses.

Table 1 is an example of how missing data may skew the predictions of a regression model when a variable has missing values. The example utilizes two datasets based on the 2014 wave of the CCHS to build an ordinal regression model for

"total household income" using only "highest completed education" as the explanatory variable. For this simulated example, one dataset (shown in Table 1b) has manually induced missing "total household income" values in a biased fashion; of the 5 categories for "total household income" those in the lowest and highest category are more likely to be missing than those in the three middle categories. The results in Table 1b show that the predictions from the regression model are biased towards the middle categories (as seen in Table 1c, which shows the change between Table 1a and Table 1b) due to their overrepresentation in the dataset with induced missingness (i.e. bias towards the mean). To minimize the effect of biases introduced by missing data it is first important to consider the nature of the missingness, or, in other words, why certain data are missing from the dataset in large.

Table 1a. An example table for predictive probabilities of a univariate ordinal regression model that uses Highest Completed Level of Education to predict Total Household Income for the 2014 Canadian Community Health Survey

n = 63522	Total Household Income				
Education	<20k	20k-39k	40k-59k	60k-79k	>80k
Less than SS	0.370	0.381	0.129	0.0543	0.0654
PS Graduate	0.140	0.315	0.215	0.128	0.202
Some PS	0.124	0.298	0.218	0.136	0.225
Secondary School	0.0485	0.159	0.182	0.164	0.446

Note. Higher cell values indicate a higher chance of a model predicting an observation to fall into the associated income group based on the observation's education group. (SS = secondary school, PS = postsecondary school).

Table 1b. An example table for predictive probabilities of a univariate ordinal regression model that uses Highest Completed Level of Education to predict Total Household Income for the 2014 CCHS dataset with manually induced missing data

n = 61362	Total Household Income				
Education	<20k	20k-39k	40k-59k	60k-79k	>80k
Less than SS	0.341	0.398	0.139	0.0571	0.0651
PS Graduate	0.125	0.314	0.227	0.134	0.201
Some PS	0.111	0.295	0.229	0.142	0.223
Secondary School	0.0439	0.157	0.189	0.171	0.439

Note. Higher cell values indicate a higher chance of a model predicting an observation to fall into the associated income group based on the observation's education group. This dataset includes a bias in missingness towards income groups 1 and 5, and thus the proportional distribution is skewed towards the middle groups compared to the data in Table 1a that has no missing data. (SS = secondary school, PS = postsecondary school, CCHS = Canadian Community Health Survey).

Table 1c. Change (in %) from predictive probabilities of Table 1a (whole population) to predictive probabilities of Table 1b (population with induced missingness)

	Total Household Income				
Education	<20k	20k-39k	40k-59k	60k-79k	>80k
Less than SS	- 7.9	**+ 4.3**	**+ 8.0**	**+ 5.2**	- 0.5
PS Graduate	- 10.3	- 0.5	**+ 5.3**	**+ 4.8**	- 0.7
Some PS	- 10.5	- 0.9	**+ 5.0**	**+ 4.7**	- 0.7
Secondary School	- 9.5	- 1.6	**+ 4.0**	**+ 4.5**	- 1.7

Note. Bolded values indicate increases in predictive probabilities from Table 1a (original dataset) to Table 1b (dataset with missing data). The concentration of positive values in income groups 40k-59k and 60k-79k shows that the univariate regression model based on the dataset with induced, biased, missing data is more likely, for any level of highest completed education, to predict total household income values to be in this range compared to the original dataset. The specific changes in predictive probabilities provide an example for how missing data can affect a regression model (SS = secondary school, PS = postsecondary school).

2.2 Non-response models

The most common and general terms to describe missingness are non-response models (van Buuren 2012). A model that describes the missingness as missing completely at random (MCAR) assumes errors, mistakes, and other external conditions cause missingness/non-response and thus the missing values are unassociated with the nature of the variable in question. Data that is MCAR may occur from survey respondents overlooking some questions or a data collector forgetting to record an answer. The main distinction between non-response models is whether there is complete randomness or specific randomness. In terms of information bias, MCAR data are non-differential measurement errors. On the other hand, data that is missing for a specific reason theoretically occurs by a certain factor influencing a

survey respondent to refuse to provide the answer to a question. The concept of social desirability bias is a common example; a respondent may become reluctant to provide information that may produce an unfavorable judgment upon them (e.g. income that is very low or very high). A further distinction between non-response models depends on the associations between missing and observed values. Missing at random (MAR) assumes that the missingness of a variable is associated with the distribution(s) of other variables. Missing not at random (MNAR) assumes that the missingness of a variable is associated with the value of the variable itself. Differential measurement error occurs in MNAR data as the rates of missingness differ between respondents of different classes/categories.

The amount of data MCAR in real world data is usually negligible as long as the quality of data or sample size is high enough. In some cases, different definitions of MCAR exist, as in for longitudinal data (Verbeke & Molenberghs 2000), for which it becomes identical to MAR. Most missing data tend to be either MAR or MNAR, but one cannot distinguish these two missing data types based on observable data alone (Sterne et al. 2009). This is problematic because while MAR assumptions allow other variables in the dataset to provide a basis for imputing the missing values, MNAR assumptions isolate the missing values from the rest of the observed values in the dataset. Nonetheless, the difficulty of dealing with missing survey data lies in determining the extent to which the missingness is MAR (i.e. explainable via associations with other variables) and MNAR (i.e. the product of a systematic bias that is inherent and unique to the variable in question).

2.3 Imputation

There are different methods a researcher can use to deal with missing data. The simplest method is to discard all observations with missing values and conduct analyses on only the complete cases. Complete case analysis has its flaws because it assumes that values are MCAR, which is rarely the case for real data and, as a result, tends to introduce systematic bias into the results (van Buuren 2012). Complete case analysis also reduces sample size (sometimes drastically), which may not be an allowable option regarding statistical power for some datasets. An alternative to

consider is imputation, the procedure of replacing missing values with one or more plausible values (i.e. imputed values). Imputation creates complete datasets that provide point estimators that are, for the most part, better representations of the population average than point estimators based on data with missing observations. Also, data providers can share imputed datasets to avoid inconsistency born from different missing data processing methods across data users. National Statistical Offices such as Statistics Canada most commonly use single imputation methods, in which a single dataset with imputed values to replace the missing values is created, when publishing data (Chen & Haziza 2018a). Point estimator calculations are simplified in imputed datasets by assuming that values are MAR, thus removing the need to account for the non-response bias. If this assumption is held, an imputed dataset, built on inferences from observable data, serves as a better representative of the true values than an unimputed dataset. There are several procedures to obtain imputed values, some being more robust than others, and ultimately the quality of imputed datasets (and the subsequent analyses made on the imputed data) rely on the integrity of these procedures. Imputation procedures are commonly generalized into two categories: deterministic imputation and random imputation.

Deterministic imputation procedures produce the same imputed dataset if repeated on the same population. Parametric methods of deterministic imputation apply specific models, such as linear regression, to generate imputed values but can also be as simple as replacing all missing values for a variable with its mean or median value. These methods are likely to distort the distribution of a variable that has missing values unless other variables and complete cases hold enough information to provide a robust estimation model (Chen & Haziza 2018a). Models that do not restrict the types of associations and interactions that may exist between variables are better at preserving the original distribution of a variable. Non-parametric imputation methods are more flexible because these do not specify model function and rather are 'greedy' in their use of available information. Predictive mean matching (Little 1988), nearest-neighbour imputation, and kernel smoothing (Silverman 1986) are examples of non-parametric methods. These are able to account for multiple types of interactions and associations between variables but are prone to overfit or to become

overcomplex when too many variables are added (the curse of dimensionality). Models that are overfitted become sensitive to noise and outliers in the data. This sensitivity can cause problems in the consistency of results if one wishes to implement random imputation procedures as well as deterministic procedures.

Random imputation procedures produce different imputed datasets each time. The most common source of variance in imputed values is the random drawing of donor values in procedures such as random hot-deck imputation. The variance in these imputation procedures, imputation variance, is purely artificial and is not a favourable replacement or simulation of the discrepancy between observed and unobserved values (i.e. non-response error) (Chen & Haziza 2018a). More advanced methods such as fractional imputation and balanced imputation, both originally proposed by Kalton & Kish (1984), work to reduce the imputation variance by assigning weights or constraints to imputed values and/or donor values. The reduction of imputation variance with the use of methods more sophisticated than completely random draws provides a better simulation of the unobservable non-response model. Methods that apply weights to random drawing are potentially more robust for simulating distributions that are difficult to express as functions, as required by most deterministic methods.

Many imputation procedures, however, often assume a total MAR situation for the sake of simplicity in statistical methods. This, as stated in the non-response mechanism section, is not necessarily the case for real data and violations in the MAR assumption give rise to false accuracy and inaccurate estimates. For example, Kim et al. (2012) concludes that there are demographic differences between two separate non-response categories for survey questions regarding income ("Don't Know" and "Refused [to answer]") from the General Social Survey (conducted by the National Opinion Research Center at the University of Chicago). The findings from the 2012 report and another article from 2007 (Kim et al.), this one using data from the Maternal and Infant Health Assessment in California, suggest that; those missing income data are not a random sample of a survey population but instead are a specific demographic that is, consequentially, underrepresented compared to the rest of the survey population. Underestimation of variance in the imputed dataset leads to

confidence intervals that are naively narrow for the resulting parameter estimates, which themselves may be skewed by models assuming a distribution different from the variable's actual distribution of its values. Advanced imputation procedures take this into account and aim to simulate the unknown/unobservable missingness mechanism (the MNAR nature of values) by techniques such as variance estimation procedures and multiple imputation.

Variance estimation for single imputation procedures have two main leagues of thought. The two-phase framework works with individual variance terms, originally proposed by Sarndal (1992) through the decomposition of the total error in estimator calculations, which include the sampling variance, the non-response variance, and a mixed value that represents the covariance between the sampling and non-response error. The reverse framework, proposed by Fay (1991), addresses response probabilities, decomposed into its own variance term, by assuming its contribution to total variance negligible if the imputation model is correctly specified. Compared to variance estimation procedures, however, multiple imputation is more common in both literature and practice.

2.4 Multiple Imputation

The now widely used procedure of multiple imputation was originally proposed by Rubin (1978, 1987). While single imputation procedures produce a single imputed dataset, multiple imputation procedures impute each missing value more than once and thus create multiple imputed datasets. The calculation of population estimates has two distinct steps: 1) the individual analysis of all imputed datasets to produce point estimates and 2) the pooling of point estimates to produce a multiply imputed population estimate. The two-step approach incorporates variances arising from imputation models (within-model variance) and from dataset population selection procedures (between-model variance). Nonetheless, multiple imputation requires specific conditions for its inferences to remain valid.

Murray (2018) provides a review of the conditions and justifications that are sometimes taken for granted by analysts, especially when they are hidden behind the commands and codes of statistical software. **Bayesian validity and congeniality**:

Rubin derived the functions of multiple imputation using Bayesian arguments e.g. $P(X|Y_{obs})$ denotes the probability of a missing value being X given the observed value of Y. This method assumes that one can estimate or predict the distribution of a value by drawing inferences from a previously established (i.e. prior) distribution. Bayesian arguments are not ideal for validating the results of multiple imputation in survey data because inferences regarding unobserved populations that are based on observed populations (i.e. posterior predictive distributions) are likely to be underinformed or misinformed for multi-step/multi-user analyses. For example, a prior distribution of a variable that an imputation step specified by the distributor of the survey data (in this case the distributed dataset is an imputed dataset) may not be the same as the prior distribution an end-user of the data may specify for their variance calculations. The mismatch between imputation model (made by the data distributor) and analysis model (made by the data end-user) is further explored in the concept of congeniality, introduced by Meng (1994). A recent update by Xie and Meng (2017) provides evidence that, even under conditions of uncongeniality (i.e. the mismatch of prior distributions in multi-step/multi-user analyses), inferences can still hold as long as "the imputer's model is more saturated than the analysts' [model]" (Murray 2018). In other words, the flexibility of an imputed dataset is limited by the robustness of its imputation model. **Frequentist validity:** The efficacy of multiple imputation lies in its Frequentist properties; the more imputed datasets there are, the more likely the final imputed value will properly reflect the population estimate. The process of repeated sampling increases the robustness of a sample by normalizing the sampling distribution (central limit theorem) and maximizing coverage from the observed values. Of course, the observed values in the dataset must have sufficient coverage on their own for repeated samplings to provide any benefit.

Rubin's rules for proper imputation: Rubin (1987) defines certain rules required for *proper* multiple imputation. Firstly, the predictive distribution of a parameter estimate and its sampling variance should remain unbiased after repeated sampling (up to ∞ imputed datasets, i.e. $M = \infty$). These conditions are usually met as long as the proportion of missing data is not large enough to suggest an under-representation of a specific group within the observed values. Additional adjustments

to address concerns of bias, regardless of the use of multiple imputation, include the specification of the missingness mechanism within the imputation procedures. A separate condition regards the uncertainty of the imputation procedures themselves. Random imputation procedures can produce a value for between-model variance that can itself be a significant source of bias that strays results from the actual variance of the variable these methods aim to approximate. Murray's review communicates the message that future research on imputation will benefit from improved models that "reflect uncertainty about missing values and about the imputation model" (2018).

2.5 Bias variance trade-off

Multiple imputation procedures usually use a single imputation model and run them multiple times. Valid inferences from imputation models (especially those that intend to hold across various samples and datasets) require an imputation model to accurately specify the distribution of the missing variable, or its relationship with complete variables, to reduce non-response bias (Chen & Haziza 2018b). However, imputation models also include their own sources of variance and hence an optimal model aims to minimize its prediction error, often measured with the mean squared error for continuous outcomes. The mean squared error (e) of a parameter estimate ($\hat{y}$) for a true value (y) can be decomposed into three components: Irreducible error (ε), which is the inherent variability in the data, bias (b), and variance (v).

$$\hat{y} = y + e$$

$$e = \varepsilon + b + v$$

Bias is a measure of the difference between the true conditional value of an object and the average prediction of the learned classifier across different training sets. Variance is a measure of the consistency of predicted values across different training sets, regardless of the true conditional probability. (Cambridge Online 2009) Of the three components researchers aim to limit the bias and variance of a model since the irreducible error is, as its name suggests, out of the scope of model improvement. Bias and variance, however, are not easy to reconcile because improving one component usually leads to the exacerbation of the other; the term for this condition is the *bias variance trade-off*.

The bias variance trade-off is exemplified in the comparison between linear models and non-linear models. Linear models are more likely to have high bias because the errors are consistent in respect to the predicted model (a straight line). Regarding variance, the results of a linear model are likely to have little variance across different samples because noise and outliers have less influence on the prediction model. On the other hand, non-linear methods are usually lower in bias and greater in variance. A decrease in bias can result from three conditions (Cambridge Online 2009): 1) the model being consistently correct, 2) probabilities for error being relatively equal across all predictions, or 3) multiple predictions for the same missing value resulting in errors in opposite directions, causing the magnitude of error to average out to 0. The sources of low bias (especially 2 and 3) indicate that, in general, an increase in variability is synonymous with a decrease in bias. Therefore the sensitivity to unique trends of non-linear models both benefit and harm complex models with its flexibility and tendency to overfit (Cambridge Online 2009). The concept of a trade-off between bias and variance provides an important basis for evaluating the efficacy of a model, and hence is convenient to keep in mind when searching for effective imputation models.

2.6 Moving beyond single model imputation

The most common imputation procedures, whether multiple or single, rely on a single imputation model. For example, statistical programs such as STATA ("*mi*") and R ("*mice*") use a single outcome regression model as the default setting for their multiple imputation procedures (StataCorp 2017, van Buuren & Groothuis-Oudshoorn 2011). However, single model imputation procedures are vulnerable to model misspecification because bias within the dataset can influence the imputed values if the model does not correctly emulate the missingness mechanism. Long, Hsu, & Li (2012) were one of the forerunners to deal with model misspecification by performing doubly robust non-parametric imputation. Their imputation procedures utilize both an outcome regression model and a propensity score model to increase the chances of correctly specifying the missingness model. However, others (Kang & Schafer 2007, Chen & Haziza 2018a) have criticized the use of double models as they

are likely to decrease model efficiency while they still remain liable to misspecification. Thus, the logical next-step was to increase the number of regression and propensity score models. Chen and Haziza (2018b) propose the use of multiply robust imputation as the basis for multiple imputation (multiply robust multiple imputation) because as long as any single model correctly specifies the missingness model, multiply robust imputation satisfies Rubin's rules for proper imputation.

Missingness mechanisms in survey data are likely to have complex associations with the remaining data observed in the dataset that may be difficult to capture using regression models or other commonly used imputation models. Complex models assisted by data processing and machine learning are likely to be more suitable to specify missingness models, thus ensuring the robustness of estimates and applicability for multiple imputation. The following sections briefly review the use of machine learning in health literature and examples of machine learning algorithms for imputing survey data.

2.7 Machine learning in health literature

Technological advancements have enabled the collection and communication of large datasets. Data analysis methods should also move beyond traditional modeling practices to match the increasing speed and volume of data collection. Machine learning algorithms are powerful data mining tools that incorporate the supervised learning of data. Unlike conventional data analyses methods that are unsupervised, supervised learning uses labeled data to actively inform and improve the models they produce. The effectiveness of machine learning algorithms increases with the availability of labeled data that act as exemplary cases from which the model building process can draw inferences. A review of health analytics by Islam et al. (2018) shows that data mining strategies on health data is useful for the creation of guidelines for clinical and administrative decision support. Furthermore, Morgenstern et al. (2020) suggest in their review that a greater focus should be placed on the utilisation of large datasets for machine learning in population health research. The report by Rey de Castillo (2014) was one of the first to compare various machine learning algorithms on their accuracy for imputing survey data.

Rey de Castillo (2014) compared conventional imputation models (least median squared errors, regression imputation, and predictive mean matching) with machine learning models, characterized by their supervised learning mechanisms that inform the prediction model. The author used one decision tree algorithm (M5P) and two neural networks (multilayer perceptron regressor and radial basis function) to impute artificially missing wage values from anonymous microdata files of the European Union Statistics on Income and Living Conditions. The comparison of conventional imputation and machine learning showed that predictive mean matching performed the best in terms of producing parameter estimates closest to the original values while the machine learning algorithms produced better one-to-one likeness of data. One issue the author made evident about data mining methods was that the resultant variances were underestimated with a bias towards the mean. Nonetheless this study proved valuable in the sense that it showed the direction along the bias variance trade-off in which future studies should pursue; to test more complex models that lower bias and increase variance. Additionally, the incorporation of multiple models over single model imputation should also improve the results obtained by machine learning algorithms as they do for conventional models.

Large-scale survey data such as the CCHS dataset with tens of thousands of observations contain great potential for supervised learning approaches to utilize. As well as benefitting from large amounts of labeled data, data mining methods can incorporate the nature of the collected data to customise the model building approach. Survey data includes information of several different types and categories and often have questions that are linked via follow-up questions that only apply to a selected portion of the survey population, thus resulting in large proportions of null observations. The heterogeneity of information in survey data provides the opportunity for meta-analytical procedures such as meta-path construction. Shang et al. (2016) and their proposed framework *ESim* provides an example for how heterogeneous information networks are parsed out to identify semantics within the dataset that can serve as foundations for clustering and classifying similar variables.

Lastly, ensemble methods (i.e. the use of multiple models and predictions to generate a consensus) are common in data mining practices as well.

Complex machine learning algorithms are susceptible to overfitting in the presence of outliers. Ensemble methods work to reduce the effects of overfitting in a similar approach to how multiple imputation works to improve single imputation; an algorithm learns and creates a model for each dataset subsample, commonly produced by bootstrapping the original dataset. Voting, regression, or other data type specific integration methods then calculate the mean model that, if successful, is more stable than each individual model. Also similar to multiple imputation, though, most ensemble methods depend on only a single algorithm and thus may not be able to use large amounts of data to its fullest potential.

2.8 Survey data imputation

It is important for data analysts to first consider the non-response model before imputing survey data. MCAR data in surveys are non-differential measurement errors, for which simple imputation methods will usually suffice as long as the amount of missing data is small (e.g. less than 10%). On the other hand, tables 1a-c and Kim et al. (2007, 2012) provide examples of how data with differential measurement error (MNAR mechanism) influence certain types of analyses. While the simplicity of complete case analyses (i.e. removing observations with missing values from a dataset) is tempting, there will always remain the risk of obscuring results and any further conclusions or hypotheses one may draw from those results. This section provides examples of how imputation has the potential to improve survey data analysis.

Imputation aims to reduce the risk of drawing inappropriate conclusions from survey data with missing data by providing a reliable estimation of what the missing values could be. A series of simulation studies with the 2002-2003 Los Angeles County Health Survey data (Zeng 2009) showed that multiple imputation can provide reliable statistical inference for datasets with up to 15% of data missing, performing better with demographic variables (e.g. age, education, race) rather than health outcome variables (e.g. arthritis, hypertension, diabetes). Dipnall et al. (2016) shows

evidence for the use of multiple imputation in imputing non-MCAR data for a machine learning prediction algorithm in the National Health and Nutrition Examination Survey 2009-2010 dataset. De Silva et al. (2019) and Lee et al. (2020) utilize multiple imputation to reduce bias in imputed datasets that address missing values in longitudinal data.

Surveys with large amounts of variables can reach a point where unique variables may capture similar information about the population. Variables that capture similar information decrease in value for imputation models because the scope of inference the model gains from including them may not outweigh the model variance the additional variables can introduce. The presence of confounding variables further adds to the mixture of potential model candidates but are less of a problem for imputation because the ultimate goal of imputation does not require the identity of the true causal pathway between dependent and independent variables. Since the selection process for explanatory variables is duplicated between the data provider and the data user, Murray (2018) advises that the imputation model should always be more complex (i.e. contain more variables) than the analysis model to ensure congeniality between the two models. This provides another avenue by which machine learning can benefit imputation, as machine learning algorithms are likely to create complex models to incorporate any useful information a dataset can provide.

Congeniality is an essential ingredient for increasing the research value of imputed survey data. The incorporation of multiple variables increases the capability of imputation models to emulate the structural features of data. Krenzke & Judkins (2008) tested hot-deck imputation (using age, race, and sex) against semi-parametric imputation (stepwise regression followed by a clustering algorithm) on the National Education Longitudinal Survey (NELS) to find that the more complex model resulted in a closer one-to-one likeness to the original data. The authors concluded that the complex semi-parametric approach was able to better preserve structural features that are lost in simple models and therefore retain more explanatory power in the variables they impute. The difficulty, however, arises when the time comes for the imputation model to decide when to stop adding complexity, especially when provided with a vast number of variables from large survey data.

To summarise, survey datasets provide large amounts of data for analyses, but analyses must include robust methods to address missing data to avoid any loss of statistical power and to mitigate any bias caused by incomplete coverage of the sample population. While many data imputation methods exist to replace missing data, machine learning algorithms can improve the efficiency and accuracy of data imputation by automatically generating models that can fit to complex associations that may exist between the large amount of variables in a survey dataset.

3 Objectives

This thesis evaluates imputation model performance on an ordinal variable from a large-scale survey dataset. The primary hypothesis question is: are machine learning algorithms better than traditional algorithms for the imputation of missing values in an ordinal variable (income)? Secondary research questions are: do different patterns of missingness have an impact on the performance of the imputation models and, if so, how does the relative performance of the machine learning algorithms compare with the traditional algorithms in each pattern of missingness?

This thesis provides a comparison between 7 imputation models to determine whether imputation models produced by machine learning algorithms perform better (i.e. provide more accurate imputations) than imputation models produced by non-machine learning algorithms, particularly in complex patterns of missingness. The thesis concludes with a suggestion, based on the results of a simulation study, for which imputation models are more suitable for imputing large-scale survey data.

4 Methods

This thesis project is a cross-sectional simulation study of data from the 2014 wave of the CCHS that evaluates the performances of several imputation algorithms. The simulation study consists of multiple steps including data preprocessing to provide a foundation for the machine learning algorithms. This thesis includes a variable selection step for practical reasons (to ensure the machine learning algorithms can converge in a timely manner), nonetheless the selection is guided by survey imputation literature (especially those pertaining to income imputation) and other simulation studies.

4.1 Dataset

CCHS 2014

The Canadian Community Health Survey (CCHS) 2014 provides national-scale data from its target population of all 110 health regions of Canada, 107 from the Provinces and 1 from each Territory (Statistics Canada, 2016). The annual sample population consists of approximately 65,000 respondents selected via a multi-stage sample allocation strategy (moving from Province/Territory to health region, each sample proportional to their respective populations). A total of 63522 observations are available to the public via public use micro-data files (PUMFs). Data distributors, in this case Statistics Canada, release PUMFs that are modified from the original data to ensure no identifiable information about individuals or organizations are included. This type of microdata is accessible by faculty and students of post-secondary institutions via the Data Liberation Initiative (Statistics Canada 2020) and does not require approval from a research ethics board.

Household Income

Household income is a common variable collected in surveys but is nonetheless a sensitive topic that usually has a relatively large proportion of missing values. For this reason household income is a common target for imputation. The baseline assumption for imputation is that other variables in the dataset provide enough information for a reasonable imputation. However, single model imputation

procedures with variance correction techniques (such as variance estimation and multiple imputation) are not ideal approaches for imputing household income. The sensitive nature of the variable suggests that the missingness mechanism is likely to include MNAR characteristics. Variables with partial MNAR characteristics are difficult to model because the observed data may not capture the distribution of the variable well enough to provide a basis for valid inferences. Single model imputations do not perform well when the model cannot sufficiently emulate the missingness mechanism. The establishment of a powerful prediction or classification algorithm has the potential to increase the validity of analyses that use imputed values to accommodate survey variables that are prone to missing data/non-response.

The CCHS 2014 PUMF provides total household income as an imputed variable based on nearest neighbour donor imputation (Yeung & Thomas 2012) and thus has very few missing values. A simulation study requires an environment in which the analyst can easily control the environment (in this case the dataset) and thus the total household income variable from the CCHS 2014 PUMF is an appropriate candidate for the simulation study. The imputation models should nonetheless work for any dataset because the algorithms that create the models are not survey specific but rather incorporate aspects of all types of data into their models (i.e. the models created by the algorithms are specific to a dataset, but the algorithms themselves are not). In theory, an algorithm that creates a successful imputation model for total household income in one survey should be able to create an imputation model for total household income in any survey, providing the survey dataset has sufficient information to create an imputation model.

4.2 Process

Data preprocessing

The CCHS provides a comprehensive data dictionary and the data have a small proportion of missing values overall. One point of interest may be the differentiation between values that are not available (i.e. missing) and those that are not applicable (i.e. null observations due to questions that follow-up on specific answers). The majority of items in the CCHS annual component, in fact, are follow-up questions. One

can identify follow-up questions using the data dictionary but are also characteristically unique from stand alone questions due to their high proportions of not applicable/invalid responses. The simulation study will not include follow-up questions as they are not independent variables. Also, questions with high proportions of invalid responses will, in general, not be as informative than those with little or no invalid responses. The variable selection process intends to reduce the number of variables to conserve time on model building while preserving the advantages of machine learning (i.e. the capability to process large amounts of data). For practical reasons the study uses 30 independent variables, which is a reduction from the total of 1129 variables available in the CCHS 2014 PUMF. The variables are focused around demographics, geography, socio-demographics, health care usage and access, general health, mental health, and education. While mainly based on imputation methods of other national surveys (see section on Inducing Missingness below), the selection also includes items in the CCHS that studies have identified to have associations with income (e.g. oral health (Farmer et al. 2017), and access to health care services (Lasser et al. 2006).

The data preprocessing step also includes the transformation of data structures to methods specific to the study. For example, this simulation study transforms the CCHS dataset using dummy variables to replace categorical data with binary variables that represent each response category because the process of inducing missing data, and some machine learning algorithms, requires a completely numerical dataset.

Simulation Study

This thesis follows the design of a simulation study performed by Scheffer (2002) in which 3 types of missing data (MCAR, MAR, and MNAR) are present for imputation. The R function "*ampute*" in the "*mice*" package (van Buuren & Groothuis-Oudshoorn 2011) can induce all 3 types of missing data, as well as mixtures of different types of missing data with specific proportions (Schouten et al. 2017). The simulation study utilizes data from the CCHS and induce missingness in the "total household income" variable in order to evaluate the performance of imputation methods. The completeness of the CCHS data allows for a simulation study to induce missingness in the variable with almost complete control over the missingness mechanism. The ideal

imputation algorithm performs well (i.e. shows little bias and preserves the original variance structure of the variable) in all types of missingness. Therefore the ability to measure the performances of imputation algorithms in various missingness mechanisms assists in the selection of the best, or most robust, algorithm.

The process of generating missing data can range in complexity. In simple cases where the mechanism of missingness is not of particular interest, one may simply set the desired proportion of missingness then use a random number generator to choose which observations become missing (this would be an example of MCAR). However, for this simulation study we are concerned about the effects of missing data type on imputation models and therefore must be able to simulate different missingness mechanisms. The results from Scheffer (2002) indicate that the trends in estimation error are mostly linear and are observable even at small proportions. The variety in simulated datasets provides different situations for which an imputation model may perform well or poorly. For example, based on the findings of Scheffer (2002) and Krenzke & Judkins (2008,) we expect multiple imputation via ordinal regression to perform poorly with datasets that have complex missingness mechanisms (i.e. increased likeliness of imputation model misspecification) and/or large proportions of missing data.

Inducing missingness

The R package "*mice*" contains a function "*ampute"* that is able to generate MAR, MCAR, and MNAR patterns. The "*ampute*" function is able to manipulate the proportion of missingness, the missingness mechanism, and the proportion of a missingness mechanism in a dataset with multiple missing data types ("*ampute*" is capable of adjusting more than this, such as simulating survey design, but that is beyond the scope of this thesis) (Schouten et al 2017). Once a user identifies the variable (denoted as A_m for this explanation) in which the function induces missingness the user can then distribute the *weight* each variable in the dataset will have on the calculations of missingness probabilities for observations of A_m. If the user gives *weight* to variables other than A_m (e.g. B, C, D ...), the missingness mechanism will be MAR because a prediction model can draw inferences on the missing values by looking at associations between A_m and the other variables. If the

user gives *weight* to A_m itself the missingness mechanism will be MNAR because the values that inform the pattern of missingness are themselves, missing from the data, and cannot inform prediction models. Furthermore the user can introduce a mix of both MAR and MNAR mechanisms by distributing partial *weight* to A_m and the rest to *B*, *C*, *D*, and so on. An MCAR mechanism will induce missingness in A_m by random and will therefore not utilize any *weights* assigned to variables in the dataset.

The simulation study produces 4 sets of missing data, each at 3 proportions of missingness: 5%, 15%, and 35%, for a total of 12 simulated datasets: one dataset with a MNAR missingness mechanism (control) and 3 datasets with varying complexities of MAR missingness mechanisms. In the case of this dataset, the *ampute* function bases the weighting scheme for selecting values to become missing solely on the total household income variable. To emulate patterns suggested in literature (explained below), this simulation study utilizes a two-tailed distribution pattern as the principle missingness model. Thus, the missingness for the MNAR dataset is more likely to occur in both higher and lower ends of observations in the total household income variable. In comparison, the MAR datasets will have the *ampute* function base their weighting schemes on the selected independent variables rather than the total household income variable. One similarity between the MAR and MNAR datasets are that they all share the general two-tailed distribution model for missing values. In theory, the general missingness model (two-tailed) is stronger in the MNAR dataset where it and the total household income variable are the only sources of variability in missingness. Therefore, the MAR datasets will not show as strong as a two-tailed missingness pattern in the total household income variable. Appendix item 1 (R code) includes the full details for the *ampute* functions for the missing data induction steps. The MAR datasets, shown in Table 2, will be the main points of comparison between the imputation models and will consist of a simple mechanism (5 independent variables and no interaction terms), a moderately complex mechanism (8 variables + 1 interaction term), and a highly complex mechanism (15 variables + 3 interaction terms). The higher complexity models will provide situations in which machine learning methods are likely to perform better than classical methods due to the

presence of interaction terms, which are difficult for models to effectively incorporate unless an efficient variable selection mechanism is utilized.

Regarding the proportions of missingness, simulations studies for imputation often utilize high proportions of missingness (≃30% - Berg et al. 2016; 1 - 50% Scheffer 2002; 50% - Quartagno et al. 2019; 28 - 33% Ogundimu & Collins 2019) to emphasize the effects of the imputation models. This thesis uses the proportions 5%, 15%, and 35% to better reflect actual missingness proportions of income variables in survey data, which typically ranges from 5% to 35% (≃6% National Health and Nutrition Examination Survey 2015-2016; ≃30% Jamaica National Healthy Lifestyle Survey 2000-2001; ≃35% CCHS 2014).

Studies of large-scale surveys around the world (Germany - Frick & Grabka 2014) (UK - Fisher et al. 2019) (Jamaica - Wilks et al. 2007) (US - Ogundmu & Collins 2019) have identified several factors that are associated with non-response for questions regarding individual and household income. The most common factors identified are: age, sex/gender, education level, race/ethnicity, employment status, and regionality (health care region/geographic). Other interesting factors of note are apparent U-shaped patterns of income non-response (higher rates of non-response at both the low and high ends) across income itself (GSOEP - Frick & Grabka 2014) and across age (JNHLS - Wilks et al. 2007). While classical imputation models rarely incorporate interaction terms along with independent, unique variables some studies speculate their impact, especially in studies that identify non-linear distributions of non-response. The findings of the aforementioned studies guided the selection of factors the "inducing missingness" step uses for the MAR datasets along with guidance for the non-machine learning imputation algorithm (ordinal logistic regression and predictive mean matching).

Finally, the simulation study proceeds to an imputation step that creates imputed datasets by all proposed algorithms (section 4.3) and then evaluates the performance using an evaluation method specifically designed to assess ordinal variables (section 4.4). This thesis explores different machine learning algorithms such as decision trees, support vector machines, and neural networks to find in what ways machine learning may improve imputation models. In this thesis the machine learning

algorithms do not utilise multiple imputation, which the non-machine learning algorithms include. While future studies can investigate the effectiveness of machine learning to provide models for single imputation on which multiple imputation methods can build on, it is important to find out in what ways machine learning models work differently, or similarly, from current models to determine whether or not the models can benefit from or are compatible with multiple imputation.

4.3 Algorithms

Ordinal logistic regression (*mice polr*)

The R "*mice*" package (van Buuren & Groothuis-Oudshoorn 2011) serves as comparator imputation algorithms as they are widely available for use by survey data analysts. The R function "*mice*" performs imputation for ordinal variables based on logistic regression.

The *polr* method in the mice library applies a proportional odds logistic regression, which utilizes successive logistic regressions to address the ordinal nature of the target variable (van Buuren 2020). The proportional odds model creates a hierarchical odds structure across the classes of the target variable. The change in odds in crossing each class-border is the same (i.e. proportional) so the model may not be adequate for data that does not follow a proportional odds structure (Ford 2015).

Regression models, unlike machine learning models, require a set of pre-selected variables to avoid certain factors from overinfluencing the prediction model and biasing the results. Since different variables may represent similar or overlapping information (e.g. smoking status and lung cancer rate) it is important for regression models to, as much as possible, receive input from variables that provide unique information. Data mining methods such as principal component analysis can assist in the selection of input variables. More traditional methods for choosing input variables include stepwise regression and reference to prior studies.

For the purposes of the simulation study I select variables for the regression model in a way that would, as much as possible, emulate a model created from traditional variable selection methods. This means that the regression model will

perfectly match the missingness mechanism of the simple-MAR dataset, closely match the missingness mechanism of the moderately-complex-MAR dataset, and partially match the missingness mechanism of the highly-complex-MAR dataset (Table 2).

The concept of survey weights is unignorable when analyzing survey data. Survey weights are necessary for the calculation of parameter estimates that represent the target population, rather than representing the sample population. Although the calculation of parameter estimates is not necessary for imputation of sample data, the inclusion of survey weights, or its factors, in the imputation model is needed to ensure congeniality between imputation and analysis models. While the CCHS 2014 PUMF includes survey weights as a variable in the PUMF, the ordinal logistic regression imputation model does not include the survey weights as an explanatory variable because including both the survey weights along with its factors (geography variables in the case of CCHS 2014) would be redundant. Similarly, the machine learning algorithms (expanded on below) may experience difficulties with including survey weights over the geography variables in their models as variable selection is automatic. Additional modifications to the imputation model, such as including the survey weights as a separate level that influences predictive probabilities (Quartagno et al. 2019), exist in principle but introduce great complexity to models (especially for non-binary imputation). As the field of covariance adjustment via survey weights in machine learning is a relatively complex concept, it is out of the scope of this thesis.

Predictive Mean Matching (*mice pmm*)

The predictive mean matching method (PMM) uses the same library, *mice*, as the ordinal logistic regression and therefore shares the multiple imputation aspect. PMM builds off of regression imputation by using models to determine distance measures and donor values (Morris, White, & Royston 2015). Distance represents the similarity between observations, for example how many same values they share across same variables, and observations within a specific distance of the recipient observation are matched to form donor pools. Initially introduced by Little (1988), the closest observation serves as the sole donor value.

The method in the *mice* library (*method = "pmm"*) generates multiple sets that each draw a random sample from a donor pool to provide an imputed value, repeated

for each observation that requires imputation for the variable specified in the model (van Buuren 2011).

Ordinal decision trees (*rpart* and *rpartScore*)

Decision trees work by passing all non-missing observations in a dataset through a series of splits, eventually providing a step-by-step tree that divides the observations into their respective class(es). Algorithms determine splits by comparing the purity of the resulting divided datasets, with purer datasets containing less heterogeneity of classes (relative to other potential splits). Decision trees have the potential to create visual interpretations of the data that assist in the identification of patterns or even serve as practical applications as well (e.g. quick screening tools). The mechanism by which decision trees organize data (i.e. make splits to divide the data into distinct groups) can vary and some work better with certain types of data. Rey de Castillo (2014) found a bias towards the mean when imputing survey data via classification and regression tree (CART) imputation. Therefore this simulation applies a decision tree algorithm that addresses both the ordinality and class-wise proportionality of the imputed variable. The *rpartScore* library (Galimberti, Soffritti, & Di Maso 2012) provides adjustments to *rpart* by utilising an ordinal impurity function based on the generalized Gini impurity function. The addition of linear costs, a loss function that scales misclassification by increasing distance between the correct and incorrect class (e.g. estimating a class 1 observation as class 2 is less costly than estimating it as class 5), in determining splits also helps to address the ordinality of the income variable the simulation study targets. The simulation study also includes a variation of the *rpart* function that utilizes an ordinal cost matrix to guide the generation of decision trees (section 4.4, below, explains the concept of ordinal cost in detail).

Ordinal Random Forest (*ordinalForest*)

Random forests classify observations by making multiple CARTs, usually introducing variance via sample selection strategies such as boosting or bagging, and choosing a final classification based on the collective results of the trees (multiple trees, hence the name random forest). The *ordinalForest* function utilizes bagging (bootstrap aggregating) which is a sampling method that takes random samples with

replacement from a pre-specified training set. The random sampling produces sample datasets with a variety of class representation. This provides the opportunity for variable classes with lower proportional representation to have more influence on the model because some sample datasets can have, by random selection, a higher proportion of a specific class compared to the original variable. Random forest algorithms then amalgamate the trees generated on the bagged samples to create a final tree that incorporates the best performing aspects from the individual trees. The unique aspect of the *ordinalForest* function is that it considers a continuous variable within the ordinal variable to develop regression models (Hornung 2020). The ordinality of the outcome variable is preserved by setting a range of continuous values to represent each class.

Support vector machines (*e1071*)

Support vector machines (SVM) are algorithms that utilize techniques such as dimensional transformations to simulate models for a target outcome variable. The flexibility offered by dimensional transformations in SVM provide the potential to model non-linear classification problems but otherwise hampers the interpretability of the resulting models. Developed in 1995 by Cortes and Vapnik, SVM were initially applicable to binary classifications. Meyer (2019) provides an R library, e1071, that applies the concept of SVM to multiclass classification by incorporating a voting mechanism. Simply put, the *svm()* function applies binary classification via SVM for each category of the target variable. In the case of our simulation study, the total household income variable generates 10 subclassifiers that provide the basis for voting.

Neural Networks (*nnet*)

Neural networks are similar to SVM as they are very useful for non-linear classification problems but do not provide interpretable models. In other words, the classification models from neural networks and SVM are black boxes that utilize transformations and weightings to represent the relationship between the explanatory variable(s) and the outcome variable. The model itself, therefore, does not provide any information for understanding relationships between variables and is instead a tool for estimating values. The *nnet* function creates classification models

through a feed-forward neural network (Venables & Ripley 2002). Neural networks are a series of nodes that apply weights to inputs they receive from sources of data, ending with a final output(s) that provides the classification of an observation. A neural network can have any number of nodes as well as multiple layers of nodes to increase complexity. Complexity of networks also depend on how the algorithm adjusts its weights and on how the network uses correct classifications to guide model creation (supervised learning). Although, similar to the *svm* function in e1071, the *nnet* function is not specifically adjusted for ordinal variables the function is able to classify multiclass variables by changing the number of output nodes to match the target variable. The impact of ordinality in model creation is an important topic of this thesis and is discussed in section 6.

4.4 Evaluation

A great advantage of a simulation study is the ability to measure a model's performance. Since the original, non-missing data is available, one can assess the quality of an imputation by how many times it was correct (accuracy) and how many times it was incorrect (instances of misclassification). Misclassification can be a ratio or one can assign a cost for incorrect classifications. Performance measures can incorporate both accuracy and a misclassification cost to evaluate the imputation algorithms.

The most common performance measure of evaluating the performance of imputation methods is the mean squared error (section 2.5). However, the definition of classification error does not consider the ordinal nature of variables such as self-rated general health and total household income. For example, one can say a model that imputes "good health" for an observation that was originally "excellent health" performs better than a model that imputes "poor health" for the same observation. The nuance of ordinal classification performance has been a topic of interest since 1970 when Murphy proposed the ranked probability score for evaluating probability forecasts for weather. The ranked probability score functions by assigning weights according to distance calculated in a symmetric matrix of predicted and observed values.

An ideal classifier would have 100% accuracy and 0% misclassification. Perfect accuracy (or inaccuracy) is highly unlikely and in the case of ordinal variables the varying degree of inaccuracy becomes a unique source of information about the classifier. Therefore, this thesis uses a performance measure that incorporates both accuracy and a misclassification cost to evaluate the imputation algorithms.

When assigning a cost to misclassification, one must consider the nature of the variable being imputed. A categorical variable may not have an equal distribution of observations across classes. This can be adjusted with inverse probabilities of misclassification (Figure 1. Appendix). An ordinal variable, as mentioned previously, has classes that are closer together than others. This can be adjusted with a linear absolute value loss function (Figure 2, Appendix). An ordinal variable with class imbalance, therefore, would incorporate both these elements when determining the cost of an incorrect classification. The product of these two concepts creates something called a cost matrix (Figure 3, Appendix), which varies in size by how many classes are in the variable of interest.

To use both accuracy and cost, George, Lu, & Chang (2016) proposed a statistic that is a measure of the Euclidean distance between the performance of a given classification model and the ideal classification model (100% accuracy, 0% misclassification or 0 cost). For this statistic, named the d-statistic, a smaller value indicates better performance by the classifier. Nyongesa (2016) provides a comparison of evaluation metrics for ordinal variable imputation in which the d-statistic appears to capture more information on the imputed data than metrics such as Kendall's Tau-b. Therefore this thesis evaluates the performance of all 7 imputation algorithms via the d-statistic.

With regards to a reference point for income imputation accuracy in the CCHS dataset, a report published by Statistics Canada (2013) provides a class transition table from different imputation methods utilized to create the PUMF. The class transition table for total household income in the CCHS compares the values predicted by the imputation models against the true, original values in the 5 classes of income. Therefore, this thesis also evaluates the class transition tables for each imputation model as well as their d-statistics.

5 Results

5.1 Datasets and Variables

The original dataset, which I based the simulation study on, is the public use microdata file of the annual component of CCHS 2014 that contains around 63500 observations. I processed this data to reduce the number variables and to only include complete cases. Then I induced missingness into the income variable with the "*ampute*" function in R, thus creating 4 different patterns of missing datasets (MAR-1, MAR-2, MAR-3, and MNAR), each with 3 proportions of missingness (5%, 15%, 35%). For each of the 12 resulting datasets I imputed income values to obtain an imputed dataset. I repeated this process for the 7 imputation algorithms I intend to compare (2 non-machine learning and 5 machine learning) for a total of 84 imputed datasets.

The final list of variables in each dataset (Appendix List 1) includes 30 independent variables along with total household income for a total of 31 variables. The Appendix item also includes a table (Table A1) that provides a summary of each variable distributed across the total household income variable. From the demographic variables: age, sex, marital status, household composition, and education show that those in groups of higher income are more likely to be younger, male, married, and postsecondary school graduates. From the self-reported health variables, those in groups of higher income are more likely to report "Very Good" for general health and "A bit stressful" for life stress. Visible patterns within across the income groups become less pronounced in the rest of the variables, but lower income groups tend to report more health issues (e.g. arthritis, back problems, and high blood pressure). Another notable difference between the income groups appears in the dentist/orthodontist utilization variable, with a higher proportion of utilization (i.e. having ever visited) by the higher income groups (from 40k-59k and above).

The processed CCHS datasets do not include respondents under the age of 18 to increase the consistency of response rate across the included survey items and to avoid surrogate respondents. The resulting processed dataset has 90.1% of the

responses as complete, in other words a decrease of 9.1% to a total of 53,065 observations as this thesis only uses complete cases.

For the MAR datasets, I used an increasing number of variables to induce different patterns of missingness in income. Table 2 shows the variables used to induce missingness for each MAR dataset along with the variables each non-machine learning imputation model incorporates as explanatory variables. The simulation study includes an intentional disparity between the two sets of variables, mainly in the complex-MAR dataset, to provide an observation about whether missingness mechanism misspecification (section 2.6) impacts the accuracy of the imputation models.

Table 2. Variables in the missingness mechanisms for each MAR dataset compared to example to the variables in the regression models for each MAR dataset

Dataset	Mechanism	Model
Simple-MAR	Total household income, Age, Sex, Education, Province, Self-perceived health	Total household income, Age, Sex, Education, Province, Self-perceived health
Moderate-MAR	Total household income, Age, Sex, Education, Province, Self-perceived health, Diabetes, Restriction of activity, Cultural/racial origin **Interaction: Age x Cultural/racial origin**	Total household income, Age, Sex, Education, Province, Self-perceived health, Diabetes, Restriction of activity, Cultural/racial origin
Complex-MAR	Total household income, Age, Sex, Marital status, Education, Province, Self-perceived health, Perceived life stress, **High blood pressure**, Diabetes, Cancer, **Dentist/orthodontist**, **Unmet healthcare needs**, Alcohol, Restriction of activity, Cultural/racial origin	Total household income, Age, Sex, Marital status, Education, Province, Self-perceived health, Perceived life stress, Diabetes, Cancer, Alcohol, Restriction of activity, Cultural/racial origin

	Interactions: Age x Cultural/racial origin, Marital status x Education, High blood pressure x Province

Note. **Bold** terms indicate variables that are in the missingness mechanism but not in the regression model. (MAR = Missing at Random)

MAR datasets 2 and 3 also have interaction terms to increase the complexity of the missingness mechanism. The missingness in the MNAR dataset is solely based on income.

5.2 Total Household Income

The CCHS income variable in the 2014 PUMF is total household income. As evident in Figure 4 the distribution of observations across the classes is not equal. The lowest class (total household income under 20,000 for the year of 2014) has the fewest observations and the highest class (total household income over 80,000 for the year of 2014) has the most observations. The large quantity of observations in the >80k class may indicate that the total household income variable has lost some information due to setting the highest income value too low as it is unlikely that all observations in the >80k class represent a similar demographic. While the numerical nature of income practically necessitates a class that has a soft boundary (i.e. a class that includes all values equal to and above a certain threshold value), it would be ideal to set a threshold so that the highest income class includes a similar number of observations with the other classes. Aside from demographic representation, the large number of observations in the >80k class had a large influence on the models created by the machine learning algorithms. Since machine learning algorithms are "greedy" in the sense that the algorithms reward accurate classifications over punishing misclassifications, the final models tend to be biased towards classes with many observations because doing so would be correct more often even if just by

chance. It is likely that this tendency is exacerbated when the classes with fewer observations are difficult to model.

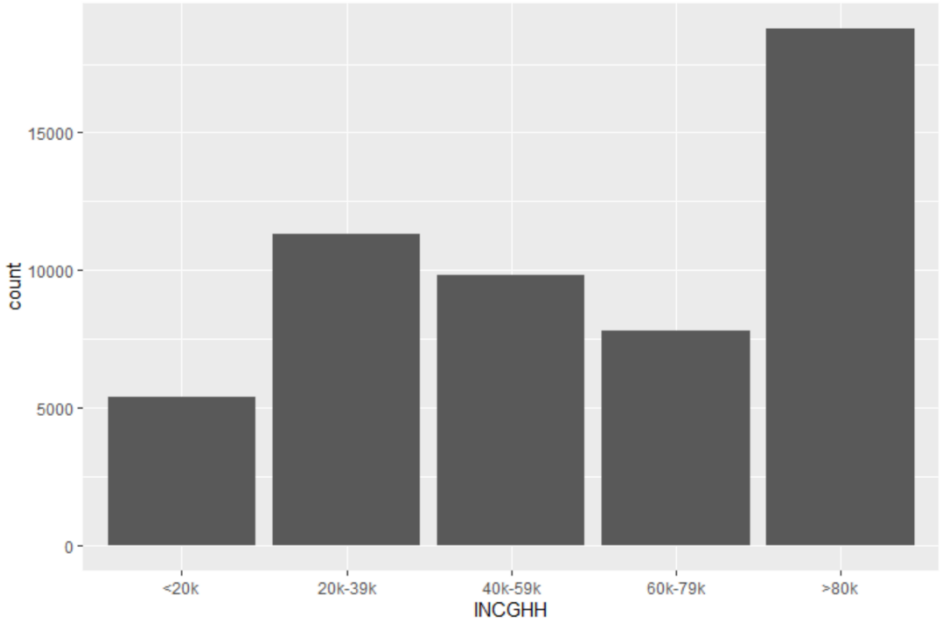


Figure 4. Number of observations for total household income (INCGHH) of the processed Canadian Community Health Survey 2014 dataset by income class.

Classification models can vary in their performance across classes within a target variable. Class-specific accuracy is dependent on how unique the observations within the class are. The target variable must have enough variability across the explanatory variables for each class to have a unique fingerprint that classification models can identify. The Pearson correlation coefficient measures the linear correlation between the target variable and a specific explanatory variable. Some prominent correlations exist between: age (higher income classes are younger), marital status and living situation (higher income classes are married), and education (higher income classes have more post-secondary school graduates). Correlations between binary variables exist in: sex (highest income class has more males than females while all other income classes are predominantly female), arthritis, back problems, and high blood pressure (income classes 40k - 59k and below have higher proportions of these health issues), and with dentist/orthodontist utilization (income classes 40k - 59k and above have a higher proportion of utilization).

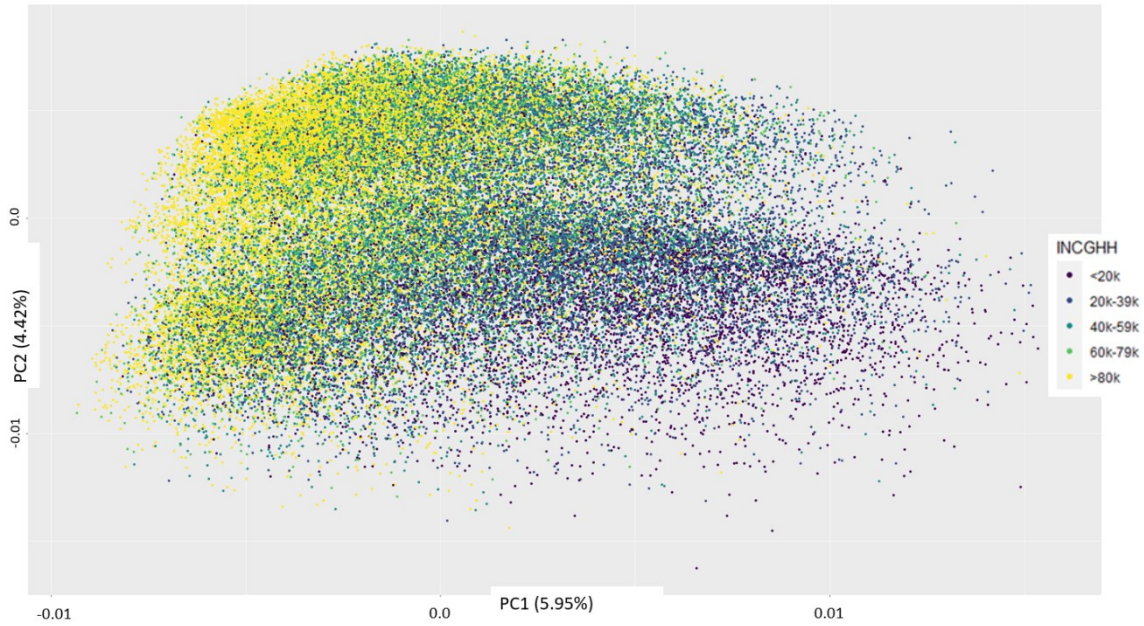


Figure 5. Principal component analysis of total household income (INCGHH). The graph represents the variability of observations within the total household income variable across the two most explanatory principal components out of all 30 covariates.

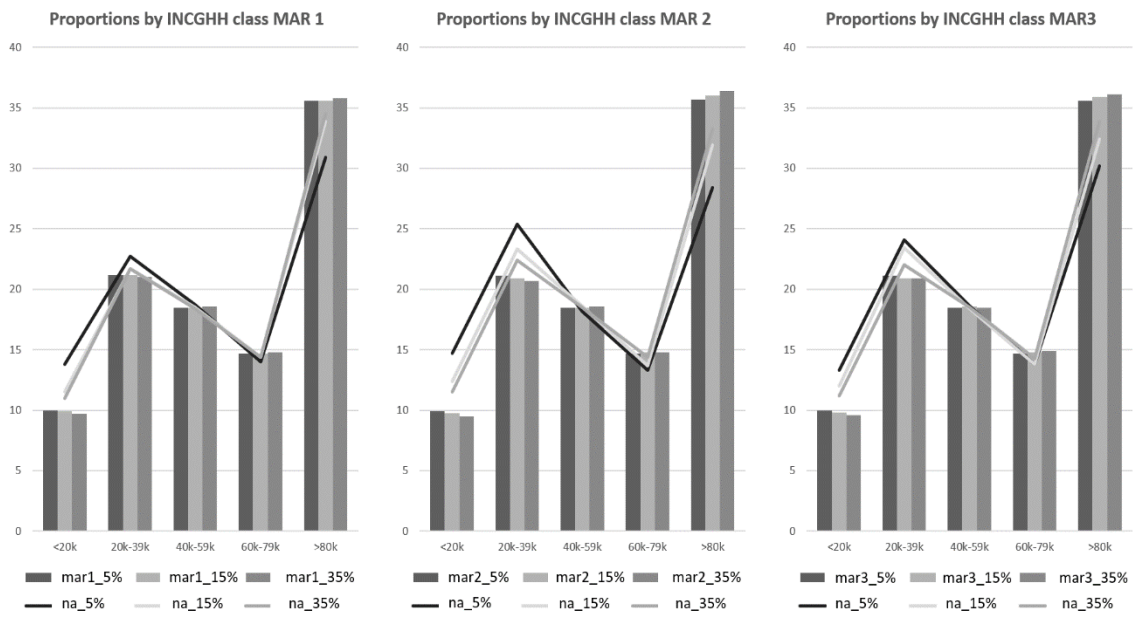


Figure 6. Distribution of observations across the 5 total household income classes in the 3 Missing at Random (MAR) datasets. The bars for each class represent the proportion of the observed values they make up out of all observed values. The lines for each class represent the proportion of the missing values they make up out of all missing values. The shades of grey represent the proportions of missing data with darkest shade showing the results for datasets with 5% missing, the lightest shade for 15% missing, and the middle shade for 35% missing.

Another method for investigating the modelling potential of a variable is principal component analysis. Figure 5 represents the two most influential principal components (explanatory variables transformed in order to maximize the amount of variability of the target they represent) on income out of all 30 explanatory variables. The axis titles indicate that, while being the most influential, each component only accounts for less than 6% of the total variance for observations within the total household income variable. The pattern in colours show the distribution across income classes (yellow representing the >80k class and a darkening of hue represents the subsequent, lower income classes). The figure also shows that total household income is somewhat differentiable by a visible pattern but no easy, clear cuts exist between all classes. The difficulty for clearly distinguishing income classes using the 30 explanatory variables poses a problem to the imputation models because this

means that there are observations that may not be consistently classified. High variability in classification leads to inconsistent results and increased computation for attaining model convergence (i.e. when the algorithm agrees on a specific model out of the many iterations it generates).

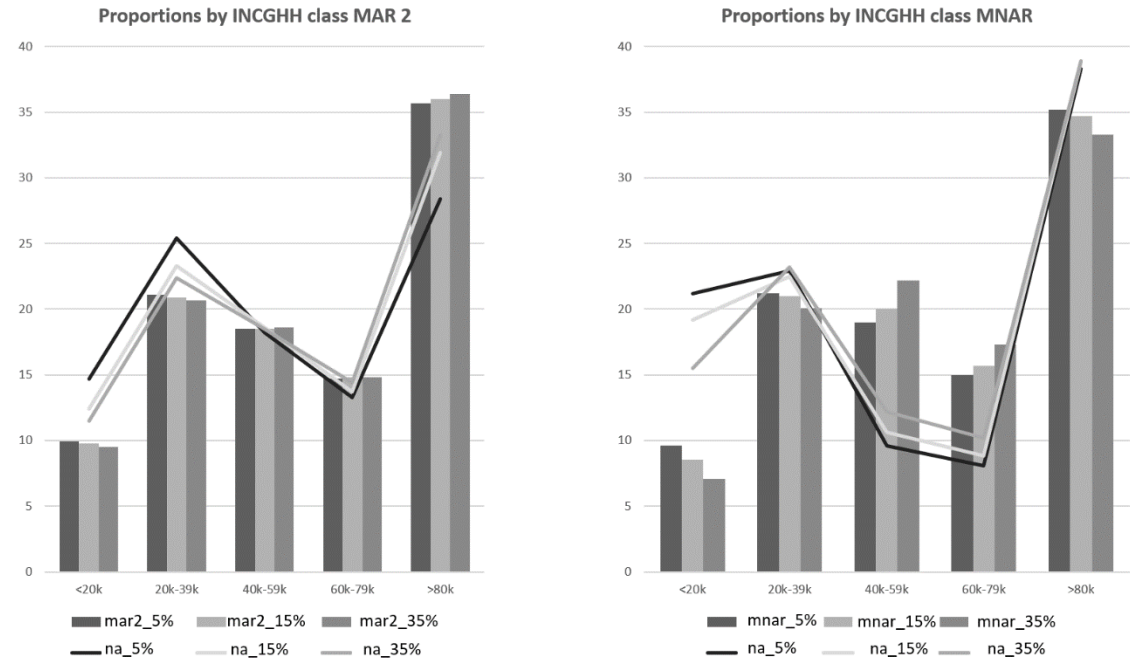


Figure 7. Distributions of observations across the 5 total household income classes in the Missing Not at Random (MNAR) dataset (right) compared to the Missing at Random (MAR) datasets (left), with the MAR2 dataset as a reference. The bars for each class represent the proportion of the observed values they make up out of all observed values. The lines for each class represent the proportion of the missing values they make up out of all missing values. The shades of grey represent the proportions of missing data with darkest shade showing the results for datasets with 5% missing, the lightest shade for 15% missing, and the middle shade for 35% missing.

Figure 6 shows the distribution of observations across the income classes for each missing dataset. The bars represent the proportion of observed values and the lines represent the proportion of missing values for each income class. There is a slight decrease in the proportions of the first two income classes, and a slight increase in the later three classes as the proportion of overall missingness increases.

The MNAR dataset (Figure 7), however, shows a significant change in the distribution across income classes as the proportion of overall missingness increases;

the first, second, and fifth classes decrease in proportion while the third and fourth increase. The patterns in proportional distribution of observations can influence the imputation model in many ways. Multiple imputation methods especially depend on unbiased prior distributions to ensure assumptions are met for their variance correction methods (Section 2.4).

5.3 Algorithm Performance

The machine learning algorithms generally produced imputation models that have better d-statistics than the imputation models produced by the non-machine learning methods, with exceptions in the MNAR data. Figure 8 and Table 3 show the d-statistics for each algorithm by missingness proportion. The machine learning algorithms have smaller d-statistics than the non-machine learning algorithms. Also, there seems to be little change in d-statistic across missingness proportion for all algorithms, which indicates all models performed consistently across proportions of 5%, 15%, and 35% missingness. While imputation on higher proportions of missingness may start to show changes in performance, it is unlikely for practical reasons that one would impute data with such high proportions of missingness (section 4.2). The MNAR datasets, however, show a pattern of increasing d-statistic (i.e. worse performance) with increasing missingness proportion compared to the slight downwards trends visible in some algorithms in the MAR datasets. Along with information provided by the class transition tables, it is likely the unique increase of d-statistics in MNAR data is due to the change in proportional representation of income classes in the MNAR data. As shown in Figure 7, the change in observed total household income values by missingness proportion has a bias towards keeping observations in the middle classes and to induce more missingness in the end classes. This bias appears to be reflected in the decreasing performance of classification models because MNAR data with larger proportions of missingness end up more biased, which leaves the algorithms with a less accurate representation of the original data to base their models on.

Table 3. d-statistics of imputation models for total household income in Canadian Community Health Survey 2014 by missingness mechanism and missingness proportion.

	Missingness mechanism by missingness proportion											
	MAR1			MAR2			MAR3			MNAR		
Algorithm	5	15	35	5	15	35	5	15	35	5	15	35
Rpart	0.57	0.58	0.59	0.55	0.55	0.55	0.58	0.58	0.60	0.71	0.67	0.69
RpartSc	0.55	0.54	0.54	0.55	0.54	0.54	0.55	0.54	0.54	0.47	0.49	0.53
SVM	0.53	0.51	0.51	0.52	0.52	0.52	0.53	0.51	0.52	0.48	0.50	0.51
NNET	0.52	0.50	0.51	0.50	0.50	0.51	0.52	0.50	0.50	0.44	0.46	0.48
O-Forest	0.57	0.51	0.51	0.51	0.51	0.51	0.52	0.51	0.51	0.45	0.46	0.48
O-Reg	0.76	0.76	0.76	0.77	0.77	0.75	0.76	0.76	0.76	0.75	0.75	0.76
PMM	0.70	0.70	0.70	0.71	0.70	0.70	0.67	0.66	0.66	0.64	0.65	0.66

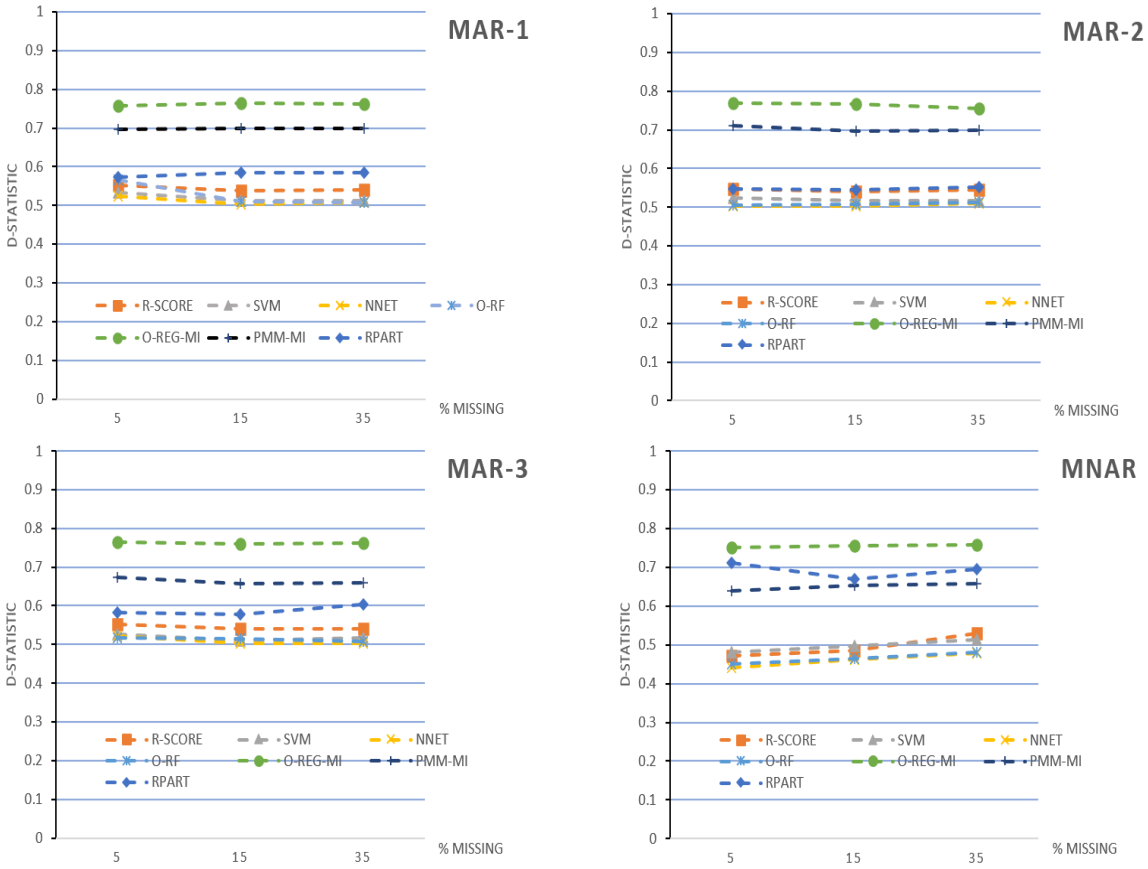


Figure 8. d-statistic of imputation models for total household income in Canadian Community Health Survey 2014. Figures 9-12 in the Appendix show each graph individually. Table 3 displays the numbered values. Legend: R-SCORE (*RpartScore*), SVM (*e1071* support vector machine), NNET (*nnet*), O-RF (*OrdinalForest*), O-REG-MI (*mice* regression based multiple imputation), PMM-MI (*mice* predictive mean matching based multiple imputation), *RPART* (*rpart*).

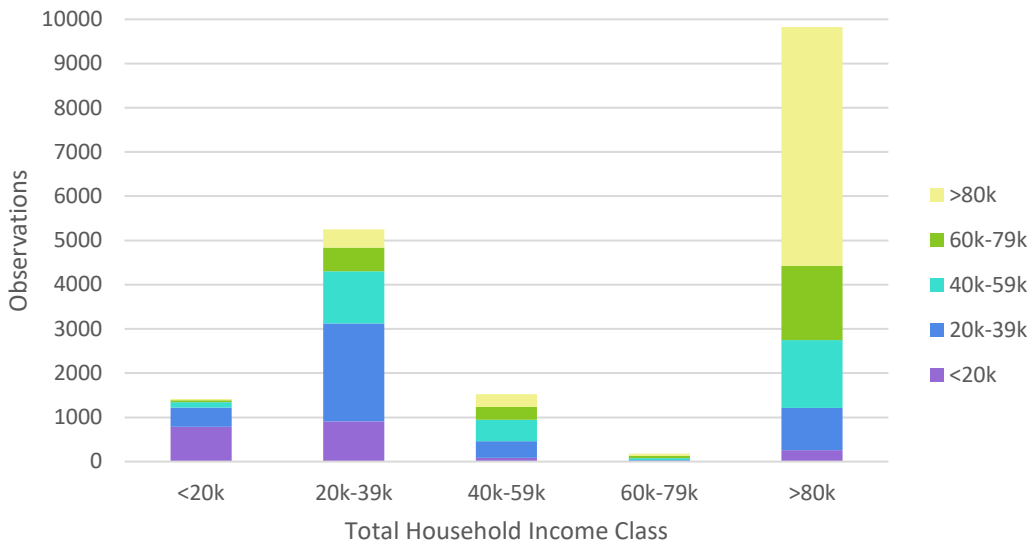


Figure 13. Breakdown of imputed total household income values from the imputation model generated by the *ordinalForest* algorithm on the Missing at Random 3 (MAR3) dataset with 35% missingness. The columns represent the number of imputed observations in each income class and the colours represent the original classes of the observations.

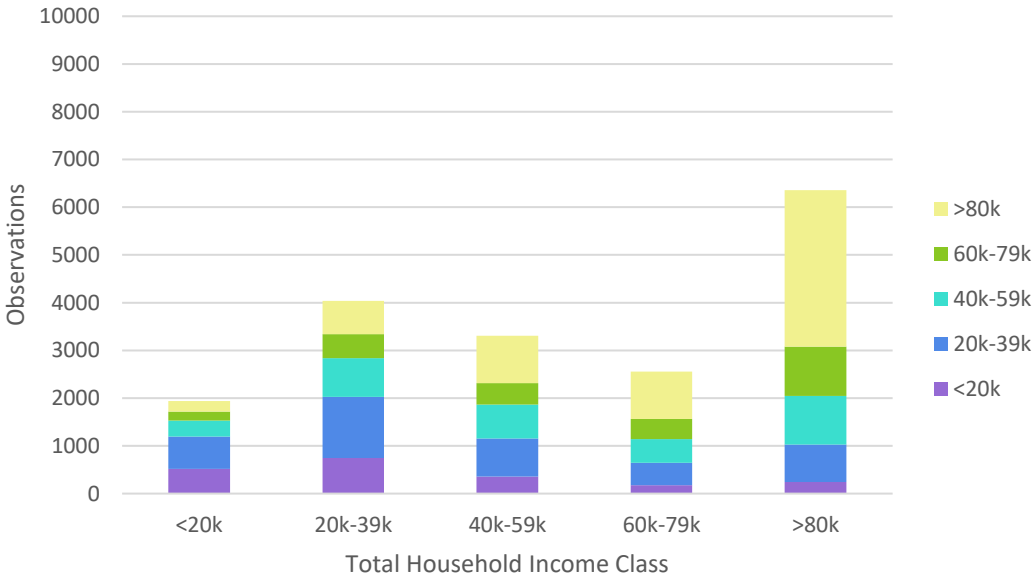


Figure 14. Breakdown of imputed total household income values from the imputation model generated by the *mice pmm* algorithm on the Missing at Random 3 (MAR3) dataset with 35% missingness. The columns represent the number of imputed observations in each income class and the colours represent the original classes of the observations.

Class-by-class comparisons between true values and imputed values provide details not captured within the d-statistic. Figures 13 (machine learning) and 14 (non-machine learning) provide an overview of how two of the best performing, in their respective categories, imputation models imputed values from the MAR3 dataset with 35% missingness. The missing observations in the MAR3 dataset follow a similar pattern to the distribution of observations across the original total household income variable (Figure 4) and thus an ideal imputation would reflect this pattern in the imputed values. At first glance, the shape of the bar graph in Figure 14 closely resembles that of Figure 4. However, the colours in each bar indicate that the *mice pmm* imputation model does not necessarily place the imputed observations into their original (i.e. correct) classes. In comparison, the *ordinalForest* imputation (Figure 13) does a better job, although not perfect, of placing the imputed values into their original classes. On the other hand, the shape of the bar graph in Figure 13 is not as similar to Figure 4 than is Figure 14, especially notable is the underrepresentation of the middle classes with almost no imputed observations placed in the 60k - 79k income class. Instead, machine learning algorithms tend to overpopulate classes that are better defined in the data. For example, Figure 5 provides a visual representation of how the income variable is plotted across 2 primary components (transformed explanatory variables). The >80k class (yellow) and the 20k-39k (dark blue) occupy distinct areas of the plot while the 60k -79k class (green) and 40k -59k class (turquoise) are more spread out across areas that are also occupied by other classes. Therefore, it is likely that *ordinalForest* algorithm decided to overpopulate specific income classes to increase the imputation model's accuracy rather than to replicate the proportional distribution of observations across the classes in the target variable. Finally, the figures also display the ordinality of the imputed values through the colours in each bar. Compared to the *mice pmm* imputation model, the *ordinalForest* imputation model has bars that are largely represented by colours of the corresponding income class and the next closest income class(es). This pattern is most evident in a comparison between the <20k income class from each figure; in Figure 13 (*ordinalForest*) the bar is mostly purple (<20k) and dark blue (20k - 39k)

with almost none of the other colours while in Figure 14 (*mice pmm*) the bar is mostly dark blue (20k - 39k), then purple (<20k), but also includes more of the three other income classes than the bar in Figure 13. In conclusion, the preference towards select classes for increased overall accuracy and a stronger tendency of ordinality in the imputed values of the machine learning models (represented by the *ordinalForest* model) are the likely factors for the stronger performance (Figure 8) over the non-machine learning models (represented by the *mice pmm* model).

Tables 4 through 15 (Appendix) are the detailed class transition tables of the *ordinalForest* models and the *mice pmm* models. The columns contain percentages that represent the proportion of true values in each imputed class, which are the rows. Figure 15 is a class transition table from a 2013 Statistics Canada report that provides a reference point for the accuracy values.

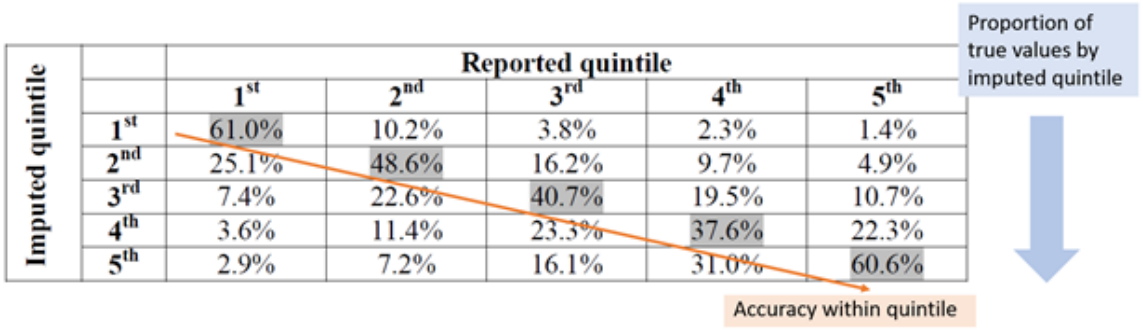


Figure 15. Quintile (class) transition table from Statistics Canada Income imputation for the Canadian Community Health Survey by Chi Wai Yeung & Steven Thomas, Household Survey Methods Division (April 2013).

The official imputation method utilized by Statistics Canada on the CCHS 2014 data consists of multiple steps and multiple models (Yeung & Thomas 2013). The initial steps group together both observed and missing values to form imputation classes and then a regression model finds the closest observed value within the groups to serve as the imputed value. The steps for forming these imputation classes varies with the amount of available data external to the CCHS survey data. The steps

most comparable to the methods in this thesis are those that do not incorporate information external to the CCHS survey data.

Most machine learning imputation models in this thesis outperform the Statistics Canada imputation model in the specific cases of the accuracies in the 20k -39k income class and the >80k income class (quintiles 2 and 5 in the example, respectively). The example model outperforms the machine learning methods in quintiles 1 (>20k), 3 (40k - 59k), and 4 (60k - 79k) and outperforms the non-machine learning models in all quintiles/classes. The *rpart* with cost matrices algorithm created imputation models that have the most similar results (supplementary material) to the example model, which suggests similarities between the model creation processes (for example, the formation of imputation classes and the formation of cost matrices). Nonetheless, the relative simplicity of implementation for the model creating algorithms in this thesis (i.e. the use of published statistical packages) provides some advantage over the multi-step process utilized by the official Statistics Canada imputation method.

The MNAR datasets (Tables 10 through 15) show different patterns from the MAR datasets and especially in higher proportions of missingness. The non-machine learning algorithms no longer preserve the class structure (such as in Figure 14) and become biased towards the middle classes. A similar bias is evident in the machine learning methods, but to a smaller degree. The following subsections describe each algorithm's performance in detail (tables attached to supplementary document).

5.3.1 CART (rpartScore)

The CART created by the *rpartScore* function performed well in the two income classes with the largest proportions of observations, the 80k (largest) and 20k-39k (second largest) annual total household income groups. The model correctly imputed observations in the >80k group consistently with more than 80% accuracy. However, there is no ordinality in the incorrect imputations. The proportion of observations that have a true value of 20k-39k is the second highest after accurate imputations rather than observations from the 60k-79k group, which is closer in income value to the 80k group. The *rpartScore* model did not perform as well in the 20k-39k group,

mostly scoring an accuracy of 50% and sometimes dropping to 40%. Similar problems with ordinality exist as the most common true value among the incorrect imputations was the >80k group instead of <20k and 40k-59k.

The model did not perform well in the other three income classes. The majority of imputed observations were incorrect for the <20k, 40k-59k, and 60k-79k groups. In fact, the model did not impute any observations to be in the 60k-79k group. The model imputed the majority of observations to be instead in the 20k-39k group (for the <20k group) and the >80k group (for the 40k-59k and 60k-79k groups) These patterns show that the rpartScore algorithm (specific settings listed in supplemental code) created a model that preferred to gain a higher accuracy score by overrepresenting imputations in the >80k income group and, to a lesser degree, the 20k-39k income group. This overrepresentation resulted in high accuracy for the classes that, together, contained the vast majority of total observations for the income variable. Unfortunately, the accuracy in the two clases came at the cost of poor performance in all other income classes along with a distorted variable class structure in the imputed income variable. This distortion is more prominent in the MNAR dataset as the imputed income variable not only loses all observations in the 60k-79k group but also in the <20k group.

5.3.2 Support Vector Machines (*e1071*)

The imputation models from the support vector machine algorithm *e1071* gave similar results to the *rpartScore* CART imputation models; a heavy focus on imputing observations into the 20k-39k and >80k income groups with severely reduced observations in the remaining three classes. The SVM model performed better than the *rpartScore* CART model within the 20k-39k and >80k income groups, scoring consistently above 60% and 80% accuracy respectively (except in the dataset with MNAR and 35% missingness proportion). Also, while to a lesser degree, the SVM model also performed better than the *rpartScore* model in the 40k-59k group, consistently scoring around 15% accuracy. However, the SVM model performed poorly in the <20k income group (typically less than 20% accuracy compared to 30-40% accuracy in the *rpartScore* model). Another difference between the *rpartScore*

CART model is that the SVM model had some, while very few, observations in the 60k-79k group. This is likely due to the flexibility of SVM algorithms that implement transformations to the input data. The total household income variable in the CCHS 2014 dataset has some classes with observations that are difficult to differentiate from observations in other classes. This pattern is observable in the plot of the two major principal components (Section 5.2, Figure 5) in which some classes show prominent clusters while other classes are less distinguishable. Data transformations allow SVM algorithms to classify observations that are difficult to separate from other classes within the target variable and provides an advantage over algorithms that are limited to analyzing non-transformed data. Nonetheless, the accuracy for the 60k-79k group was very low, ranging from 0.3% to 1% with the model incorrectly imputing over 60% of observations with a true value of 60k-79k as observations with >80k income.

5.3.3 Neural Networks (nnet)

While similar to the models from the *rpartScore* algorithm, the best performing imputation models (by virtue of the d-statistic) were those created by the *nnet* algorithm. The lower d-statistics for the *nnet* imputation models are likely due to the increased accuracy in the 40k-59k income group, comparable to the *e1071* imputation models, and the 20k-39k income group, while not as accurate as the *e1071* models. A common trouble shared with the *rpartScore* algorithm is the almost total lack of observations imputed as the 60k-79k income group.

5.3.4 Random Forest (*ordinalForest*)

The models produced by the *ordinalForest* algorithm were comparable to the *nnet* models in performance (d-statistic). The relative improvements in accuracy from the other machine learning algorithms are likely due to the class partitioning method and optimization steps that consider the ordinal nature of the target variable by the *ordinalForest* algorithm (Section 4.4). Since the algorithm considers a continuous range of values to then divide into bounded classes, all classes are existent in the imputed dataset. There are multiple internal optimization steps that *ordinalForest* can use when specified by the data analyst. In this simulation study I specified the

default method that evaluates the performance of the generated CARTs via a ranked probability score. The ordinal-conscious evaluation along with the added robustness from multiple trees with varying ordinal class widths are likely factors of the improved performance compared to the other algorithms.

5.3.5 Ordinal Logistic Regression (*mice polr*)

At first glance the non-machine learning methods seem to provide promising imputation results because the number of imputed observations matches the number of true observations in each income class. However, the class transition tables show that the accuracy within each class is poor. A distinct pattern is visible in the observations imputed by the multiple imputation model using ordinal logistic regression. All imputed observations consist of true observations from each income class, the proportional distribution of true classes is also consistent across all imputed classes: each imputed class consists of ~10% of observations from the <20k income class, ~20% from the 20k-39k income class, 15~20% from the 40k-59k income class, 10~15% from the 60k-79k income class, and 35~40% from the >80k income class. This pattern is consistent with the proportional distribution of the true observations across income classes. In other words, this pattern directly translates to the accuracy of the imputed dataset as all imputed classes follow this pattern (~10% accuracy, ~20% accuracy, 15~20% accuracy, 10~15% accuracy, and 35~40% accuracy). Therefore, unfortunately, the multiple imputation model using ordinal logistic regression is consistently inaccurate with the highest accuracy (around 40%) occurring in the >80k income class.

5.3.6 Predictive Mean Matching (*mice pmm*)

The multiple imputation model using predictive mean matching is an improvement from the ordinal logistic regression imputation. The proportional distributions of the imputed observations closely match the proportional distributions of the true observations for the MAR datasets. One improvement in imputation accuracy is the apparent recognition of a bimodal distribution with peaks at 20k-39k and >80k, similar to the overrepresentation of income groups 20k-39k and >80k in the machine learning methods but to a lesser degree. A trend of

increasing accuracy as the MAR mechanism becomes increasingly complex exists in the PMM imputation models. This is likely due to the increased number of variables I incorporated into the PMM models to match the increasing complexity of MAR mechanisms (Section 5.1, Table 2). Nonetheless, the accuracies of imputations within each income class are moderate at best, scoring over 50% only in the >80k income class with the next best being scores of around 30% in the 20k-39k income class. As shown in the poorer d-statistics compared to the machine learning imputation models, the lack of imputation accuracy devalues the correct proportionality of income classes in the imputed datasets.

5.3.7 CART (*rpart* with cost matrices)

The CARTs created by the *rpart* function with cost matrices provided results unlike any other imputation models. There is a unique focus on the middle class (the 40k-59k income group) rather than on the 20k-39k or the >80k income groups that other imputation models picked up. The likely reason for this focus on the middle class is the low ordinal cost from either end of the scale. The model creation algorithm evaluates incorrect imputations as less troublesome if they are closer to the true value on an ordinal scale. Therefore the middle class has the lowest likelihood of costly mistakes as it can only be off by 2 classes at most. The *rpart* algorithm for the MAR-2 dataset also includes bootstrap aggregation (bagging) as a modification as an attempt to improve results but the results are similar to the models the algorithm created without bagging. The *rpart* + cost matrix CART models have the highest accuracy in the 40k-59k as well as the 60k-79k income classes due to its unique bias towards the middle class. The <20k and 40k-59k income classes range from 30% to 40% accuracy, 60k-79k mostly around 30% accuracy, and 20k-39k and >80k have a range of 40% to 50% accuracy. The *rpart* + cost matrix CART models also show the most ordinality in their results with the majority of incorrect imputations being observations with true values one class away from the imputed class. However, the performance measured by the d-statistic indicates that the relative decrease in accuracy compared to imputation models from the *nnet* and *ordinalForest* algorithms was more costly than the improvements made in the preservation of ordinality.

5.3.8 Performance in MNAR data

Almost all imputation models performed better with the MNAR dataset. The increase in performance is the product of increased accuracy in the 40k-59k and 60k-79k income classes. The MNAR datasets are skewed to have relatively more observations missing from the <20k and >80k income classes, resulting in larger proportional representation by the 40k-59k and 60k-79k income classes (the proportion of observations in the 20k-39k remained relatively consistent across MAR and MNAR datasets). However, as missingness proportions increase from 5% to 35% the comparative loss in accuracy for the other income classes begin to outweigh the improvements made in the 40k-59k and 60k-79k income classes. This pattern resulted in the trend of worsening performance as missingness proportion increased, which is different to the slightly increasing performance trend shown by the MAR models.

The *mice polr, mice pmm*, and *rpart* + cost matrices algorithms were affected by the modified class imbalance introduced by the MNAR mechanism, resulting in higher proportions of the 40k-59k and 60k-79k income groups compared to models made by the same algorithms in MAR datasets. The machine learning algorithms (excluding the *rpart* + cost matrices) were affected to a lesser degree.

5.4 Variable importance

Section 2.8 describes the potential application of machine learning algorithms in the variable selection process for prediction models. One way this is possible is through the examination of post-modelling statistics that some algorithms provide. The CART algorithms (*rpart*, *rpartScore*, and *ordinalForest*) generate a list that ranks each explanatory variable by their importance as a by-product during the tree creation process. In other words, variables that are common and high-ranking throughout the lists are determined by the algorithms to be useful when predicting the target variable. Throughout all 36 variable importance lists (3 CART algorithms, 4 missing data mechanisms, 3 missing data proportions) there were 6 variables that consistently ranked in top 5: DHHGLVG (living arrangement/household composition), DHHGMS (marital status), EDUDR04 (highest completed education), DHHGAGE (age),

CHP_14 (consulted with dentist or orthodontist), and ALCDTTM (type of drinker, alcohol). Apart from the top 5, the main difference between the lists appeared as *ordinalForest* models preferring demographic variables such as respondent sex and geographic location compared to *rpart* and *rpartScore* preferring health variables such as diagnoses of high blood pressure and arthritis. While the household composition, marital status, age, and education variables are commonly associated with income (section 4.2), the presence of the dentist/orthodontist consultation and alcohol variables are interesting in their consistent usage by the algorithms. The high ranking of the dentist and orthodontist consultation variable is in line with the findings of authors such as Farmer et al. (2017) and may serve as an example of the potential utility of variables that are not as common in income prediction models. As in this way, the identification of useful explanatory variables through variable importance lists is an advantage that CART machine learning algorithms have over other machine learning algorithms (or non-machine learning algorithms that may require additional tests to measure variable importance).

6 Discussion

Machine learning algorithms produce imputation models for ordinal variables that are more accurate than imputation models produced by ordinal logistic regression or predictive mean matching. However, imputed data from machine learning methods tend to be biased towards classes that are easy to identify; for example, classes that represent a large proportion of the target variable's observations, classes with a population that is uniquely characterized by the explanatory variables, or a combination of both. In the case of total household income in the CCHS 2014 PUMF, while all 5 machine learning algorithms outperformed ordinal logistic regression and predictive mean matching, the income classes of 20k-39k and >80k were overrepresented in the imputed datasets. On the other hand, the non-machine learning methods were better at preserving the variable class structure by producing imputed datasets with observations distributed across the income classes in a similar pattern to the original dataset with no missing data. The non-machine learning methods were less accurate because the imputation models were less likely to classify the missing observations to their correct income classes. With regards to incorrect classifications, the machine learning models were more likely to misclassify observations as their adjacent income classes than as income classes further away from their original class. A stronger sense of ordinality in data may benefit certain cases of prediction that can find value in close, but incorrect guesses. The trade-off between accuracy and variable class structure is also a factor that requires consideration when implementing machine learning imputation models.

6.1 Strengths and Implications for using imputed values

The goal of imputation is to increase the number of usable observations in a dataset by estimating values for observations that have item nonresponse. One should consider the impacts of using imputed data to perform analyses such as whether the imputed observations are biased towards certain values, potentially skewing the associations between the variables within the dataset. The simulation study in this thesis provides evidence for the potential bias in both machine learning

and non-machine learning imputation methods. Both methods are affected by MNAR mechanisms because prediction models built on existing data are not able to account for inherent biases in the datasets without information on missingness mechanisms (section 2.6). Therefore data imputation is only practical if one can assume the missing data to be mostly MAR or unless one can correctly model the missingness mechanism with supplementary data/metadata, such as in the case of Yeung & Thomas (2013) creating imputation classes with data from tax forms external to the CCHS. Nonetheless, imputation models that are biased towards specific outcomes have potential use in situations where the value of correctly imputing one class greatly outweighs misclassification of less important classes. Machine learning algorithms also tend to have more ordinality in their imputations as misclassified observations are more likely to fall into classes that are closer to their original value. Therefore, machine learning imputation has the potential to provide better results for imputing ordinal variables with misclassification costs that scale with distance. For example, Nyongesa (2016) provides evidence for the potential applicability of imputation models from machine learning algorithms in the imputation of colon cancer data. The study uses real world data from the colon cancer dataset (GSE17536) and compares the performances of imputation models while considering the ordinality of misclassified observations. The cost of misclassifying a cancer patient as stage I instead of stage IV can be high due to the invasive nature of testing and treatment required for stage IV cancer patients. Potentially more costly would be the misclassification of a stage IV patient as stage I because a different approach for treatment can have dire consequences. The high performance of machine learning methods for colon cancer data imputation shows that imputation and data prediction from machine learning methods can serve a unique niche in health data analyses, in which situations similar to the prediction of cancer stages are abundant.

6.2 Limitations

The distribution of observations across the income classes in the total household income variable likely had an impact on the performances of the imputation models. As explored in section 5.2, the concentration of observations in the highest income

class (>80k) is problematic because there is a loss of information for a large proportion of the respondents that possibly have a total household income much greater than 80,000. Additional classes beyond a >80k class can provide room for observations to spread out more to possibly show trends in data in higher income populations. Another approach could be to increase the range of income within each class, allowing the variable to capture a wider scope of household incomes while possibly amalgamating similar income households into identifiable groups. An income variable that has a more balanced distribution of observation across classes may decrease the bias in the machine learning models as no single class will provide better accuracy through chance alone.

The missingness induction process outlined in section 4.2 also influences the structure of the income variable. Modifications to features of the *ampute* function can change the missingness mechanism from MAR to MCAR or MNAR (or include a complex mixture of mechanisms via subsection sampling) but can also change the way the MAR variables are associated to the missingness mechanism. Further modifications to the missingness mechanism can provide information on how different imputation models perform in specific situations but is out of the scope of this thesis. Also, the range of missingness proportions in this thesis (5%, 15%, and 35%) may limit the patterns of model performance to a smaller threshold than other studies that use simulations. This was a practical choice to limit the amount of analysis to an area that would more likely reflect data in which one would utilize imputation methods.

The non-machine learning models included variables chosen to reflect the variables in the missingness induction processes. However, the models did not include one specific variable (DHHGLVG, household composition) that the CART algorithms identified as an important explanatory variable. A posteriori analyses found that the non-machine learning models experienced some improvement when including this variable (data not shown) the final results and performances were similar to the models that did not include the household composition variable.

Finally, the dataset for the simulation study was a modified version of the CCHS 2014 dataset provided as a PUMF. Modifications to the data include changes to ensure

the anonymity of participants, which may cause some discrepancies between the actual data and the PUMF data. Due to this fact, a direct comparison of performance between the Statistics Canada imputation methods (Figure 15) and the imputation methods in this thesis may have some discrepancies. For example, in addition to the modification of data in the PUMF, the authors of the Statistics Canada report display their results (Figure 15) as "income quintiles" (i.e. observations divided into five equal groups with the groups defined by ranges of income) and not as "income classes" as defined by the total household income variable in the CCHS 2014 PUMF. Nonetheless, the quintile transition table (Figure 15) is still useful as a point of reference for income imputation and, even if the classes do not exactly match up with the quintiles, the general sense of ordinality exists in both income quintiles and income classes.

6.3 Future research

A definite characteristic of the machine learning imputation models is that they prefer values that are more likely to be correct and are prone to overrepresent such values over other classes in the target variable (i.e. greedy). The "greedy" nature of machine learning algorithms is likely because most machine learning processes focus on binary variables as target outcomes. Machine learning algorithms adapted to multi-class outcomes usually are multiple binary models that use voting systems (such as SVMs in the *e1071* package) to deal with multiple outcomes. Future developments for multi-class machine learning algorithms specifically focused on ordinal variables may benefit by incorporating ordinal-sensitive performance measures such as the d-statistic in their model creation processes.

The overrepresentation of specific classes in imputed data from machine learning methods are likely to influence analyses that use the imputed data. Studies on the impacts of imputed variables with different types of errors and bias can provide information for guidelines on the use of imputed data. Greater understanding of the characteristics of imputation methods can improve the utility of imputed data if one can identify a suitable imputation method for their specific dataset. Different datasets can provide new information about the algorithms by showing whether

improvements in accuracy are due to overfitting or not. Good performance across various datasets is evidence for an algorithm's versatility and generalizability by showing that it can adjust to different trends and associations that characterise different datasets. Future studies can validate machine learning algorithms for imputation by applying algorithms similar to the algorithms in this thesis to other survey datasets and, eventually, to non-survey datasets with ordinal data.

Different types of datasets exhibit different types of missingness. Working with real data (not solely simulated data) will provide opportunities to test the applicability of machine learning methods in situations of complex missingness, such as mixes of MAR and MNAR data. While for this thesis the datasets remained separated between MAR and MNAR data, it is more likely for datasets to contain both MAR and MNAR data. For some cases, single models (whether machine learning or not) may not suffice to properly address complex missingness. Multiple model approaches can include the specification of a missingness mechanism (Long, Hsiu, & Li 2012), the use of voting systems (Chen & Haziza 2018b), or multi-step imputation based on external data (Yeung & Thomas 2013). The strength of machine learning models lies in the possibility to provide robust single imputation methods that these multiple model approaches can incorporate. Similarly, multiple imputation methods based on machine learning models also have the potential to show the strength of incorporating a robust single imputation method to build off of.

One significant challenge for future studies to overcome is the presence of missing data within the explanatory variables and not just the target variable. The machine learning algorithms in this thesis all require the inputted data to not have any missing observations. The algorithms can only utilize complete cases because classification is based on comparing similarities and differences between observations across all input variables. Therefore, if there is a value that cannot be compared to any of the other values (i.e. a missing value, unless the variable specifically includes a field for null or missing values), algorithms will not know how to classify the variable that contains this incomparable value. Methods for dealing with incomplete explanatory variables were beyond the scope of this thesis but are important next steps for increasing the applicability of machine learning imputation models.

7 Conclusion

In conclusion, it is difficult to determine whether imputation models from machine learning algorithms completely outperform methods such as ordinal logistic regression or predictive mean matching. Nonetheless, improvements in accuracy for a specific group of observations and stronger ordinality in imputed data provide machine learning methods a unique upper-hand in situations where a specific type of observation holds very high value. Future research on addressing the bias in machine learning algorithms has the potential to further improve the performance and generalizability of machine learning methods for data imputation.

References

Berg, E., Kim, J., & Skinner, C. (2016). Imputation Under Informative Sampling. *Journal of Survey Statistics and Methodology, 4*(4), 436-462. doi:10.1093/jssam/smw032

Chen, S., & Haziza, D. (2018a). Multiply robust nonparametric multiple imputation for the treatment of missing data. *Statistics Sinica,* doi:10.5705/ss.202017.0126

Chen, S., & Haziza, D. (2018b). Recent developments in dealing with item non-response in surveys: A critical review. *International Statistical Review, 87*(S1), S19-S218. doi:10.1111/insr.12305

De Silva, A.P., Moreno-Betancur, M., De Livera, A.M., Lee, K.J., & Simpson, J.A. (2019). Multiple imputation methods for handling missing values in a longitudinal categorical variable with restrictions on transitions over time: A simulation study. *BMC Medical Research Methodology, 19*(14). https://doi.org/10.1186/s12874-018-0653-0

Dipnall, J.F., Pasco, J.A., Berk, M.B., Williams, L.J., Dodd, S., Jacka, F.N., & Meyer, D. (2016). Fusing data mining, machine learning and traditional statistics to detect biomarkers associate with depression. *PLoS ONE, 11*(2), e0148195. doi:10.1371/journal.pone.0148195

Farmer, J., Phillips, R. C., Singhal, S., & Quiñonez, C. (2017). Inequalities in oral health: Understanding the contributions of education and income. *Canadian Journal of Public Health, 108*(3). doi:10.17269/cjph.108.5929

Fay, R. E. (1991). A design-based perspective on missing data variance. Paper presented at the 429-440.

Fisher, P., Fumagalli, L., Buck, N., & Avram, S. (2019) Understanding Society and its income data. *Understanding Society Working Paper* 2019-08, Colchester: University of Essex

Frick, J.R. & Grabka, M.M. (2014). Missing income data in the German SOEP: Incidence, imputation and its impact on the income distribution. *SOEP Survey Papers,* 225. Deutsches Institut für Wirtschaftsforschung (DIW), Berlin

George, N.I., Lu, T., & Chang, C. (2016). Cost-sensitive performance metric for comparing multiple ordinal classifiers. *Journal of Artificial Intelligence Research*, 5(*1*), 135-143. doi:10.5430/air.v5n1p135.

Galimberti G., Soffritti G., Di Maso, M. (2012). Classification trees for ordinal responses in R: The rpartScore package. *Journal of Statistical Software*, 47(10), 1-25. URL: http://www.jstatsoft.org/v47/i10/.

Hornung, R. (2020). Ordinal forests. *Journal of Classification*, 37, 4-17. https://doi.org/10.1007/s00357-018-9302-x

Islam, M. S., Hasan, M. M., Wang, X., Germack, H. D., & Noor-E-Alam, M. (2018). A systematic review on healthcare analytics: Application and theoretical perspective of data mining. *Healthcare (Basel, Switzerland), 6*(2), 54. doi:10.3390/healthcare6020054

Kalton, G. & Kish, K. (1984) Some efficient random imputation methods. *Communications in Statistics - Theory and Methods, 13*(16), 1919-1939. doi: 10.1080/03610928408828805

Kang, J. D. & Schafer, J. L. (2007). Rejoinder: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science, 22*(4), 574-580. doi:10.1214/07-sts227rej

Kim, S., Egerter, S., Cubbin, K., Takahashi, E. R., & Braveman, P. (2007). Potential implications of missing income data in population-based surveys: An example from a postpartum survey in California. *Public Health Reports, 122,* 753-763.

Kim, S., Son, J., Kwok, P.K., Kang, J., Laken, F., Daquilanea, J., Shin, H., & Smith, T.W. (2012). Trends and correlates of income nonresponse: Forty years of the General Social Survey. *GSS Methodological Report 120*.

Krenzke, T. & Judkins, D. (2008). Filling in the blanks: Some guesses are better than others. *Chance, 21*(3), 7-13. doi:10.1007/s144-008-0003-9

Lasser, K. E., Himmelstein, D. U., & Woolhandler, S. (2006). Access to care, health status, and health disparities in the United States and Canada: results of a cross-national population-based survey. *American journal of public health, 96*(7), 1300–1307. https://doi.org/10.2105/AJPH.2004.059402

Lee, M., Rahbar, M.H., Gensler, L.S., Brown, M., Weisman, M., & Reveille, J.D. (2020). A latent class based imputation method under Bayesian quantile regression framework using asymmetric Laplace distribution for longitudinal medication usage data with intermittent missing values. *Journal of Biopharmaceutical Statistics, 30*(1), 160-177. doi:10.1080/10543406.2019.1684306

Little, R. J. A. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics, 6*(3), 287-296. doi:10.1080/07350015.1988.10509663

Long, Q., Hsu, C., & Li, Y. (2012). Doubly robust nonparametric multiple imputation for ignorable missing data. *Statistica Sinica, 22*(1), 149-172. doi:10.5705/ss.2010.069

Meng, X. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science, 9*(4), 538-558. doi:10.1214/ss/1177010269

Meyer, D. (2019). *Support vector machines: The interface to* libsvm *in package* e1071. Retrieved from: https://cran.r-project.org/web/packages/e1071/vignettes/svmdoc.pdf

Morgenstern, J.D., Buajitti, E., O'Neill, M., Piggott, T., Goel, V., Fridman, D., Kornas, K., & Rosella, L.C. (2020). Predicting population health with machine learning: A scoping review. *BMJ Open, 10*. doi:10.1136/bmjopen-2020-037860

Murphy, A. H. (1970). The ranked probability score and the probability score: A comparison. *Monthly Weather Review, 98*(12)

Murray, J. S. (2018). Multiple imputation: A review of practical and theoretical findings. *Statistical Science, 33*(2), 142-159. doi:10.1214/18-STS644

Nyongesa, D.B., Moon, H., Kim-Park, Y.H., Suaray, K., & Gao, T. (2016). Various considerations on performance measures for a classification of ordinal data: A thesis. *Department of Mathematics and Statistics, California State University; Long Beach, CA*.

Ogundimu, E. O., & Collins, G. S. (2019). A robust imputation method for missing responses and covariates in sample selection models. *Statistical Methods in Medical Research, 28*(1), 102-116. doi:10.1177/0962280217715663

Online edition (c) 2009 Cambridge UP.*Online edition (c) 2009 cambridge UP*

Quartagno, M., Carpenter, J. R., & Goldstein, H. (2019). Multiple Imputation with Survey Weights: A Multilevel Approach. *Journal of Survey Statistics and Methodology*. doi:10.1093/jssam/smz036

R Core Team (2020). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*; Vienna, Austria. URL: https://www.R-project.org/

Raghunathan, T. (2010). *Survey Inference with Incomplete Data.* Retrieved from https://www.cdc.gov/nchs/ppt/nchs2010/34_raghunathan.pdf

Rey de Castillo, P. (2014). On the use of data mining for imputation. Paper presented at the doi:10.13140/2.1.2780.2404 Retrieved from https://search.datacite.org/works/10.13140/2.1.2780.2404

Rubin, D. B. (1978). Multiple imputations in sample surveys: A phenomenological bayesian approach to nonresponse. *Annals of Statistics*, 6(*1*), 34-58

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley.

Sarndal, C.E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology,* 18(*2*), 241-252

Scheffer, J. (2002). Dealing with missing data. *Research Letters in the Information and Mathematical Sciences, 3*, 153-160

Schouten, R., Lugtig, P., Brand, J. & Vink, G. (2017). Generating missing values with ampute. Retrieved from https://www.gerkovink.com/Amputation_with_Ampute/Vignette/ampute.html

Shang, J., Qu, M., Liu, J., Kaplan, L. M., Han, J., & Peng, J. (2016). Meta-path guided embedding for similarity search in large-scale heterogeneous information networks Retrieved from https://www.openaire.eu/search/publication?articleId=od18::0d6dee94664d898c271cb1e8d7d139a9

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London; New York: Chapman and Hall.

StataCorp. (2017). Stata statistical software: Release 15 [computer software]. College Station, TX: StataCorp LLC:

Statistics Canada. (2016). Canadian Community Health Survey: Complement to the user guide public use microdata files 2014 and 2013-2014. Retrieved from https://www.statcan.gc.ca

Statistics Canada (2020). Data Liberation Initiative (DLI). Retrieved from https://www.statcan.gc.ca/eng/dli/dli

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., … Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ (Clinical research ed.)*, *338*, b2393. doi:10.1136/bmj.b2393

van Buuren, S., & Groothuis-Oudshoorn, K. (2011, 2020). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3), 1-67. Retrieved from https://www.jstatsoft.org/v45/i03/

van Buuren, S. (2012). *Flexible imputation of missing data*. Retrieved from https://ebookcentral.proquest.com

Verbeke, G. & Molenberghs, G. (2000). Linear mixed models for longitudinal data. *Springer series in statistics*. New York: Springer.

Venables, W. N. & Ripley, B. D. (2002). *Modern applied statistics with S. Fourth edition*. Springer, New York. ISBN 0-387-95457-0

Wilks, R., Younger, N., Mullings, J., Zohoori, N., Figueroa, P., Tulloch-Reid, M., . . . Ashley, D. (2007). Factors affecting study efficiency and item non-response in health surveys in developing countries: The Jamaica national healthy lifestyle survey. *BMC Medical Research Methodology, 7*(1). doi:10.1186/1471-2288-7-13

Xie, X., & Meng, X. (2017). Dissecting multiple imputation from a multi-phase inference perspective: What happens when god's, imputer's and analyst's models are uncongenial? *Statistica Sinica,* doi:10.5705/ss.2014.067

Yeung, C.W. & Thomas, S. (2013) Income imputation for the Canadian Community Health Survey. Statistics Canada. Retrieved from https://www.statcan.gc.ca

Zeng, Z. (2009) Multiple imputation for missing income data in population-based health surveillance. *Journal of Public Health Management Practice, 15*(6), E12-E21.

Appendix

Appendix Item 1. R code for missingness induction steps.

```
load("cchs_var5.RData")

cchs.dat = cchs.dat5.cc

cchs.datg = cchs.dat

str(cchs.datg)

### Label multiclass variables to have intelligible class names in dummy codes

cchs.datg$INCGHH = ordered(cchs.dat$INCGHH, levels = c(1,2,3,4,5),
                    labels = c("<20k","20k-39k","40k-59k","60k-79k",">80k"))

cchs.datg$DHHGAGE = ordered(cchs.dat$DHHGAGE, levels = c(1,2,3,4,5,6,7),
                     labels = c("18-29","30-39","40-49","50-59","60-69",
                          "70-79","80+"))

cchs.datg$DHHGMS = factor(cchs.dat$DHHGMS, levels = c(1,2,3,4),
                   labels = c("MARRIED","COMLAW","WSD","SINGLE"))

cchs.datg$DHHGLVG = factor(cchs.dat$DHHGLVG, levels = c(1,2,3,4,5,6,7,8),
                    labels = c("I_ALON","I_W/OT","W/SPPA","PS_W/C","1P_W/C",
                         "C_W/1P","C_W/PS","OTHER"))

cchs.datg$GEOGPRV = factor(cchs.dat$GEOGPRV, levels = c(10,11,12,13,24,35,
                                  46,47,48,59,60),
                    labels = c("NL","PE","NS","NB","QC","ON","MB","SK","AB",
                         "BC","YTNTNU"))

cchs.datg$EDUDR04 = factor(cchs.dat$EDUDR04, levels = c(1,2,3,4),
                    labels = c("<SS","SSGRAD","SOMEPS","PSGRAD"))

cchs.datg$GEN_01 = factor(cchs.dat$GEN_01, levels = c(1,2,3,4,5),
                   labels = c("EXCELL","VERY_G","GOOD","FAIR","POOR"))

cchs.datg$GEN_07 = factor(cchs.dat$GEN_07, levels = c(1,2,3,4,5),
                   labels = c("NO_STR","NV_STR","AB_STR","QB_STR","EX_STR"))

cchs.datg$RAC_1 = factor(cchs.dat$RAC_1, levels = c(1,2,3),
                  labels = c("SOME","OFTEN","NEVER"))
```

```
cchs.datg$SMK_202 = factor(cchs.dat$SMK_202, levels = c(1,2,3),
                           labels = c("DAILY","OCCASN","DO_NOT"))
cchs.datg$ALCDTTM = factor(cchs.dat$ALCDTTM, levels = c(1,2,3),
                           labels = c("REGULR","OCCASN","NOT12M"))
cchs.datg$SDC_5A_1 = factor(cchs.dat$SDC_5A_1, levels = c(1,2,3,4),
                            labels = c("ENG","FR","ENGFR","NEITHR"))
str(cchs.datg)
### Save dataset as original with no missing values
save(cchs.datg, file =
"~/School/MSc_CHE/Thesis/DataMiningAndMultipleImputation/Rscripts/cchs_dat
a_2\\cchs_ori.RData")
### Data set for simple MAR (5 variables inducing missingness in income)
### INCGHH.mar1 induced via DHHGAGE, DHH_SEX, GEOGPRV, EDUDR04, and
GEN01
#need 5, 15, 35% missing
vars1 = c("INCGHH", "DHHGAGE", "DHH_SEX", "EDUDR04", "GEOGPRV", "GEN_01")
cchs.mar1 = cchs.datg[, vars1]
### Dummy coding variables with more than 2 categories
cchs.mar1.dc = dummy_cols(cchs.mar1, select_columns = c("DHHGAGE","EDUDR04",
                                                        "GEOGPRV","GEN_01"),
                          remove_selected_columns = TRUE)
str(cchs.mar1.dc)
#creating columns with MAR-simple data
marpattern = c(0,
               1,1,1,1,1,
               1,1,1,1,1,
               1,1,1,1,1,
               1,1,1,1,1,
               1,1,1,1,1,
               1,1,1)
```

```
marweight = c(0,
        1,1,1,1,1,
        1,1,1,1,1,
        1,1,1,1,1,
        1,1,1,1,1,
        1,1,1,1,1,
        1,1,1)
### 5% Missing
set.seed(2019);mar1_5 = ampute(cchs.mar1.dc, patterns = marpattern, weights = marweight, mech = "MAR", type = "TAIL", cont = TRUE, prop = 0.05)
cchs.mar1_5 = mar1_5$amp
cchs.dat.mar1_5 = cchs.datg
cchs.dat.mar1_5$INCGHH = ordered(cchs.mar1_5$INCGHH, levels = c(1,2,3,4,5),
                labels = c("<20k","20k-39k","40k-59k","60k-79k",">80k"))
addmargins(table(cchs.dat.mar1_5$INCGHH))
# 50515/53065 = 0.952 = 95.2%
save(cchs.dat.mar1_5, file = "~/School/MSc_CHE/Thesis/DataMiningAndMultipleImputation/Rscripts/cchs_data_2\\mar1_05.RData")
### 15% Missing
set.seed(2019);mar1_15 = ampute(cchs.mar1.dc, patterns = marpattern, weights = marweight, mech = "MAR", type = "TAIL", cont = TRUE, prop = 0.15)
cchs.mar1_15 = mar1_15$amp
cchs.dat.mar1_15 = cchs.datg
cchs.dat.mar1_15$INCGHH = ordered(cchs.mar1_15$INCGHH, levels = c(1,2,3,4,5),
                labels = c("<20k","20k-39k","40k-59k","60k-79k",">80k"))
addmargins(table(cchs.dat.mar1_15$INCGHH))
# 45200/53075 = 0.852 = 85.2%
```

```
save(cchs.dat.mar1_15, file =
"~/School/MSc_CHE/Thesis/DataMiningAndMultipleImputation/Rscripts/cchs_dat
a_2\\mar1_15.RData")

### 35% Missing

set.seed(2019);mar1_35 = ampute(cchs.mar1.dc, patterns = marpattern, weights =
marweight, mech = "MAR", type = "TAIL", cont = TRUE, prop = 0.35)

cchs.mar1_35 = mar1_35$amp

cchs.dat.mar1_35 = cchs.datg

cchs.dat.mar1_35$INCGHH = ordered(cchs.mar1_35$INCGHH, levels = c(1,2,3,4,5),

                    labels = c("<20k","20k-39k","40k-59k","60k-79k",">80k"))

addmargins(table(cchs.dat.mar1_35$INCGHH))

# 34743/53075 = 0.655 = 65.5%

save(cchs.dat.mar1_35, file =
"~/School/MSc_CHE/Thesis/DataMiningAndMultipleImputation/Rscripts/cchs_dat
a_2\\mar1_35.RData")

### Data set for mid-complex MAR (8 variables + 1 interaction term)

### Variables: DHHGAGE, DHH_SEX, GEOGPRV, EDUDR04, GEN01, RAC_1,
SDCGCGT, CCC_101

### Interaction: DHHGAGE x SDCGCGT

#need 5, 15, 35% missing

vars2 = c("INCGHH", "DHHGAGE", "DHH_SEX", "EDUDR04", "GEOGPRV", "GEN_01",
"CCC_101",

      "RAC_1", "SDCGCGT")

cchs.mar2 = cchs.datg[, vars2]

str(cchs.mar2)

### Create an interaction term between DHHGAGE and SDCGCGT

cchs.mar2$AGE_RACE = as.factor(paste(cchs.mar2$DHHGAGE,
as.factor(cchs.mar2$SDCGCGT), sep=""))

### Dummy coding variables with more than 2 categories

cchs.mar2.dc = dummy_cols(cchs.mar2, select_columns = c("DHHGAGE","EDUDR04",

                                "GEOGPRV","GEN_01",

                                "RAC_1","AGE_RACE"),
```

```
                remove_selected_columns = TRUE)
str(cchs.mar2.dc)
#creating columns with MAR-mid-complex data
marpattern2 = c(0,
          1,1,1,1,1,
          1,1,1,1,1,
          1,1,1,1,1,
          1,1,1,1,1,
          1,1,1,1,1,
          1,1,1,1,1,
          1,1,1,1,1,
          1,1,1,1,1,
          1,1,1,1,1,
          1,1)
marweight2 = c(0,
          1,1,1,1,1,
          1,1,1,1,1,
          1,1,1,1,1,
          1,1,1,1,1,
          1,1,1,1,1,
          1,1,1,1,1,
          1,1,1,1,1,
          1,1,1,1,1,
          1,1,1,1,1,
          1,1)
### 5% Missing MAR2
set.seed(2019);mar2_5 = ampute(cchs.mar2.dc, patterns = marpattern2, weights =
marweight2, mech = "MAR", type = "TAIL", cont = TRUE, prop = 0.05)
cchs.mar2_5 = mar2_5$amp
addmargins(table(cchs.mar2_5$INCGHH))
```

```
cchs.dat.mar2_5 = cchs.datg

cchs.dat.mar2_5$INCGHH = ordered(cchs.mar2_5$INCGHH, levels = c(1,2,3,4,5),

                    labels = c("<20k","20k-39k","40k-59k","60k-79k",">80k"))

addmargins(table(cchs.dat.mar2_5$INCGHH))

# 50517/53065 = 0.952 = 95.2%

save(cchs.dat.mar2_5, file =
"~/School/MSc_CHE/Thesis/DataMiningAndMultipleImputation/Rscripts/cchs_dat
a_2\\mar2_05.RData")

### 15% Missing MAR2

set.seed(2019);mar2_15 = ampute(cchs.mar2.dc, patterns = marpattern2, weights =
marweight2, mech = "MAR", type = "TAIL", cont = TRUE, prop = 0.15)

cchs.mar2_15 = mar2_15$amp

addmargins(table(cchs.mar2_15$INCGHH))

cchs.dat.mar2_15 = cchs.datg

cchs.dat.mar2_15$INCGHH = ordered(cchs.mar2_15$INCGHH, levels = c(1,2,3,4,5),

                    labels = c("<20k","20k-39k","40k-59k","60k-79k",">80k"))

addmargins(table(cchs.dat.mar2_15$INCGHH))

# 45335/53065 = 0.854 = 85.4%

save(cchs.dat.mar2_15, file =
"~/School/MSc_CHE/Thesis/DataMiningAndMultipleImputation/Rscripts/cchs_dat
a_2\\mar2_15.RData")

### 35% Missing MAR2

set.seed(2019);mar2_35 = ampute(cchs.mar2.dc, patterns = marpattern2, weights =
marweight2, mech = "MAR", type = "TAIL", cont = TRUE, prop = 0.35)

cchs.mar2_35 = mar2_35$amp

addmargins(table(cchs.mar2_35$INCGHH))

cchs.dat.mar2_35 = cchs.datg

cchs.dat.mar2_35$INCGHH = ordered(cchs.mar2_35$INCGHH, levels = c(1,2,3,4,5),

                    labels = c("<20k","20k-39k","40k-59k","60k-79k",">80k"))

addmargins(table(cchs.dat.mar2_35$INCGHH))
```

```
# 35026/53065 = 0.661 = 66.1%

save(cchs.dat.mar2_35, file =
"~/School/MSc_CHE/Thesis/DataMiningAndMultipleImputation/Rscripts/cchs_dat
a_2\\mar2_35.RData")

### Data set for complex MAR (15 variables + 3 interaction term)

### Variables: DHHGAGE, DHH_SEX, GEOGPRV, EDUDR04, GEN_01, RAC_1,
SDCGCGT,

###   CCC_101, DHHGMS, GEN_07, UCN_010, ALCDTTM, CCC_31A, CHP_14, CCC_071

### Interactions: DHHGAGE x SDCGCGT, DHHGMS x EDUDR04, CCC_071 x
GEOGPRV

#need 5, 15, 35% missing

vars3 = c("INCGHH", "DHHGAGE", "DHH_SEX", "DHHGMS", "EDUDR04", "GEOGPRV",

      "GEN_01", "GEN_07", "CCC_071", "CCC_101", "CCC_31A", "CHP_14",

      "UCN_010", "ALCDTTM","RAC_1", "SDCGCGT")

cchs.mar3 = cchs.datg[, vars3]

str(cchs.mar3)

### Create an interaction term between DHHGAGE and SDCGCGT

cchs.mar3$AGE_RACE = as.factor(paste(cchs.mar3$DHHGAGE,
as.factor(cchs.mar3$SDCGCGT), sep=""))

### Create an interaction term between DHHGMS and EDUDR04

cchs.mar3$MAR_EDU = as.factor(paste(cchs.mar3$DHHGMS, cchs.mar3$EDUDR04,
sep=""))

### Create an interaction term between GEOGPRV and CCC_071

cchs.mar3$GEO_HBP = as.factor(paste(cchs.mar3$GEOGPRV,
as.factor(cchs.mar3$CCC_071), sep=""))

### Dummy coding variables with more than 2 categories

cchs.mar3.dc = dummy_cols(cchs.mar3, select_columns = c("DHHGAGE","DHHGMS",

                                    "EDUDR04","GEOGPRV",

                                    "GEN_01","GEN_07",

                                    "ALCDTTM","RAC_1",

                                    "AGE_RACE","MAR_EDU",

                                    "GEO_HBP"
```

```
),
remove_selected_columns = TRUE)
str(cchs.mar3.dc)
#creating columns with MAR-complex data
marpattern3 = c(0,
                1,1,1,1,1,
                1,1,1,1,1,
                1,1,1,1,1,
                1,1,1,1,1,
                1,1,1,1,1,
                1,1,1,1,1,
                1,1,1,1,1,
                1,1,1,1,1,
                1,1,1,1,1,
                1,1,1,1,1,
                1,1,1,1,1,
                1,1,1,1,1,
                1,1,1,1,1,
                1,1,1,1,1,
                1,1,1,1,1,
                1,1,1,1,1,
                1,1,1,1,1,
                1,1,1,1,1,
                1,1,1,1,1,
                1)
marweight3 = c(0,
               1,1,1,1,1,
               1,1,1,1,1,
```

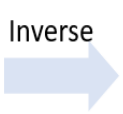
```
1,1,1,1,1,
1,1,1,1,1,
1,1,1,1,1,
1,1,1,1,1,
1,1,1,1,1,
1,1,1,1,1,
1,1,1,1,1,
1,1,1,1,1,
1,1,1,1,1,
1,1,1,1,1,
1,1,1,1,1,
1,1,1,1,1,
1,1,1,1,1,
1,1,1,1,1,
1,1,1,1,1,
1,1,1,1,1,
1,1,1,1,1,
1,1,1,1,1,
1,1,1,1,1,
1)
### 5% Missing MAR3
set.seed(2019);mar3_5 = ampute(cchs.mar3.dc, patterns = marpattern3, weights =
marweight3, mech = "MAR", type = "TAIL", cont = TRUE, prop = 0.05)
cchs.mar3_5 = mar3_5$amp
addmargins(table(cchs.mar3_5$INCGHH))
cchs.dat.mar3_5 = cchs.datg
cchs.dat.mar3_5$INCGHH = ordered(cchs.mar3_5$INCGHH, levels = c(1,2,3,4,5),
                    labels = c("<20k","20k-39k","40k-59k","60k-79k",">80k"))
addmargins(table(cchs.dat.mar3_5$INCGHH))
# 50440/53065 = 0.951 = 95.1%
```

```
save(cchs.dat.mar3_5, file =
"~/School/MSc_CHE/Thesis/DataMiningAndMultipleImputation/Rscripts/cchs_dat
a_2\\mar3_05.RData")

### 15% Missing MAR3

set.seed(2019);mar3_15 = ampute(cchs.mar3.dc, patterns = marpattern3, weights =
marweight3, mech = "MAR", type = "TAIL", cont = TRUE, prop = 0.15)

cchs.mar3_15 = mar3_15$amp

addmargins(table(cchs.mar3_15$INCGHH))

cchs.dat.mar3_15 = cchs.datg

cchs.dat.mar3_15$INCGHH = ordered(cchs.mar3_15$INCGHH, levels = c(1,2,3,4,5),

                    labels = c("<20k","20k-39k","40k-59k","60k-79k",">80k"))

addmargins(table(cchs.dat.mar3_15$INCGHH))

# 45213/53065 = 0.852 = 85.2%

save(cchs.dat.mar3_15, file =
"~/School/MSc_CHE/Thesis/DataMiningAndMultipleImputation/Rscripts/cchs_dat
a_2\\mar3_15.RData")

### 35% Missing MAR3

set.seed(2019);mar3_35 = ampute(cchs.mar3.dc, patterns = marpattern3, weights =
marweight3, mech = "MAR", type = "TAIL", cont = TRUE, prop = 0.35)

cchs.mar3_35 = mar3_35$amp

addmargins(table(cchs.mar3_35$INCGHH))

cchs.dat.mar3_35 = cchs.datg

cchs.dat.mar3_35$INCGHH = ordered(cchs.mar3_35$INCGHH, levels = c(1,2,3,4,5),

                    labels = c("<20k","20k-39k","40k-59k","60k-79k",">80k"))

addmargins(table(cchs.dat.mar3_35$INCGHH))

# 34872/53065 = 0.657 = 65.7%

save(cchs.dat.mar3_35, file =
"~/School/MSc_CHE/Thesis/DataMiningAndMultipleImputation/Rscripts/cchs_dat
a_2\\mar3_35.RData")

### Create MNAR dataset

cchs.mnar = cchs.datg
```

```
mnarpattern = c(0,
        1,1,1,1,1,
        1,1,1,1,1,
        1,1,1,1,1,
        1,1,1,1,1,
        1,1,1,1,1,
        1,1,1,1,1
        )


mnarweights = c(1,
        0,0,0,0,0,
        0,0,0,0,0,
        0,0,0,0,0,
        0,0,0,0,0,
        0,0,0,0,0,
        0,0,0,0,0
        )
#MNAR 5%
set.seed(2019);mnar05 = ampute(cchs.mnar, patterns = mnarpattern, weights =
mnarweights, mech = "MNAR", type = "TAIL", cont = TRUE, prop = 0.05)
cchs.mnar05 = mnar05$amp
addmargins(table(cchs.mnar05$INCGHH))
cchs.dat.mnar05 = cchs.datg
cchs.dat.mnar05$INCGHH = ordered(cchs.mnar05$INCGHH, levels = c(1,2,3,4,5),
                labels = c("<20k","20k-39k","40k-59k","60k-79k",">80k"))
addmargins(table(cchs.dat.mnar05$INCGHH))
# 50495/53065 = 0.952 = 95.2%
```

```
save(cchs.dat.mnar05, file =
"~/School/MSc_CHE/Thesis/DataMiningAndMultipleImputation/Rscripts/cchs_dat
a_2\\mnar05.RData")

#MNAR 15%

set.seed(2019);mnar15 = ampute(cchs.mnar, patterns = mnarpattern, weights =
mnarweights, mech = "MNAR", type = "TAIL", cont = TRUE, prop = 0.15)

cchs.mnar15 = mnar15$amp

addmargins(table(cchs.mnar15$INCGHH))

cchs.dat.mnar15 = cchs.datg

cchs.dat.mnar15$INCGHH = ordered(cchs.mnar15$INCGHH, levels = c(1,2,3,4,5),

                    labels = c("<20k","20k-39k","40k-59k","60k-79k",">80k"))

addmargins(table(cchs.dat.mnar15$INCGHH))

# 44884/53065 = 0.846 = 84.6%

save(cchs.dat.mnar15, file =
"~/School/MSc_CHE/Thesis/DataMiningAndMultipleImputation/Rscripts/cchs_dat
a_2\\mnar15.RData")

#MNAR 35%

set.seed(2019);mnar35 = ampute(cchs.mnar, patterns = mnarpattern, weights =
mnarweights, mech = "MNAR", type = "TAIL", cont = TRUE, prop = 0.35)

cchs.mnar35 = mnar35$amp

addmargins(table(cchs.mnar35$INCGHH))

cchs.dat.mnar35 = cchs.datg

cchs.dat.mnar35$INCGHH = ordered(cchs.mnar35$INCGHH, levels = c(1,2,3,4,5),

                    labels = c("<20k","20k-39k","40k-59k","60k-79k",">80k"))

addmargins(table(cchs.dat.mnar35$INCGHH))

# 50495/53065 = 0.631 = 63.1%

save(cchs.dat.mnar35, file =
"~/School/MSc_CHE/Thesis/DataMiningAndMultipleImputation/Rscripts/cchs_dat
a_2\\mnar35.RData")
```

	Class 1	Class 2	Class 3	N
Class 1	0	10/80	10/30	10
Class 2	20/90	0	20/30	20
Class 3	70/90	70/80	0	70
N	10	20	70	100



	Class 1	Class 2	Class 3	N
Class 1	0	80/10	30/10	10
Class 2	90/20	0	30/20	20
Class 3	90/70	80/70	0	70
N	10	20	70	100

Figure 1. Example table of how to calculate the inverse probability of misclassification for a 3-class categorical variable.

Cost	Class 1	Class 2	Class 3
Class 1	0	1	2
Class 2	1	0	1
Class 3	2	1	0

Figure 2. Example table of how to represent linear absolute value loss for a 3-class categorical variable.

Figure 3. Ordinal misclassification cost calculated as the product of inverse probability of misclassification and linear absolute value loss.

Appendix List 1. Variables selected in processed Canadian Community Health Survey dataset

Income = INCGHH (total household income from all sources) | 5 categories, 66 Not stated

---

Demographics - 6

Age = DHHGHAGE (remove under 18) | 16 categories, 0 missing, 4948 u18

Sex = DHH_SEX | binary, 0 missing

Marital status = DHHGMS | 4 categories, 130 Not stated

Living arrangement = DHHGLVG (household make-up) | 8 categories, 325 Not stated

Education = EDUDR04 (highest completed by respondent) | 4 categories, 1021 Not stated

Province = GEOGPRV (province of residence) | 11 categories, 0 missing

---

General Health - 2

Self-perceived health = GEN_01 | 5 categories, 90 Don't know, 8 Refusal

Perceived life stress = GEN_07 | 5 categories, 352 Don't know, 31 Refusal

---

Chronic Conditions - 11

Asthma = CCC_031 | binary, 73 Don't know, 13 Refusal

Fibromyalgia = CCC_041 | binary, 113 Don't know, 7 Refusal, 13 Not stated

Arthritis = CCC_051 | binary, 1588 Not applicable, 184 Don't know, 7 Refusal, 829 Not stated

Back problems = CCC_061 | binary, 108 Don't know, 8 Refusal, 9 Not stated

High blood pressure = CCC_071 | binary, 192 Don't know, 11 Refusal, 13 Not stated

Migraine headaches = CCC_081 | binary, 58 Don't know, 8 Refusal, 13 Not stated

Diabetes = CCC_101 | binary, 68 Don't know, 6 Refusal, 13 Not stated

Heart disease = CCC_121 | binary, 194 Don't know, 8 Refusal, 13 Not stated

Ever had cancer = CCC_31A | (Y/N + currently has cancer as Not Applicable), 44 Don't know, 3 Refusal, 21 Not stated

Mood disorder = CCC_280 | binary, 94 Don't know, 25 Refusal, 13 Not stated

Anxiety disorder = CCC_290 | binary, 124 Don't know, 24 Refusal, 13 Not stated

---

Healthcare Utilization - 4

Overnight patient = CHP_01 (within the past 12 months) | binary, 35 Don't know, 13 Refusal

Family doctor = CHP_03 (ever visited?) | binary, 88 Don't know, 9 Refusal, 13 Not stated

Dentist = CHP_14 (ever visited?) | binary, 54 Don't know, 3 Refusal, 13 Not stated

Unmet healthcare needs = UCN_010 (within the past 12 months) | binary, 141 Don't know, 20 Refusal

---

Other Health-Related - 4

Injury = INJ_01 (within the past 12 months) | binary, 78 Don't know, 66 Refusal

Restriction of activity = RAC_1| 3 categories, 91 Don't know, 16 Refusal

Smoking = SMK_202| 3 categories, 26 Don't know, 10 Refusal, 312 Not stated

Alcohol = ALCDTTM (type of drinker) | 3 categories, 1359 Not stated

---

Socio-demographics - 3

Country of birth = SDCGCB13 | binary (Canada/other), 1946 Not stated

Knowledge of official languages = SDC_5A_1 | 4 categories, 9 Don't know, 34 Refusal, 1607 Not stated

Cultural/racial origin = SDCGCGT | binary (white or visible minority), 2294 Not stated

Table A1. Summary of explanatory variables by total household income group.

	<20k (N=5391)	>80k (N=18768)	20k-39k (N=11287)	40k-59k (N=9834)	60k-79k (N=7785)	Overall (N=53065)
Age						
18-29	551 (10.2%)	3423 (18.2%)	1153 (10.2%)	1250 (12.7%)	1188 (15.3%)	7565 (14.3%)
30-39	363 (6.7%)	3582 (19.1%)	715 (6.3%)	919 (9.3%)	976 (12.5%)	6555 (12.4%)
40-49	376 (7.0%)	3138 (16.7%)	689 (6.1%)	874 (8.9%)	912 (11.7%)	5989 (11.3%)
50-59	875 (16.2%)	4290 (22.9%)	1456 (12.9%)	1692 (17.2%)	1576 (20.2%)	9889 (18.6%)
60-69	1299 (24.1%)	2953 (15.7%)	2804 (24.8%)	2626 (26.7%)	1857 (23.9%)	11539 (21.7%)
70-79	1060 (19.7%)	1029 (5.5%)	2724 (24.1%)	1693 (17.2%)	918 (11.8%)	7424 (14.0%)
80+	867 (16.1%)	353 (1.9%)	1746 (15.5%)	780 (7.9%)	358 (4.6%)	4104 (7.7%)
Sex						
Male	1788 (33.2%)	9399 (50.1%)	4292 (38.0%)	4266 (43.4%)	3655 (46.9%)	23400 (44.1%)
Female	3603 (66.8%)	9369 (49.9%)	6995 (62.0%)	5568 (56.6%)	4130 (53.1%)	29665 (55.9%)
Marital Status						
Common Law	190 (3.5%)	2427 (12.9%)	575 (5.1%)	812 (8.3%)	840 (10.8%)	4844 (9.1%)
Married	666 (12.4%)	11447 (61.0%)	3930 (34.8%)	4564 (46.4%)	4165 (53.5%)	24772 (46.7%)
Single	1644 (30.5%)	3638 (19.4%)	2508 (22.2%)	2202 (22.4%)	1683 (21.6%)	11675 (22.0%)
Widowed, Separated, Divorced	2891 (53.6%)	1256 (6.7%)	4274 (37.9%)	2256 (22.9%)	1097 (14.1%)	11774 (22.2%)
Household Composition Living Arrangement						
Single parent with children	293 (5.4%)	273 (1.5%)	487 (4.3%)	331 (3.4%)	238 (3.1%)	1622 (3.1%)
Child living with one parent	101 (1.9%)	337 (1.8%)	292 (2.6%)	285 (2.9%)	258 (3.3%)	1273 (2.4%)
Child living with parents	86 (1.6%)	2006 (10.7%)	271 (2.4%)	453 (4.6%)	471 (6.1%)	3287 (6.2%)
Individual living alone	3993 (74.1%)	1697 (9.0%)	5457 (48.3%)	3092 (31.4%)	1583 (20.3%)	15822 (29.8%)
Individual living with others	145 (2.7%)	460 (2.5%)	387 (3.4%)	328 (3.3%)	225 (2.9%)	1545 (2.9%)
Other	69 (1.3%)	1002 (5.3%)	226 (2.0%)	303 (3.1%)	290 (3.7%)	1890 (3.6%)

	<20k (N=5391)	>80k (N=18768)	20k-39k (N=11287)	40k-59k (N=9834)	60k-79k (N=7785)	Overall (N=53065)
Parents living with children	171 (3.2%)	5316 (28.3%)	586 (5.2%)	936 (9.5%)	1204 (15.5%)	8213 (15.5%)
Living with spouse/partner	533 (9.9%)	7677 (40.9%)	3581 (31.7%)	4106 (41.8%)	3516 (45.2%)	19413 (36.6%)
Highest Completed Education						
Less than secondary school	2291 (42.5%)	1073 (5.7%)	3451 (30.6%)	1654 (16.8%)	825 (10.6%)	9294 (17.5%)
Post-secondary school graduate	1751 (32.5%)	13368 (71.2%)	4693 (41.6%)	5355 (54.5%)	4803 (61.7%)	29970 (56.5%)
Some post-secondary school	273 (5.1%)	902 (4.8%)	534 (4.7%)	468 (4.8%)	382 (4.9%)	2559 (4.8%)
Secondary school	1076 (20.0%)	3425 (18.2%)	2609 (23.1%)	2357 (24.0%)	1775 (22.8%)	11242 (21.2%)
Province of Residence						
Alberta	309 (5.7%)	2469 (13.2%)	821 (7.3%)	802 (8.2%)	731 (9.4%)	5132 (9.7%)
British Columbia	592 (11.0%)	2306 (12.3%)	1397 (12.4%)	1227 (12.5%)	993 (12.8%)	6515 (12.3%)
Manitoba	278 (5.2%)	1104 (5.9%)	635 (5.6%)	608 (6.2%)	479 (6.2%)	3104 (5.8%)
New Brunswick	292 (5.4%)	458 (2.4%)	613 (5.4%)	419 (4.3%)	301 (3.9%)	2083 (3.9%)
Newfoundland and Labrador	239 (4.4%)	541 (2.9%)	417 (3.7%)	341 (3.5%)	208 (2.7%)	1746 (3.3%)
Nova Scotia	325 (6.0%)	600 (3.2%)	586 (5.2%)	404 (4.1%)	337 (4.3%)	2252 (4.2%)
Ontario	1596 (29.6%)	6339 (33.8%)	3489 (30.9%)	3279 (33.3%)	2610 (33.5%)	17313 (32.6%)
PEI	112 (2.1%)	183 (1.0%)	210 (1.9%)	220 (2.2%)	138 (1.8%)	863 (1.6%)
Quebec	1200 (22.3%)	2795 (14.9%)	2331 (20.7%)	1886 (19.2%)	1402 (18.0%)	9614 (18.1%)
Saskatchewan	291 (5.4%)	1271 (6.8%)	597 (5.3%)	497 (5.1%)	424 (5.4%)	3080 (5.8%)
Yukon, Northwest Territories, Nunavut	157 (2.9%)	702 (3.7%)	191 (1.7%)	151 (1.5%)	162 (2.1%)	1363 (2.6%)
Self-Perceived Health						
Excellent	562 (10.4%)	4510 (24.0%)	1536 (13.6%)	1730 (17.6%)	1460 (18.8%)	9798 (18.5%)
Fair	1170 (21.7%)	1026 (5.5%)	1714 (15.2%)	1034 (10.5%)	654 (8.4%)	5598 (10.5%)

	<20k (N=5391)	>80k (N=18768)	20k-39k (N=11287)	40k-59k (N=9834)	60k-79k (N=7785)	Overall (N=53065)
Good	1859 (34.5%)	4869 (25.9%)	3991 (35.4%)	3136 (31.9%)	2315 (29.7%)	16170 (30.5%)
Poor	547 (10.1%)	271 (1.4%)	587 (5.2%)	332 (3.4%)	172 (2.2%)	1909 (3.6%)
Very Good	1253 (23.2%)	8092 (43.1%)	3459 (30.6%)	3602 (36.6%)	3184 (40.9%)	19590 (36.9%)
Perceived Stress						
A bit stressful	1988 (36.9%)	8101 (43.2%)	4153 (36.8%)	3845 (39.1%)	3118 (40.1%)	21205 (40.0%)
Extremely stressful	241 (4.5%)	480 (2.6%)	289 (2.6%)	242 (2.5%)	186 (2.4%)	1438 (2.7%)
Not at all stressful	1031 (19.1%)	1789 (9.5%)	2209 (19.6%)	1539 (15.6%)	1046 (13.4%)	7614 (14.3%)
Not very stressful	1232 (22.9%)	4592 (24.5%)	3103 (27.5%)	2821 (28.7%)	2157 (27.7%)	13905 (26.2%)
Quite a bit stressful	899 (16.7%)	3806 (20.3%)	1533 (13.6%)	1387 (14.1%)	1278 (16.4%)	8903 (16.8%)
Asthma						
No	4727 (87.7%)	17394 (92.7%)	10313 (91.4%)	9067 (92.2%)	7216 (92.7%)	48717 (91.8%)
Yes	664 (12.3%)	1374 (7.3%)	974 (8.6%)	767 (7.8%)	569 (7.3%)	4348 (8.2%)
Fibromyalgia						
No	5089 (94.4%)	18543 (98.8%)	10933 (96.9%)	9588 (97.5%)	7649 (98.3%)	51802 (97.6%)
Yes	302 (5.6%)	225 (1.2%)	354 (3.1%)	246 (2.5%)	136 (1.7%)	1263 (2.4%)
Arthritis						
No	3172 (58.8%)	15898 (84.7%)	7261 (64.3%)	7072 (71.9%)	6047 (77.7%)	39450 (74.3%)
Yes	2219 (41.2%)	2870 (15.3%)	4026 (35.7%)	2762 (28.1%)	1738 (22.3%)	13615 (25.7%)
Back Problems						
No	3711 (68.8%)	15410 (82.1%)	8410 (74.5%)	7719 (78.5%)	6162 (79.2%)	41412 (78.0%)
Yes	1680 (31.2%)	3358 (17.9%)	2877 (25.5%)	2115 (21.5%)	1623 (20.8%)	11653 (22.0%)
High Blood Pressure						
No	3430 (63.6%)	15693 (83.6%)	7133 (63.2%)	6929 (70.5%)	5875 (75.5%)	39060 (73.6%)
Yes	1961 (36.4%)	3075 (16.4%)	4154 (36.8%)	2905 (29.5%)	1910 (24.5%)	14005 (26.4%)
Migraine Headaches						
No	4763 (88.4%)	17017 (90.7%)	10339 (91.6%)	8953 (91.0%)	7050 (90.6%)	48122 (90.7%)

	<20k (N=5391)	>80k (N=18768)	20k-39k (N=11287)	40k-59k (N=9834)	60k-79k (N=7785)	Overall (N=53065)
Yes	628 (11.6%)	1751 (9.3%)	948 (8.4%)	881 (9.0%)	735 (9.4%)	4943 (9.3%)
Diabetes						
No	4496 (83.4%)	17725 (94.4%)	9649 (85.5%)	8816 (89.6%)	7148 (91.8%)	47834 (90.1%)
Yes	895 (16.6%)	1043 (5.6%)	1638 (14.5%)	1018 (10.4%)	637 (8.2%)	5231 (9.9%)
Heart Disease						
No	4685 (86.9%)	18015 (96.0%)	9945 (88.1%)	9020 (91.7%)	7289 (93.6%)	48954 (92.3%)
Yes	706 (13.1%)	753 (4.0%)	1342 (11.9%)	814 (8.3%)	496 (6.4%)	4111 (7.7%)
Ever had Cancer						
No	4584 (85.0%)	17505 (93.3%)	9600 (85.1%)	8609 (87.5%)	6994 (89.8%)	47292 (89.1%)
Yes	807 (15.0%)	1263 (6.7%)	1687 (14.9%)	1225 (12.5%)	791 (10.2%)	5773 (10.9%)
Mood Disorder						
No	4396 (81.5%)	17621 (93.9%)	10143 (89.9%)	8971 (91.2%)	7211 (92.6%)	48342 (91.1%)
Yes	995 (18.5%)	1147 (6.1%)	1144 (10.1%)	863 (8.8%)	574 (7.4%)	4723 (8.9%)
Anxiety Disorder						
No	4628 (85.8%)	17773 (94.7%)	10388 (92.0%)	9141 (93.0%)	7290 (93.6%)	49220 (92.8%)
Yes	763 (14.2%)	995 (5.3%)	899 (8.0%)	693 (7.0%)	495 (6.4%)	3845 (7.2%)
Overnight Patient (within past 12 months)						
No	4501 (83.5%)	17424 (92.8%)	9921 (87.9%)	8875 (90.2%)	7149 (91.8%)	47870 (90.2%)
Yes	890 (16.5%)	1344 (7.2%)	1366 (12.1%)	959 (9.8%)	636 (8.2%)	5195 (9.8%)
Family Doctor (ever visited)						
No	1087 (20.2%)	4163 (22.2%)	2107 (18.7%)	2025 (20.6%)	1611 (20.7%)	10993 (20.7%)
Yes	4304 (79.8%)	14605 (77.8%)	9180 (81.3%)	7809 (79.4%)	6174 (79.3%)	42072 (79.3%)
Dentist Orthodontist (ever visited)						
No	3295 (61.1%)	4245 (22.6%)	5784 (51.2%)	3842 (39.1%)	2405 (30.9%)	19571 (36.9%)
Yes	2096 (38.9%)	14523 (77.4%)	5503 (48.8%)	5992 (60.9%)	5380 (69.1%)	33494 (63.1%)

	<20k (N=5391)	>80k (N=18768)	20k-39k (N=11287)	40k-59k (N=9834)	60k-79k (N=7785)	Overall (N=53065)
Unmet Healthcare Needs						
No	4581 (85.0%)	16925 (90.2%)	10093 (89.4%)	8800 (89.5%)	7010 (90.0%)	47409 (89.3%)
Yes	810 (15.0%)	1843 (9.8%)	1194 (10.6%)	1034 (10.5%)	775 (10.0%)	5656 (10.7%)
Injury (within the past 12 months)						
No	4634 (86.0%)	15709 (83.7%)	9880 (87.5%)	8564 (87.1%)	6652 (85.4%)	45439 (85.6%)
Yes	757 (14.0%)	3059 (16.3%)	1407 (12.5%)	1270 (12.9%)	1133 (14.6%)	7626 (14.4%)
Restriction of Activity						
Never	2642 (49.0%)	14475 (77.1%)	6626 (58.7%)	6495 (66.0%)	5506 (70.7%)	35744 (67.4%)
Often	1385 (25.7%)	1371 (7.3%)	2053 (18.2%)	1270 (12.9%)	821 (10.5%)	6900 (13.0%)
Sometimes	1364 (25.3%)	2922 (15.6%)	2608 (23.1%)	2069 (21.0%)	1458 (18.7%)	10421 (19.6%)
Type of Smoker						
Daily	1387 (25.7%)	2158 (11.5%)	1856 (16.4%)	1486 (15.1%)	1025 (13.2%)	7912 (14.9%)
Does not smoke	3787 (70.2%)	15791 (84.1%)	9037 (80.1%)	7953 (80.9%)	6442 (82.7%)	43010 (81.1%)
Occasional	217 (4.0%)	819 (4.4%)	394 (3.5%)	395 (4.0%)	318 (4.1%)	2143 (4.0%)
Type of Drinker (Alcohol)						
Not within the past 12 months	2147 (39.8%)	2290 (12.2%)	3431 (30.4%)	2072 (21.1%)	1385 (17.8%)	11325 (21.3%)
Occasional	1218 (22.6%)	2460 (13.1%)	2453 (21.7%)	1901 (19.3%)	1282 (16.5%)	9314 (17.6%)
Regular	2026 (37.6%)	14018 (74.7%)	5403 (47.9%)	5861 (59.6%)	5118 (65.7%)	32426 (61.1%)
Country of Birth						
Other	816 (15.1%)	2619 (14.0%)	1955 (17.3%)	1634 (16.6%)	1196 (15.4%)	8220 (15.5%)
Canada	4575 (84.9%)	16149 (86.0%)	9332 (82.7%)	8200 (83.4%)	6589 (84.6%)	44845 (84.5%)
Knowledge of Official Languages						
English	3737 (69.3%)	14095 (75.1%)	8088 (71.7%)	7121 (72.4%)	5761 (74.0%)	38802 (73.1%)

	<20k (N=5391)	>80k (N=18768)	20k-39k (N=11287)	40k-59k (N=9834)	60k-79k (N=7785)	Overall (N=53065)
English and French	739 (13.7%)	3534 (18.8%)	1497 (13.3%)	1548 (15.7%)	1247 (16.0%)	8565 (16.1%)
French	844 (15.7%)	1101 (5.9%)	1614 (14.3%)	1115 (11.3%)	747 (9.6%)	5421 (10.2%)
Neither	71 (1.3%)	38 (0.2%)	88 (0.8%)	50 (0.5%)	30 (0.4%)	277 (0.5%)
Cultural / Racial Origin						
Visible minority	920 (17.1%)	2421 (12.9%)	1494 (13.2%)	1255 (12.8%)	978 (12.6%)	7068 (13.3%)
White	4471 (82.9%)	16347 (87.1%)	9793 (86.8%)	8579 (87.2%)	6807 (87.4%)	45997 (86.7%)

d-statistic of imputation models for total household income in CCHS2014 with induced MAR (mechanism 1: simple) by proportion of missing data

MAR-1	% Missing		
Imputation model	5	15	35
RPART	0.5737	0.5837	0.5855
RPART_SC	0.5518	0.5381	0.5396
SVM	0.5322	0.5124	0.5115
NNET	**0.5231**	**0.5027**	0.5087
O-RF	0.5655	0.5092	**0.5085**
O-REG-MI	0.7584	0.7634	0.7626
PMM-MI	0.6973	0.7001	0.6998

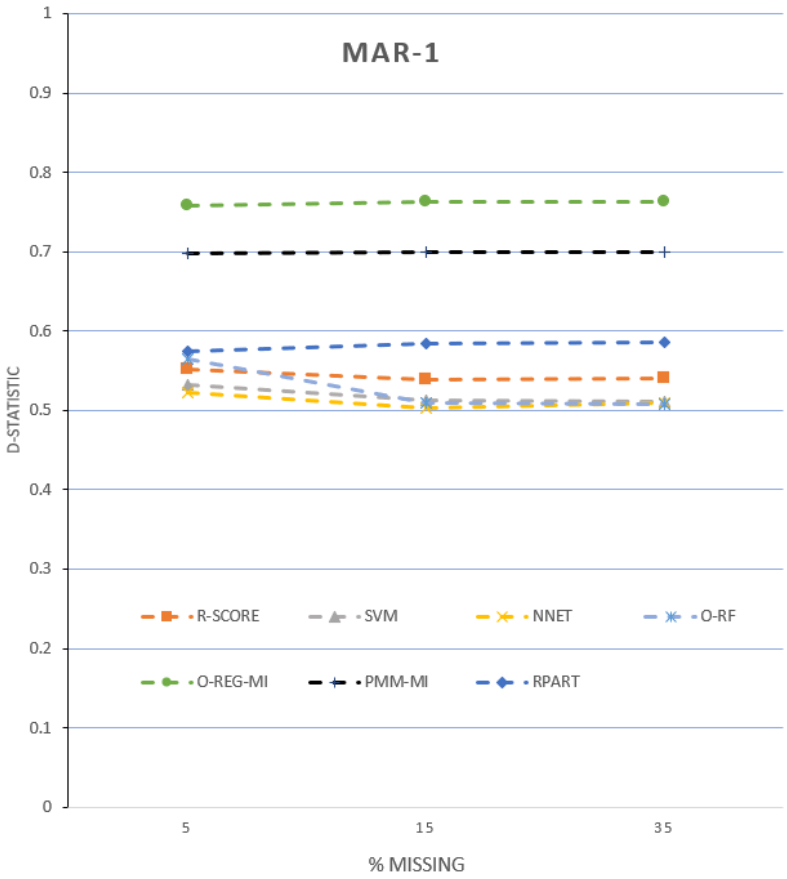


Figure 9. d-statistics of imputation models for total household income in Canadian Community Health Survey 2014 with induced MAR (mechanism 1: simple) by proportion of missing data

d-statistic of imputation models for total household income in CCHS2014 with induced MAR (mechanism 2: mid-complex) by proportion of missing data

MAR-2	% Missing		
Imputation model	5	15	35
RPART_BAG	0.5463	0.5461	0.5516
RPART_SC	0.5483	0.5395	0.5443
SVM	0.5239	0.5162	0.5178
NNET	**0.5020**	**0.5035**	**0.5098**
O-RF	0.5055	0.5082	0.5112
O-REG-MI	0.7696	0.7664	0.7545
PMM-MI	0.7116	0.6968	0.6990

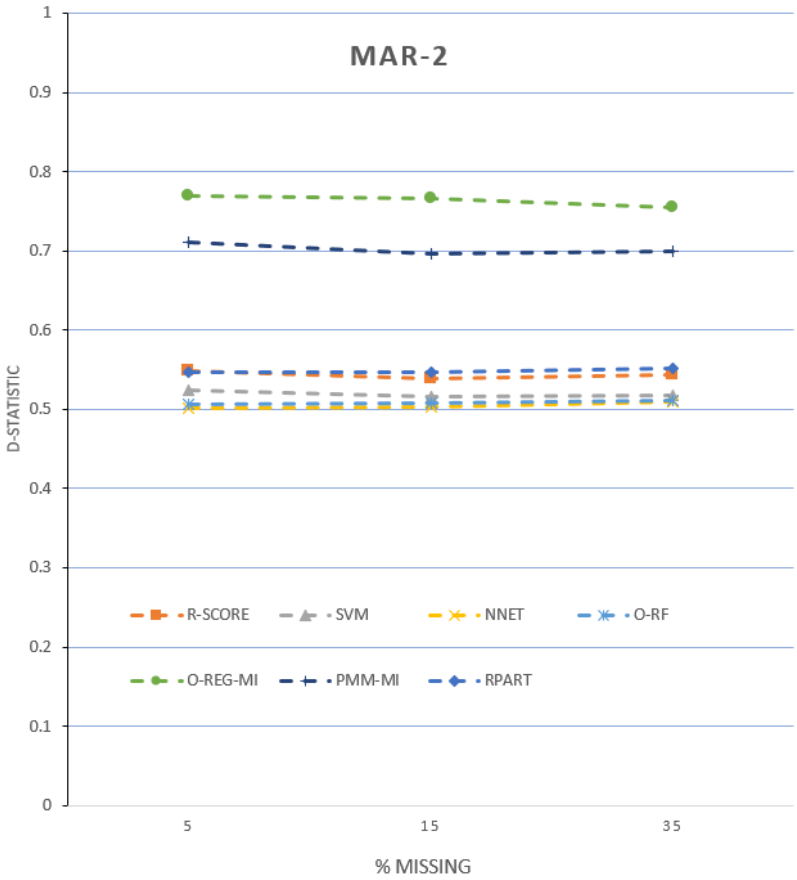


Figure 10. d-statistics of imputation models for total household income in Canadian Community Health Survey 2014 with induced MAR (mechanism 2: mid-complex) by proportion of missing data.

d-statistic of imputation models for total household income in CCHS2014 with induced MAR (mechanism 3: complex) by proportion of missing data

MAR-3	% Missing		
Imputation model	5	15	35
RPART	0.5817	0.5788	0.6042
RPART_SC	0.5512	0.5415	0.5394
SVM	0.5257	0.5107	0.5173
NNET	0.5223	**0.5022**	**0.5036**
O-RF	**0.5166**	0.5139	0.5089
O-REG-MI	0.7638	0.7606	0.7614
PMM-MI	0.6743	0.6581	0.6586

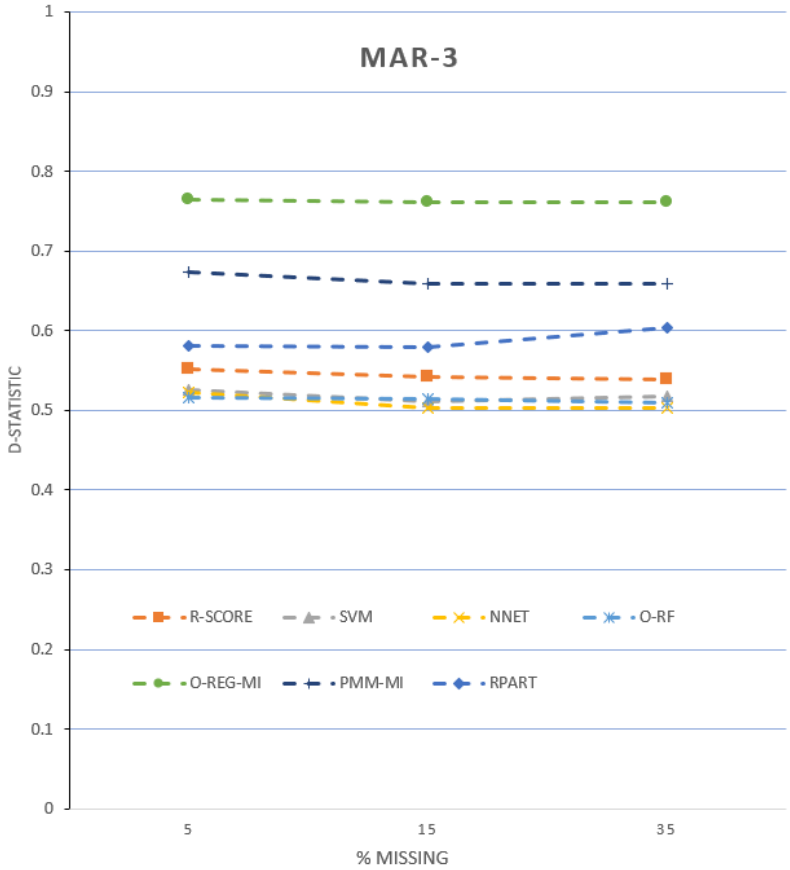


Figure 11. d-statistics of imputation models for total household income in Canadian Community Health Survey 2014 with induced MAR (mechanism 3: complex) by proportion of missing data.

d-statistic of imputation models for total household income in CCHS2014 with induced MNAR by proportion of missing data

MNAR	% Missing		
Imputation model	5	15	35
RPART	0.7121	0.6696	0.6948
RPART_SC	0.4704	0.4858	0.5292
SVM	0.4809	0.4970	0.5128
NNET	**0.4405**	**0.4611**	**0.4789**
O-RF	0.4494	0.4635	0.4796
O-REG-MI	0.7518	0.7549	0.7572
PMM-MI	0.6397	0.6540	0.6573

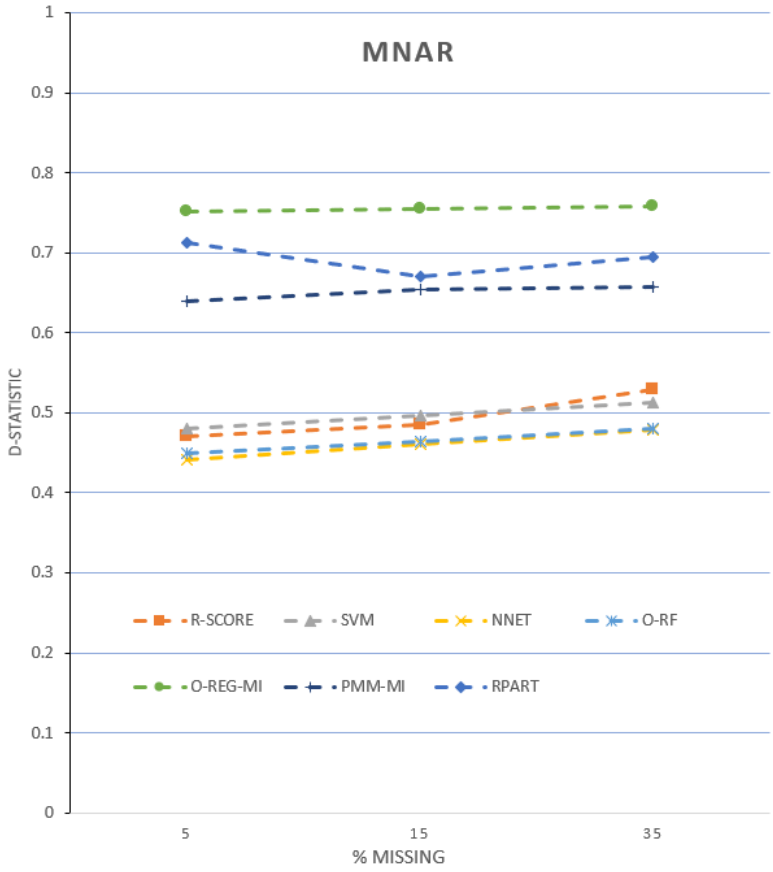


Figure 12. d-statistics of imputation models for total household income in Canadian Community Health Survey 2014 with induced MNAR by proportion of missing data.

Table 4. Proportions (%) of true CCHS2014 total household income value by its imputed value in induced MAR (mechanism 1: simple) with 35% missingness, imputed via O-RF

O-RF: MAR-1 (35%)	True					
Imputed	<20k	20k-39k	40k-59k	60k-79k	>80k	Sum
<20k	38.4	11.1	3.1	1.4	0.4	1381
20k-39k	**44.6**	**55.3**	35.5	20.6	7.0	5283
40k-59k	3.4	9.3	13.5	11.1	4.4	1467
60k-79k	0.2	0.5	1.6	1.8	0.7	171
>80k	13.2	23.9	**46.2**	**65.0**	**87.5**	10020
Sum	2009	3978	3377	2635	6323	18322

Note. Shaded cells indicate the diagonal (i.e. accurately imputed values). Bold figures indicate the highest proportion true values in each income category

Table 5. Proportions (%) of true CCHS2014 total household income value by its imputed value in induced MAR (mechanism 2: mid-complex) with 35% missingness, imputed via O-RF

O-RF: MAR-2 (35%)	True					
Imputed	<20k	20k-39k	40k-59k	60k-79k	>80k	Sum
<20k	40.7	12.1	3.5	1.5	0.4	1507
20k-39k	**44.1**	**55.7**	36.1	21.1	7.7	5376
40k-59k	3.5	9.1	14.7	11.4	4.5	1495
60k-79k	0.3	0.5	1.6	1.3	0.9	166
>80k	11.3	22.6	**44.1**	**64.6**	**86.6**	9495
Sum	2066	4040	3331	2596	6006	18039

Note. Shaded cells indicate the diagonal (i.e. accurately imputed values). Bold figures indicate the highest proportion true values in each income category

Table 6. Proportions (%) of true CCHS2014 total household income value by its imputed value in induced MAR (mechanism 3: complex) with 35% missingness, imputed via O-RF

O-RF: MAR-3 (35%)	True					
Imputed	<20k	20k-39k	40k-59k	60k-79k	>80k	Sum
<20k	38.4	10.9	3.6	1.8	0.4	1415
20k-39k	**44.4**	**55.3**	35.1	20.7	6.7	5253
40k-59k	4.1	9.4	14.6	11.1	4.6	1522
60k-79k	0.4	0.6	1.3	1.8	0.8	179
>80k	12.7	23.9	**45.5**	**64.6**	**87.4**	9824
Sum	2046	4005	3366	2604	6172	18193

Note. Shaded cells indicate the diagonal (i.e. accurately imputed values). Bold figures indicate the highest proportion true values in each income category

Table 7. Proportions (%) of true CCHS2014 total household income value by its imputed value in induced MAR (mechanism 1: simple) with 35% missingness, imputed via PMM-MI

PMM-MI: MAR-1 (35%)	True					
Imputed	<20k	20k-39k	40k-59k	60k-79k	>80k	Sum
<20k	19.2	18.1	12.3	9.1	5.8	2132
20k-39k	**29.5**	**25.4**	20.1	18.1	13.6	3618
40k-59k	20.1	19.4	21.4	20.0	17.6	3538
60k-79k	10.8	12.4	13.3	14.1	15.4	2505
>80k	20.5	24.6	**32.9**	**38.6**	**47.6**	6529
Sum	2009	3978	3377	2635	6323	18322

Note. Shaded cells indicate the diagonal (i.e. accurately imputed values). Bold figures indicate the highest proportion true values in each income category

Table 8. Proportions (%) of true CCHS2014 total household income value by its imputed value in induced MAR (mechanism 2: mid-complex) with 35% missingness, imputed via PMM-MI

PMM-MI: MAR-2 (35%)	True					
Imputed	<20k	20k-39k	40k-59k	60k-79k	>80k	Sum
<20k	20.3	16.8	12.2	9.8	6.7	2162
20k-39k	**32.9**	**29.9**	24.4	20.3	15.4	4156
40k-59k	18.5	18.7	20.3	18.0	17.1	3309
60k-79k	10.4	12.9	13.7	15.3	15.3	2511
>80k	17.9	21.6	**29.5**	**36.6**	**45.4**	5901
Sum	2066	4040	3331	2596	6006	18039

Note. Shaded cells indicate the diagonal (i.e. accurately imputed values). Bold figures indicate the highest proportion true values in each income category

Table 9. Proportions (%) of true CCHS2014 total household income value by its imputed value in induced MAR (mechanism 3: complex) with 35% missingness, imputed via PMM-MI

PMM-MI: MAR-3 (35%)	True					
Imputed	<20k	20k-39k	40k-59k	60k-79k	>80k	Sum
<20k	25.4	16.9	9.9	7.4	3.6	1942
20k-39k	**36.4**	**32.1**	24.1	19.2	11.2	4035
40k-59k	17.6	19.8	21.0	17.4	16.0	3304
60k-79k	8.6	11.7	14.8	16.2	16.0	2554
>80k	12.0	19.5	**30.2**	**39.8**	**53.1**	6358
Sum	2046	4005	3366	2604	6172	18193

Note. Shaded cells indicate the diagonal (i.e. accurately imputed values). Bold figures indicate the highest proportion true values in each income category

Table 10. Proportions (%) of true CCHS2014 total household income value by its imputed value in induced MNAR with 5% missingness, imputed via O-RF

O-RF: MNAR (5%)	True					
Imputed	<20k	20k-39k	40k-59k	60k-79k	>80k	Sum
<20k	34.9	9.9	2.0	1.4	0.4	260
20k-39k	**47.9**	**55.8**	30.9	20.8	6.0	767
40k-59k	4.4	10.2	17.1	11.1	5.7	205
60k-79k	0.4	0.3	2.0	2.4	1.5	29
>80k	12.5	23.8	**48.0**	**64.3**	**86.4**	1309
Sum	545	588	246	207	984	2570

Note. Shaded cells indicate the diagonal (i.e. accurately imputed values). Bold figures indicate the highest proportion true values in each income category

Table 11. Proportions (%) of true CCHS2014 total household income value by its imputed value in induced MNAR with 15% missingness, imputed via O-RF

O-RF: MNAR (15%)	True					
Imputed	<20k	20k-39k	40k-59k	60k-79k	>80k	Sum
<20k	27.7	6.7	1.3	1.5	0.1	585
20k-39k	**52.6**	**55.9**	34.5	21.2	6.3	2508
40k-59k	6.5	14.7	19.9	14.6	6.4	855
60k-79k	0.3	0.7	2.2	3.1	1.3	100
>80k	12.8	22.1	**42.1**	**59.6**	**85.9**	4133
Sum	1573	1840	869	721	3178	8181

Note. Shaded cells indicate the diagonal (i.e. accurately imputed values). Bold figures indicate the highest proportion true values in each income category

Table 12. Proportions (%) of true CCHS2014 total household income value by its imputed value in induced MNAR with 35% missingness, imputed via O-RF

O-RF: MNAR (35%)	True					
Imputed	<20k	20k-39k	40k-59k	60k-79k	>80k	Sum
<20k	23.3	5.1	1.2	0.2	0.2	980
20k-39k	**55.2**	**53.3**	28.9	17.1	5.1	5507
40k-59k	11.3	21.0	29.1	22.3	10.2	3206
60k-79k	0.3	1.4	2.6	3.2	1.9	342
>80k	9.9	19.2	**38.3**	**57.2**	**82.7**	9526
Sum	3023	4544	2380	2004	7610	19561

Note. Shaded cells indicate the diagonal (i.e. accurately imputed values). Bold figures indicate the highest proportion true values in each income category

Table 13. Proportions (%) of true CCHS2014 total household income value by its imputed value in induced MNAR with 5% missingness, imputed via PMM-MI

PMM-MI: MNAR (5%)	True					
Imputed	<20k	20k-39k	40k-59k	60k-79k	>80k	Sum
<20k	22.6	15.1	10.2	4.3	3.9	284
20k-39k	**34.7**	**31.0**	21.5	19.8	10.1	564
40k-59k	19.1	21.6	19.5	18.4	15.2	467
60k-79k	10.1	11.2	16.3	15.5	15.9	349
>80k	13.6	21.1	**32.5**	**42.0**	**55.0**	906
Sum	545	588	246	207	984	2570

Note. Shaded cells indicate the diagonal (i.e. accurately imputed values). Bold figures indicate the highest proportion true values in each income category

Table 14. Proportions (%) of true CCHS2014 total household income value by its imputed value in induced MNAR with 15% missingness, imputed via PMM-MI

PMM-MI: MNAR (15%)	True					
Imputed	<20k	20k-39k	40k-59k	60k-79k	>80k	Sum
<20k	22.4	15.0	8.1	5.0	3.3	839
20k-39k	**34.8**	**30.4**	24.6	21.1	12.6	1875
40k-59k	19.2	21.7	22.9	21.4	17.4	1607
60k-79k	9.9	13.3	16.6	17.5	16.5	1194
>80k	13.7	19.6	**27.8**	**35.1**	**50.2**	2666
Sum	1573	1840	869	721	3178	8181

Note. Shaded cells indicate the diagonal (i.e. accurately imputed values). Bold figures indicate the highest proportion true values in each income category

Table 15. Proportions (%) of true CCHS2014 total household income value by its imputed value in induced MNAR with 35% missingness, imputed via PMM-MI

PMM-MI: MNAR (35%)	True					
Imputed	<20k	20k-39k	40k-59k	60k-79k	>80k	Sum
<20k	19.0	12.5	6.9	4.5	2.3	1570
20k-39k	**34.5**	**31.0**	22.4	19.2	10.6	4173
40k-59k	22.0	24.7	25.5	21.2	19.4	4300
60k-79k	11.7	14.4	18.6	18.2	18.4	3220
>80k	12.8	17.4	**26.6**	**37.0**	**49.3**	6298
Sum	3023	4544	2380	2004	7610	19561

Note. Shaded cells indicate the diagonal (i.e. accurately imputed values). Bold figures indicate the highest proportion true values in each income category