Author/s:
Nguyen, CD;Moreno-Betancur, M;Rodwell, L;Romaniuk, H;Carlin, JB;Lee, KJ

Title:
Multiple imputation of semi-continuous exposure variables that are categorized for analysis

Date:
2021-11-30

Citation:
Nguyen, C. D., Moreno-Betancur, M., Rodwell, L., Romaniuk, H., Carlin, J. B. & Lee, K. J. (2021). Multiple imputation of semi-continuous exposure variables that are categorized for analysis. Statistics in Medicine, 40 (27), pp.6093-6106. https://doi.org/10.1002/sim.9172.

Persistent Link:
https://hdl.handle.net/11343/298871

Multiple imputation of semi-continuous exposure variables that are categorised for analysis

Cattram D. Nguyen[1,2], Margarita Moreno-Betancur[1,2], Laura Rodwell[1,2,3], Helena Romaniuk[4], John B. Carlin[1,2], Katherine J. Lee[1,2]

1. Clinical Epidemiology and Biostatistics Unit, Murdoch Children's Research Institute, Parkville, Australia

2. Department of Paediatrics, Faculty of Medicine, Dentistry and Health Sciences, The University of Melbourne, Melbourne, Australia

3. Department for Health Evidence, Radboud Institute for Health Sciences, Radboud University Medical Center, Nijmegen, The Netherlands

3. Centre for Adolescent Health, The Royal Children's Hospital, Parkville, Australia

4. Deakin University, Faculty of Health, Biostatistics Unit, Geelong, Australia

Correspondence to: Cattram Nguyen, Clinical Epidemiology and Biostatistics Unit, Murdoch Children's Research Institute, Royal Children's Hospital, 50 Flemington Road, Parkville, Victoria, 3052, Australia

email: cattram.nguyen@mcri.edu.au

Formatted: Highlight

Abstract

Semi-continuous variables are characterised by a point mass at one value and a continuous range of values for remaining observations. An example is alcohol consumption quantity, with a spike of zeros representing non-drinkers and positive values for drinkers. If multiple imputation is used to handle missing values for semi-continuous variables, it is unclear how this should be implemented within the standard approaches of fully conditional specification (FCS) and multivariate normal imputation (MVNI). This question is brought into focus by the use of categorised versions of semi-continuous exposure variables in analyses (e.g. no drinking, drinking below binge level, binge drinking, heavy binge drinking), raising the question of how best to achieve congeniality between imputation and analysis models.

We performed a simulation study comparing nine approaches for imputing semi-continuous exposures requiring categorisation for analysis. Three methods imputed the categories directly: ordinal logistic regression, and imputation of binary indicator variables representing the categories using MVNI (with two variants). Six methods (predictive mean matching, zero-inflated binomial imputation and two-part imputation methods with variants in FCS and MVNI) imputed the semi-continuous variable, with categories derived after imputation.

The ordinal and zero-inflated binomial methods had good performance across most scenarios, while MVNI methods requiring rounding after imputation did not perform well. There were mixed results for predictive mean matching and the two-part methods, depending on whether the estimands were proportions or regression coefficients. The results highlight the need to consider the parameter of interest when selecting an imputation procedure.

243 words

Key words: missing data, multiple imputation, semi-continuous, ordinal categorical variable, zero inflated data

BACKGROUND

Epidemiological variables often exhibit semi-continuous distributions, which are characterised by a point mass at one value and a continuous range of values for the remaining observations. An example of this is alcohol consumption quantity, where non-drinkers are assigned zero alcohol units and drinkers have a positive range of units. Other examples include hours spent exercising, number of cigarettes smoked, viral loads and duration of breastfeeding. When semi-continuous variables are integer counts with a point mass at zero, then they can also be considered zero-inflated count variables. When used in analyses as the exposure of interest, semi-continuous variables are often categorised for analysis due to their non-symmetric distributions, or to create classifications that are relevant to policy-making.[1,2] For example, the number of alcohol units can be categorised to create an ordinal variable that identifies different levels of drinking (e.g. no drinking, below binge drinking, binge drinking, and heavy binge drinking).[3] As with most variables used in epidemiological research, semi-continuous variables can be subject to missing data and it is unclear how the missing values for this type of measure should be handled in analyses.

One popular statistical approach for handling missing data is multiple imputation (MI).[4] To use MI, an imputation procedure is first specified to impute values for the variables with missing data. This is repeated multiple times to create multiple completed datasets, reflecting the uncertainty surrounding the missing values. Each completed dataset is then analysed using standard methods, with the estimates obtained from each of these analyses pooled using 'Rubin's rules' to obtain overall estimates and inferences for the parameters of interest.[5]

There are currently two standard imputation approaches for imputing multiple incomplete variables: multivariate normal imputation (MVNI)[6] and fully conditional specification (FCS).[7] MVNI assumes that the variables in the imputation model follow a joint normal distribution. Several studies have demonstrated that MVNI works well even if some of the imputed variables are not normally distributed,[6,8] with rounding methods or latent normal variable specifications used to handle categorical variables.[9] Under FCS, imputation is performed using a series of univariate imputation models, one for each variable with missing data. A range of univariate models are available, including logistic regression for the imputation of a binary variable, linear regression for a continuous variable, and ordinal logistic regression for an ordinal categorical variable.

When there are missing values in semi-continuous variables that are categorised for analysis, the missing values can either be imputed at the raw data level (i.e. in the semi-continuous variable, with the ordinal variable derived following imputation), or the ordinal variable could be imputed directly.

A number of approaches are available for imputing semi-continuous variables at the raw data level. Schafer and Olsen proposed a two-part method that first uses a logistic regression model to impute the binary component of the semi-continuous variable (i.e. drinker/non-drinker).[10] Then, a linear regression model is specified to impute the (usually log-transformed) continuous component (i.e. number of units of alcohol consumed). Because of the two-step conditional nature of this approach, it is suited to the FCS framework. However, it is also possible to implement the two-part method using MVNI. This involves imputing the binary and continuous components as separate variables, and obtaining the binary imputed values either via post-imputation rounding or through a latent normal variable specification. If the semi-continuous data are counts with a large proportion of zero values, then another option is to impute the missing values using methods for handling zero-inflated count data, e.g. zero-inflated Poisson models or zero-inflated negative binomial models when there is overdispersion (i.e. where the variance of the data is greater than the mean).[11] ~~Similar to the two-part methods, the~~The zero-inflated imputation approaches are mixture modelling methods consisting of a model for the zeros and a model for the counts (e.g. Poisson or negative binomial).[12] There are also dedicated methods for imputing semi-continuous variables, such as the blocked general location model developed by Javaras and van Dyk.[13] However, we do not consider this method in the current study, as it is not readily accessible to users in practice.

It is also possible to impute semi-continuous variables using general-purpose methods for continuous variables. One method that can be implemented within the FCS framework is predictive mean matching (PMM).[14,15] PMM replaces the missing value with an observed value 'borrowed' from a donor (or pool of donors) with the closest predicted value from a linear regression model. The appeal of PMM is that the observed range and distribution of the variable are preserved, because only values that have been observed are used for the imputation. Another option is to impute the semi-continuous variable using simple linear regression assuming normality. This method was considered by Yu et al.[16] and resulted in a large number of negative values being imputed, which were replaced with zero. Yu et al.[16]

reported that this method performed poorly, and we therefore do not consider this method in the current study.

Alternatively, the missing values can be imputed directly on the ordinal scale required for analysis. One option is to impute the ordinal variable using an ordinal logistic regression model, which can be used within FCS in a general multivariable missingness setting.[7,17] Another option is to create a set of indicator variables that represent the categories of the ordinal variable, impute these indicators using MVNI. To obtain binary (rather than continuous) imputed values, the indicators can either be imputed using a latent normal specification for the binary indicators,[18] or a rounding method can be applied after imputation to allocate each observation to one of the ordinal categories. Studies that have compared methods of rounding have identified the method of projected distance-based rounding proposed by Allison[19] to be the best method for rounding indicator variables.[20,21] Another approach is to treat the ordinal values as continuous and impute them using a linear regression imputation model, followed by rounding of the imputed values. However, we do not consider this method here, because the values assigned to ordinal categories (and the distance between them) may not be meaningful.

A few studies have evaluated methods for imputing semi-continuous variables.[10,16,22] Schafer and Olsen[10] performed a simulation study to evaluate their two-part method. They found the two-part method had good coverage and low bias for estimating the proportion in the point mass, correlation coefficients and regression coefficients, but there was moderate bias and over-coverage in estimates of log-odds ratios. Vink et al.[22] compared PMM with dedicated methods for semi-continuous data including the two-part method and the blocked general location model. They found that PMM had the best performance across a number of scenarios; it estimated the size of the point mass accurately and preserved correlations and distributional shapes. Yu et al.[16] compared a number of routines in available software packages where semi-continuous variables were imputed either using PMM or a normal imputation model (either replacing negative values with zero or pre-specifying a boundary to avoid out of bound values). They found the PMM methods performed well and retained underlying distributions of semi-continuous variables, while methods assuming normality led to biased results and poor coverage. Kleinke and Reinecke[12] compared a number of methods for imputing zero-inflated count data (e.g. Poisson, quasi-Poisson, zero-inflated Poisson and zero-inflated negative binomial multiple imputation). Estimates were biased or had poor coverage when a restrictive model was used (e.g. if there were excess zeros then these needed

to be accommodated using a zero-inflated model). However, there was no harm in using a more general model (e.g. a zero-inflated negative binomial imputation model performed well even in the absence of overdispersion).

~~previous studies have not evaluated available methods for imputing semi-continuous variables that are used as categorical exposure variables in regression analyses, and they have not examined methods that directly impute the categorised version that is required for analysis. In terms of congeniality, it is generally recommended that the imputation model should align with the analysis model, which includes imputing variables in the same form as they appear in the analysis model.[17,23,24] Therefore, the methods that directly impute the categorised version may potentially have better performance due to the resulting congeniality of the imputation and the analysis models.[24,25]~~ A few studies have examined methods for imputing continuous data when the target analysis is performed on the categorised variable.[26-29] ~~This is supported by previous research examining methods for imputing continuous data that are categorised for analysis.~~ In two simulation studies, Demirtas ~~performed a simulation study that~~ compared two ~~imputation~~ approaches for imputing continuous data that were categorised for analysis : i) impute under multivariate normality followed by categorisation for analysis, and ii) categorise then impute under a log-linear model.[26,27] In both studies, c~~C~~ategorisation followed by imputation was the preferred strategy across most simulation scenarios, suggesting that it may be preferable to impute data in the form in which they will be analysed. In contrast, in their simulation study, Floden and Bell found imputing before dichotomisation had similar performance to dichotomisation then imputation.[28] However, with large amounts of missing data (50%), imputing the continuous variable before dichotomisation produced less biased results. Previous studies have not evaluated available methods for imputing semi-continuous variables that are used as categorical exposure variables in regression analyses, and they have not examined methods that directly impute the categorised version that is required for analysis. In terms of congeniality, it is generally recommended that the imputation model should align with the analysis model, which includes imputing variables in the same form as they appear in the analysis model.[17,23,24] Therefore, the methods that directly impute the categorised version may potentially have better performance due to the resulting congeniality of the imputation and the analysis models;[24,25] however, this is unclear given mixed results of previous research for imputing dichotomised variables.[26,28]

The aim of the current study is to compare the performance of the readily available methods described above for imputing semi-continuous exposure variables that are categorised for analysis: PMM for the semi-continuous variable, two-part imputation of the semi-continuous variable (with two variants each in FCS and MVNI), zero-inflated negative binomial imputation of the semi-continuous variable, ordinal logistic regression for the categorised variable, and imputation of indicator variables for the categorised variable using MVNI (using either rounding or a latent normal specification to obtain binary imputed values). To compare the imputation methods, we simulated data that broadly reflect the patterns of drinking behaviour observed within the Victorian Adolescent Health Cohort Study (VAHCS). We also illustrate the methods using a case study from the VAHCS.

METHODS

Introduction to case study: The Victorian Adolescent Health Cohort Study

The VAHCS is a longitudinal study of 1943 young people who were recruited as adolescents through schools in Victoria, Australia, which was initiated in August 1992.[30,31] Participants were recruited when they were 14-15 years old at one of two entry points (waves 1 and 2) which were six months apart. Participants were followed up on four more occasions at 6-monthly intervals during adolescence (waves 3 to 6), and at four points in adulthood (waves 7 to 10) when the participants were on average 21, 24, 29 and 35 years of age, respectively. At each of the adolescent waves, participants completed a retrospective alcohol use diary, from which the number of alcohol units they consumed on each day/occasion in the previous week was calculated. In adulthood, alcohol consumption was measured using a beverage and quantity-specific four-day diary that included all weekend days and the most recent weekday.[31]

For the case study, we focused on males in late adolescence (mean age 17 years, wave 6) and considered a four-level categorical variable that was derived from the units of alcohol consumed on the day in the diary week with the highest alcohol consumption. The categories were: no drinking (0 units); drinking below binge level (1 to less than 5 units); binge drinking (5 to less than 11 units) and heavy binge drinking (11 or more units). Our illustrative analysis focuses on estimating the proportions in each of the four drinking categories, and the association of adolescent drinking behaviour (wave 6) with subsequent binge drinking (defined as 5 or more units on any one day) in young adulthood (mean age 21 years, wave 7). Estimating the latter required adjustment for peer drinking behaviour in adolescence, a

potentially confounding variable that was reported by participants at each wave. The confounder was incorporated as a binary variable, which was equal to one if most of the participant's peers drank at any time during adolescence (waves 2 to 6), and was equal to zero otherwise. The analysis of interest used a logistic regression model:

$$\text{logit Pr}(Binge = 1) = \alpha_0 + \alpha_1 Alcohol_2 + \alpha_2 Alcohol_3 + \alpha_3 Alcohol_4 + \alpha_4 Peer \quad (1)$$

where $Binge$ is an indicator for binge drinking in adulthood (21 years) and $Peer$ is an indicator for whether most peers drank at any wave in adolescence. $Alcohol_2$, $Alcohol_3$ and $Alcohol_4$ are indicators for drinking below binge level, binge drinking and heavy binge drinking in adolescence (17 years), respectively. At each wave in adolescence, participants also reported the number of cigarettes smoked each day during the last week. These variables were used to derive a binary auxiliary variable (i.e. a variable used to improve the imputation procedure) indicating whether the participants had smoked daily at any wave.

The analysis included all males for whom the outcome was observed at wave 7, resulting in a sample of 725 participants (thus leaving aside for this purpose the potential for selection bias due to loss to follow-up). For the purpose of the case study, the confounder (peer drinking at any wave) and auxiliary variable (daily smoking at any wave) were derived for all participants (regardless of missing data at individual waves), and were therefore complete for all 725 participants. The only incomplete variable in this example was adolescent alcohol consumption, which was missing for 18% of participants.

Multiple imputation methods

The following two methods can be used to impute the categorical exposure variable directly:

i) *"Ordinal"*: The ordinal categorical variable was imputed using an ordinal logistic regression imputation model. This was implemented as a single univariate imputation model, but it can also be carried out within the FCS framework.

ii) *"MVNI indicator - rounding"*: For a 4-level ordinal variable, three binary indicator variables were generated to represent the categories of the ordinal variable (excluding the reference category). The missing indicators were imputed using MVNI, and the resulting (continuous) imputed values were used to allocate the record to one of the four categories using projected distance based rounding [19]. Specifically, an imputed value for the reference category indicator was first

calculated by subtracting the imputed values for the other three categories from 1. The record being imputed was then allocated to the category with the largest imputed value across all categories (including the reference category) in the completed dataset [19,32].

iii) *"MVNI indicator – latent"*: This is a variant of method (ii) where instead of using post-imputation rounding, the binary indicators are represented by latent normal variables in the imputation model.[18] The latent variables are modelled jointly under the assumption of multivariate normality.[33]

The following imputation methods can be used to impute the semi-continuous variable, followed by derivation of the ordinal variable:

iv) *"Predictive mean matching" (PMM):* The semi-continuous variable was imputed using PMM using the Type I matching algorithm recommended by Morris et al.[15]. We specified the imputed value to be sampled by random selection from the 5 nearest (observed) neighbours to the linear prediction for the missing value.

v) *"Two-part FCS – conditional"*: This two-part imputation method involved two univariate imputation models within the FCS procedure [7]. A logistic regression model was first specified to impute a binary variable that represented whether the semi-continuous variable had a zero or positive value. For records with a positive value according to the imputed binary variable, the continuous value was then imputed on a log-transfomed scale using a linear regression model. The continuous values were transformed back to the original scale following FCS.

vi) *"Two-part FCS – just another variable" (JAV):* This method is a variant of method (v), also using logistic regression to impute the binary values, and linear regression to impute the log-transformed continuous values using FCS. However, unlike method (v), this approach ignores the relationship between the binary and continuous components, and imputes the continuous values as "just another variable". After the FCS procedure, the semi-continuous variable is imputed as either zero (if the imputed binary variable is zero) or the back-transformed continuous values (if the imputed binary variable is one, ignoring the imputed continuous value when the binary variable is zero).

vii) *"Two-part MVNI – rounding"*: Similar to the other two-part methods (v and vi), under this approach the semi-continuous variable is represented as two variables: a) a binary indicator for whether the values are zero or positive, and b) the log-

transformed continuous values. These two variables are then imputed as separate variables using MVNI. As the imputed values for the binary component are continuous, the imputed values are rounded to 0 or 1 using an adaptive rounding method, where the threshold for rounding is based on a normal approximation to the binomial distribution.[34] In the final step, imputed values for the semi-continuous variable are obtained by setting the missing values to zero (if the binary indicator is equal to zero) or the back-transformed continuous values (if the binary indicator is equal to one), as for method vi.

viii) *"Two-part MVNI – latent":* This method is a variant of method (vii), where instead of using adaptive rounding to handle the binary variable in the MVNI imputation, a latent normal specification is used.[35] This approach assumes that the binary variable is linked to a latent normal variable through a probit model. This underlying normal variable is modelled jointly with other incomplete variables (in this case the continuous component of the semi-continuous exposure) under the MVNI assumption of multivariate normality, and then the probit transformation is used to yield imputations of the binary variable.[33]

ix) *"Zero-inflated negative binomial (ZINB):* This imputation approach uses a mixture model, whereby a logit model is used to impute a binary variable that represents whether the semi-continuous variable has a zero or positive value, and a negative binomial model with a log link is used to impute the positive counts. ~~Note.~~ We used a zero-inflated negative binomial model rather than a zero-inflated Poisson model based on findings that a more general, rather than restrictive, modelling approach is recommended (where the negative binomial model is considered more general as it estimates an additional parameter that allows for overdispersion).[12]

Application to case study

We applied these methods to the case study. The imputation models included all analysis model variables (i.e. variables in Equation 1), as well as a binary indicator for daily smoking at any adolescent wave, which was included as an auxiliary variable. For each method, 20 imputed datasets were generated and Rubin's rules[5] were used to combine the results of the logistic regression analysis model, as well as the estimates of the marginal proportions in each drinking behaviour category. For comparison, a complete case analysis (i.e. analysis

only including participants with completely observed data) was also performed. See Results section.

Simulation study

The simulation study was designed to represent realistic scenarios based on the VAHCS case study. Further details on the choice of parameter values for the simulation study are provided in Supplementary Table 1.

Data generation

The simulated datasets included a semi-continuous exposure variable (X_1), a binary outcome variable (Y) and a binary confounding variable (X_2). In these simulations, we assumed that the exposure variable X_1 was truly semi-continuous, but was related to the outcome Y as a categorised variable. A continuous auxiliary variable (X_3) was also generated for inclusion in the imputation models to provide additional information for the prediction of the missing data. We simulated data for three scenarios where 20%, 40% and 60% of values were in the point mass of the semi-continuous variable (with a value of zero). The steps for the generation of the data were as follows:

Step 1: A confounder variable (X_2) and an auxiliary variable (X_3) were simulated as: $X_2 \sim Binomial(1,0.7)$ and $X_3 \sim N(0,1)$, respectively.

Step 2: A semi-continuous variable (X_1) was simulated by first generating a binary indicator, U, as a function of the confounder and auxiliary variables using:

$$\text{logit}\, Pr(U = 1) = \beta_0 + \beta_1 X_2 + \beta_2 X_3 \quad (2)$$

with the value for β_0 (given in Supplementary Table 1) altered to control the percentage of zeros (i.e. 20%, 40% or 60%), $\beta_1 = 1.25$ (odds ratio $(\text{OR}) = 3.5$), and $\beta_2 = 0.41$ ($\text{OR} = 1.5$).

A continuous variable (V), representing the positive values for X_1, was then generated from a Poisson regression model dependent on the confounder and auxiliary variable using:

$$V \sim Poisson(\mu) \quad (3)$$

$$\log(\mu) = \gamma_0 + \gamma_1 X_2 + \gamma_2 X_3 \quad (4)$$

where $\gamma_1 = 0.5,\ \gamma_2 = 0.3$, and γ_0 was varied to control the median of the continuous component of X_1 (see Supplementary Table 1 for details).

The semi-continuous variable, X_1, was then obtained by setting $X_1 = 0$ if $U = 0$, and $X_1 = V$ if $U = 1$.

Step 3: A four-level categorical variable, *C*, was derived from the semi-continuous variable, X_1, where:

$$C = \begin{cases} 0 \text{ if } X_1 = 0 \\ 1 \text{ if } X_1 > 0 \text{ and } < 6 \\ 2 \text{ if } X_1 \geq 6 \text{ and } < 9 \\ 3 \text{ if } X_1 \geq 9 \end{cases} \quad (5)$$

We then generated indicator variables for the non-reference (non-zero) categories defined in equation (5), i.e. $I_1 = I(C = 1)$, $I_2 = I(C = 2)$, and $I_3 = I(C = 3)$ where $I()$ is the indicator function.

Step 4: A binary outcome variable, *Y*, was generated as a function of the indicator variables (I_1, I_2, I_3), and the confounder variable, using the following logistic regression model:

$$\text{logit Pr}(Y = 1) = \delta_0 + \delta_1 I_1 + \delta_2 I_2 + \delta_3 I_3 + \delta_4 X_2 \quad (6)$$

where $\delta_0 = -0.83,\ \delta_1 = 0.69$ (OR = 2), $\delta_2 = 0.92$ (OR = 2.5), $\delta_3 = 1.39$ (OR = 4), and $\delta_4 = 0.92$ (OR = 2.5).

Missing data

Missing data were imposed on the semi-continuous variable (X_1) using two different mechanisms. Firstly, data were set to be missing completely at random (MCAR), whereby a random sample of 30% of values of X_1 were set to missing. Secondly, data were set to be missing at random (MAR), where the missingness in X_1 was dependent on the outcome, confounder and auxiliary variables, using the logistic regression model:

$$\text{logit Pr}(X_1\ missing) = \zeta_0 + \zeta_1 Y + \zeta_2 X_2 + \zeta_3 X_3 \quad (7)$$

where $\zeta_1 = 0.69$ (OR = 2), $\zeta_2 = -0.69;$ OR = 0.5), and $\zeta_3 = 0.69;$ OR = 2). The value of ζ_0 was controlled so that approximately 30% of observations were missing.

For each semi-continuous exposure variable scenario (20%, 40% and 60% zeros) and the two types of missingness (MCAR and MAR), 2000 datasets of 1000 observations were generated. The sample size of 1000 observations per dataset was chosen to be a realistic sample size for

a cohort study and was motivated by the VAHCS (which recruited n=1943 in total, with n=725 included in the case study analysis). A simulation sample size of 2000 replications was chosen to produce a standard error of 0.5% for a coverage of 95%.[36]

Multiple imputation methods

The missing values were imputed using the nine approaches described previously. For each MI analysis, 50 imputations were generated and results of target analyses combined using Rubin's rules[5] (with no transformation of estimates of regression coefficients and proportions prior to combination).[17] A larger number of imputations was used for the simulations compared to the case study as there was a larger proportion of missing data.

Target analyses and evaluation of performance

The parameters of interest were the marginal proportions in each of the 4 categories of the semi-continuous exposure variable, and the coefficients from the logistic regression for the binary outcome on the exposure indicators (adjusted for the confounder) (equation 6). These parameters were estimated using each of the nine imputation methods, along with a complete case analysis for comparison. For the marginal proportions, a pseudo-population of one million complete observations was generated for each scenario in order to obtain 'true' population values of the proportions. For the logistic regression coefficients, the 'true' values were the parameter values specified in the data generating model (equation 6).

Four measures of performance were considered for the evaluation of the methods:

i. Bias: the difference between the average of the estimates (over the 2000 replications) and the 'true' value.
ii. Empirical standard error (SE): the standard deviation of the point estimates over the 2000 datasets.
iii. Model-based SE: the average of the estimated standard errors over the 2000 replications. If an imputation procedure is performing well, the average model-based SE should be similar to the empirical SE.
iv. Coverage: the proportion of 95% confidence intervals across the 2000 replications that contain the true value.

Software

For both the case study analysis and the simulation study, the MVNI indicator-latent and two-part MVNI – latent methods were implemented in R 3.5 using the 'jomo' package.[9,37] The ZINB method was implemented using the 'countimp' package, which is an add-on for the 'mice' package in R.[38,39] The remaining analyses were conducted using Stata 15.1[40] using the mi impute command, with the exception of PMM, which was performed using the ice[41] add-on program for Stata to enable implementation of the Type I matching algorithm.[15] To avoid perfect prediction, the 'augment' option was specified when implementing the ordinal and two-part FCS methods in Stata.[40,42] Performance measures and Monte Carlo errors were calculated using the 'simsum' package in Stata.[40,43]

RESULTS

Simulation study

Marginal proportions Results for the MAR scenarios for the marginal proportions are shown in Figure 1 and Supplementary Table 3. Estimates from the complete case analysis were biased as expected, particularly for estimates of proportions in the "no drinking" and "heavy binge" categories. PMM and ZINB generally had the lowest bias, while the MVNI-indicator and two-part MVNI rounding methods had the largest bias and largest standard errors for several scenarios. In general, the two-part methods estimated the size of the point mass well, but estimates of the proportions in the other categories were biased. In addition, the two-part imputation methods produced the largest inconsistencies between empirical and model-based SEs, with empirical SEs generally being smaller than model-based SEs. No methods consistently had coverage close to nominal levels, with poor coverage for the MVNI indicator and two-part MVNI rounding methods for several scenarios. PMM was the best performing method with respect to coverage despite slight under-coverage for some of the proportions for some scenarios.

Results for MCAR scenarios for the marginal proportions are provided in Supplementary Table 2. As expected, under MCAR there was negligible bias in the results for the complete case analysis. The ZINB and PMM methods had the best performance with respect to bias. The two-part methods produced the largest biases, particularly for estimates of proportions in the binge and heavy binge categories. In terms of standard errors, the two-part MVNI rounding and MVNI indicator methods generally produced larger SEs than the other methods. For the MVNI indicator method, the model-based SEs were generally smaller than empirical SEs, leading to under-coverage. Coverage was close to nominal values for the

ordinal imputation method. Taken together, the PMM, ZINB and ordinal methods appeared to be the best performing imputation methods for estimating marginal proportions under MCAR.

Logistic regression coefficients

Results for the logistic regression coefficients are shown in Figure 2 and Supplementary Tables 5 for MAR. No imputation method consistently had the best performance across all scenarios. The ordinal method produced the smallest bias across many scenarios, and the ZINB method also performed well, but both methods had over-coverage of the 95% confidence intervals. PMM had the worst performance in terms of bias, followed by the MVNI indicator-rounding and two-part MVNI rounding methods. There were inconsistencies between model-based and empirical SEs for most imputation methods, and no methods consistently achieved nominal coverage levels, with PMM having the worst levels of under-coverage. There was a similar pattern of results for the MCAR scenario, with smaller levels of bias than under MAR (see Supplementary Table 4).

When comparing the methods that required post-imputation rounding (MVNI indicator-rounding and two-part MVNI-rounding) with the respective methods that used a latent variable approach (MVNI indicator-latent and two-part MVNI-latent), the rounding methods generally produced more biased estimates of regression coefficients.

For the MCAR scenario with 20% zeros, the imputation methods did not converge for a small number of replications (n=3 for the MVNI indicator method; n=2 for the ordinal, PMM, 2-part FCS conditional, 2-part FCS JAV and 2-part MVNI round methods). These problems occurred in simulated datasets where all cases with $X_1 = 2$ had the same outcome ($Y = 1$). There were no convergence problems for the MVNI indicator-latent, 2-part MVNI latent and ZINB methods, but the model-based standard errors were very large for the same three replications; we therefore omitted these repetitions from the results (Supplementary Tables 2 and 4).

Case study

Figures 3 and 4 present the estimates of proportions and associations respectively when a complete case analysis and the nine imputation methods were applied to the case study. The complete case analysis and nine imputation methods produced similar results for the

estimates of the proportions in each of the alcohol consumption categories (Figure 3) s. There was some variability in the estimates of associations across methods (Figure 4), but these differences would not change the substantive conclusions.

DISCUSSION

This paper compared methods for the imputation of a semi-continuous variable in a context where it is analysed as a categorical exposure variable as is commonly done in epidemiological analyses.[44-47] We examined three methods that imputed ordinal categories directly, and six methods that imputed the semi-continuous variable followed by categorisation. Across the simulation scenarios, we found that the ordinal logistic regression and ZINB imputation approaches performed well across most scenarios.

We hypothesised that imputing the ordinal categories directly would be the preferred strategy, as the imputation model would be congenial with the substantive analysis that involved an ordinal exposure variable. However, there are also arguments for imputing the semi-continuous variable followed by categorisation as there may be advantages in retaining information about the semi-continuous distribution when imputing missing values. In this study, we did not find a preferred strategy with respect to imputing before or after categorisation, as we found both good and poor performing methods within these two strategies, with possible reasons discussed below.

Previous studies have recommended the use of PMM for imputing semi-continuous variables.[16,22] Consistent with the published research, we found that PMM performed well for estimates of marginal proportions, particularly for the MAR scenarios. However, PMM produced the largest bias for estimates of logistic regression coefficients. This highlights the importance of considering the parameter of interest when selecting an imputation method. PMM may be preferable for preserving features of marginal distributions, but less so for measures of association. Of the two-part methods, the worst performing approach was MVNI with post-imputation rounding, which produced the most biased results, as well as the largest standard errors for estimates of proportions. The remaining two-part methods (two-part FCS conditional, two-part FCS JAV and two-part MVNI latent) had similar performance across most scenarios examined; they generally performed well for the estimates of associations, but there was under-coverage and bias in estimates of proportions. Of the methods that imputed the semi-continuous variable followed by categorisation, ZINB was the best performing

method~~The ZINB approach could also be considered a two-part model, but unlike the other two-part approaches it used a negative binomial model rather than a normal model to impute the continuous values~~. However, ~~I~~it is possible that our simulation design favoured the ZINB approach, as we used a count model (i.e. Poisson regression) in the data generating process for the semi-continuous variable.

On the whole, the simulation study does not support the use of the MVNI indicator approach. In particular, we observed biased estimates of associations for the MVNI indicator method. This may be due to the need to round the indicator variables following imputation, which can introduce bias.[32] These findings are consistent with other studies that caution against the rounding of imputed values. [48-50] However, the use of rounding methods is contentious and there are studies in favour of such approaches,[6] and we also had simulation scenarios where the latent variable approach had worse performance than the rounding methods.

A limitation of this study was that performance of the imputation methods was examined using a simulation study with a limited number of scenarios, ~~including only one sample size (n=1000) and~~ with only one variable with missing data. Additionally, we only evaluated the performance of these methods for a relatively large sample size (n=1000), and therefore the results may not generalise to smaller samples. In addition, our simulation study only used one set of cutpoints for categorising the semi-continuous data, which was chosen to produce the marginal proportions observed in the VAHCS case study. Although the choice of cutpoints may have influenced results, we note that a previous simulation study of imputation methods for continuous data that were ordinalised for analysis found no systematic trends across the cutpoints examined.[27] Another limitation is that we only compared methods that were available in mainstream statistical packages and did not, for example, explore the blocked general location method that was not easily accessible by users. We also did not ~~include~~ focus on other approaches for handling incomplete zero-inflated count data, such as hurdle models.[38] However, in additional simulations we found that the zero-inflated and hurdle negative binomial methods had nearly identical performance (results ~~not shown~~presented in Supplementary tables 2-5). A strength of this study is that we endeavoured to design the simulation study to replicate the features of a real cohort study. As with all simulation studies, the simplified conditions examined may not accurately represent the complexity of real observational data. However, it is important to identify which methods do not perform well in a simplified setting before considering more complex scenarios, such as extensions to analyses with more than one incomplete variable.

Although this paper used alcohol consumption as a motivating example, researchers in other epidemiological contexts encounter variables that follow similar semi-continuous distributions, such as the number of cigarettes smoked, days of hospital admission, years in a relationship, and hours of TV watching, and would face similar decisions when using MI to handle missing data in these variables.

CONCLUSION

We considered the challenges of imputing semi-continuous variables in the context where they are used as ordinal exposure variables in the analysis. Based on our simulations, we found that the ordinal and ZINB methods generally performed well across many scenarios. PMM had good performance for marginal proportions, but not for estimates of regression coefficients. In general, the MVNI methods that required rounding of the imputed values did not perform well. There were mixed results for the two-part FCS and MVNI-indicator latent methods: they generally performed well in estimating regression coefficients, but not as well for marginal proportions. Future research is required to replicate these findings, and in particular to investigate these methods across a range of more complex scenarios.

AUTHOR CONTRIBUTIONS

All authors participated in the planning of the manuscript and interpretation of the results. LR, CDN, and MMB wrote the analysis programs. CDN and MMB performed the simulations and case study analyses. LR wrote an early version of the paper and CDN led the writing of the current paper. All authors read and contributed to the final manuscript.

DATA AVAILABILITY STATEMENT

Data from the Victorian Adolescent Health Cohort Study (VAHCS) are not publicly available but those interested in replicating these findings are welcome to contact the corresponding author, or the VAHCS team (https://www.mcri.edu.au/research/projects/2000-stories/information-researchers).

ACKNOWLEDGEMENTS

This paper used unit record data from the Victorian Adolescent Health Cohort Study. For this we thank the families that participated in the VAHCS, the study research team and the Principal Investigator, Professor George Patton. This work was supported by funding from the Australian National Health and Medical Research Council: Career Development

Fellowship ID 1120571 (KJL), Project Grant ID 1127984 (KJL, JBC), Project Grant ID 1166023 (KJL, MMB, JBC, CDN) and a Centre of Research Excellence grant ID 1035261 (JBC), which funded the Victorian Centre for Biostatistics (ViCBiostat). MMB is the recipient of an Australian Research Council Discovery Early Career Award (project number DE190101326) funded by the Australian Government. The VAHCS has been supported by the National Health and Medical Research Council, the Royal Children's Hospital Foundation and the Murdoch Children's Research Institute. The authors also acknowledge support provided to the Murdoch Children's Research Institute through the Victorian Government's Operational Infrastructure Support Program.

REFERENCES

1. National Health and Medical Research Council *Australian guidelines to reduce health risks from drinking alcohol.* Canberra: 2009.
2. World Health Organization *Global status report on alcohol and health 2018.* World Health Organization;2019.
3. Hermens DF, Lagopoulos J. Binge Drinking and the Young Brain: A Mini Review of the Neurobiological Underpinnings of Alcohol-Induced Blackout. *Frontiers in Psychology.* 2018;9(12).
4. Hayati Rezvan P, Lee KJ, Simpson JA. The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. *BMC Med Res Methodol.* 2015;15(1):30.
5. Rubin DB. *Multiple imputation for nonresponse in surveys.* New York: Wiley; 1987.
6. Schafer JL. *Analysis of incomplete multivariate data.* London: Chapman & Hall; 1997.
7. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research.* 2007;16(3):219-242.
8. Lee KJ, Carlin JB. Multiple Imputation for Missing Data: Fully Conditional Specification Versus Multivariate Normal Imputation. *American Journal of Epidemiology.* 2010;171(5):624-632.
9. Quartagno M, Carpenter J. *jomo: A package for Multilevel Joint Modelling Multiple Imputation*. https://CRAN.R-project.org/package=jomo. 2018.
10. Schafer JL, Olsen MK. Modeling and imputation of semicontinuous survey variables. Paper presented at: Proceedings of the Federal Committee on Statistical Methodology Research Conference. 1999.
11. Cameron AC, Trivedi PK. *Regression analysis of count data.* Vol 53: Cambridge university press; 2013.
12. Kleinke K, Reinecke J. Multiple imputation of incomplete zero-inflated count data. *Statistica Neerlandica.* 2013;67(3):311-336.

13. Javaras KN, van Dyk DA. Multiple Imputation for Incomplete Data with Semicontinuous Variables. *Journal of the American Statistical Association.* 2003;98(463):703-715.

14. Little RJA. Missing-Data Adjustments in Large Surveys. *Journal of Business & Economic Statistics.* 1988;6(3):287-296.

15. Morris TP, White IR, Royston P. Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Med Res Methodol.* 2014;14(1):75.

16. Yu LM, Burton A, Rivero-Arias O. Evaluation of software for multiple imputation of semi-continuous data. *Statistical Methods in Medical Research.* 2007;16(3):243-258.

17. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine.* 2011;30(4):377-399.

18. Quartagno M, Carpenter JR. Multiple imputation for discrete data: Evaluation of the joint latent normal model. 2019;61(4):1003-1019.

19. Allison PD. *Missing data.* Thousand Oaks, California: Sage Publications; 2002.

20. Lee KJ, Galati JC, Simpson JA, Carlin JB. Comparison of methods for imputing ordinal data using multivariate normal imputation: a case study of non-linear effects in a large cohort study. *Statistics in Medicine.* 2012;31(30):4164-4174.

21. Galati JC, Seaton KA, Lee KJ, Simpson JA, Carlin JB. Rounding non-binary categorical variables following multivariate normal imputation: evaluation of simple methods and implications for practice. *Journal of Statistical Computation and Simulation.* 2012;84(4):798-811.

22. Vink G, Frank LE, Pannekoek J, Van Buuren S. Predictive mean matching imputation of semicontinuous variables. *Statistica Neerlandica.* 2014;68(1):61-90.

23. van Buuren S. *Flexible Imputation of Missing Data.* Boca Raton: CRC Press; 2012.

24. Meng X-L. Multiple-Imputation Inferences with Uncongenial Sources of Input. *Statistical Science.* 1994;9(4):538-558.

25. Bartlett JW, Seaman SR, White IR, Carpenter JR, for the Alzheimer's Disease Neuroimaging I. Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research.* 2015;24(4):462-487.

26. Demirtas H. Practical Advice on How to Impute Continuous Data When the Ultimate Interest Centers on Dichotomized Outcomes Through Pre-Specified Thresholds. *Communications in Statistics - Simulation and Computation.* 2007;36(4):871-889.

27. Demirtas H. On imputing continuous data when the eventual interest pertains to ordinalized outcomes via threshold concept. *Computational Statistics & Data Analysis.* 2008;52(4):2261-2271.

28. Floden L, Bell ML. Imputation strategies when a continuous outcome is to be dichotomized for responder analysis: a simulation study. *BMC Med Res Methodol.* 2019;19(1):161.

29. Lu K, Jiang L, Tsiatis AA. Multiple Imputation Approaches for the Analysis of Dichotomized Responses in Longitudinal Studies with Missing Data. 2010;66(4):1202-1208.

30. Patton GC, Coffey C, Romaniuk H, et al. The prognosis of common mental disorders in adolescents: a 14-year prospective cohort study. *The Lancet.* 2014;383(9926):1404-1411.

31. Degenhardt L, Loughlin C, Swift W, et al. The persistence of adolescent binge drinking into adulthood: findings from a 15-year prospective cohort study. *BMJ Open.* 2013;3(8):e003015.

32. Galati JC, Seaton KA, Lee KJ, Simpson JA, Carlin JB. Rounding non-binary categorical variables following multivariate normal imputation: evaluation of simple methods and implications for practice. *Journal of Statistical Computation and Simulation.* 2014;84(4):798-811.

33. Carpenter JR, Kenward MG. Multiple Imputation and its Application. 2013.

34. Bernaards CA, Belin TR, Schafer JL. Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Statistics in Medicine.* 2007;26(6):1368-1382.

35. Goldstein H, Carpenter J, Kenward MG, Levin KA. Multilevel models with multivariate mixed response types. 2009;9(3):173-197.

36. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. 2019;38(11):2074-2102.

37. Quartagno M, Carpenter JR. Multiple imputation for IPD meta-analysis: allowing for heterogeneity and studies with missing covariates. *Statistics in medicine.* 2016;35(17):2938-2954.

38. Kleinke K. *countimp: Multiple imputation of incomplete count data*. https://github.com/kkleinke/countimp. 2020.

39. Van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. *J Stat Softw.* 2011;45(3):1-67.

40. *Stata Statistical Software: Release 15*. College Station, TX: StataCorp LP; 2017.

41. Royston P, White IR. Multiple Imputation by Chained Equations (MICE): Implementation in Stata. *2011.* 2011;45(4):20.

42. White IR, Daniel R, Royston P. Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables. *Computational Statistics & Data Analysis.* 2010;54(10):2267-2275.

43. White IR. simsum: Analyses of simulation studies including Monte Carlo error. *Stata Journal.* 2010;10(3):369-385.

44. Bessem B, De Bruijn MC, Nieuwland W, Zwerver J, Van Den Berg M. The electrocardiographic manifestations of athlete's heart and their association with exercise exposure. *European Journal of Sport Science.* 2018;18(4):587-593.

45. Jaakkola MS, Aalto SAM, Hyrkäs-Palmu H, Jaakkola JJK. Association between regular exercise and asthma control among adults: The population-based Northern Finnish Asthma Study. *PLOS ONE.* 2020;15(1):e0227983.

46. Brown JE, Broom DH, Nicholson JM, Bittman M. Do working mothers raise couch potato kids? Maternal employment and children's lifestyle behaviours and weight in early childhood. *Social Science & Medicine.* 2010;70(11):1816-1824.

47. Fancourt D, Steptoe A. Television viewing and cognitive decline in older age: findings from the English Longitudinal Study of Ageing. *Scientific Reports.* 2019;9(1):2851.

48. Horton NJ, Lipsitz SR, Parzen M. A Potential for Bias When Rounding in Multiple Imputation. *American Statistician.* 2003;57(4):229-232.

49. Allison PD. Imputation of categorical variables with PROC MI. *SUGI 30 proceedings.* 2005;113(30):1-14.

50. Rodwell L, Lee K, Romaniuk H, Carlin J. Comparison of methods for imputing limited-range variables: a simulation study. *BMC Med Res Methodol.* 2014;14(1):57.

FIGURE CAPTIONS

Figure 1. Simulation results for estimating the marginal proportions for the missing at random (MAR) scenario, with 20% (left column), 40% (middle column) and 60% (right column) of zero values in the point mass of the semi-continuous variable. Results shown are bias (top row), empirical standard error (middle row) and coverage (bottom row) of the estimates of the proportions in the no drinking, drinking below binge level, binge drinking and heavy binge drinking categories. Vertical grey lines in the middle row show the difference between the model-based standard errors and the empirical standard errors.

Figure 2. Simulation results for estimating the logistic regression coefficients for the semi-continuous exposure for the missing at random (MAR) scenario with 20% (left column), 40% (middle column) and 60% (right column) of zero values in the point mass of the semi-continuous variable. Results shown are bias (top row), empirical standard error (middle row) and coverage (bottom row) of the estimates of the coefficients for the drinking below binge level, binge drinking and heavy binge drinking categories with non-drinkers as the reference category. Vertical grey lines in the middle row show the difference between the model-based standard errors and the empirical standard errors.

Figure 3. Estimates of the marginal proportions in each drinking behaviour category for males in late adolescence from the VAHCS case study. Results shown are proportions and 95% confidence intervals. n=594 for complete case analysis (CCA); n=725 for multiple imputation analyses.

Figure 4. Estimates of associations between drinking behaviour in late adolescence and any binge drinking in young adulthood for males from the VAHCS case study. n=594 for complete case analysis (CCA); n=725 for multiple imputation analyses. Results represent log-odds ratios and 95% confidence intervals for each category compared to the no drinking (reference) category adjusted for peer drinking.