Imputation of Missing Data for a Continuous Variable with an Ordinal form of Risk Function: When to Apply the Transformation?

Mohammad Reza Baneshi[1], Behshid Garrusi[2] and Saiedeh Haji-Maghsoudi[3,*]

[1]Research Center for Modeling in Health, Institute for Futures Studies in Health, Kerman University of Medical Sciences, Kerman, Kerman University of Medical Sciences, Kerman, Iran

[2]Department of Community Medicine, Neuroscience Research Center, Afzalipour Medical School, Kerman University of Medical Sciences, Kerman, Iran

[3]Research Center for Social Determinants of Health, Institute for Futures Studies in Health, Kerman University of Medical Sciences, Iran

Abstract: *Introduction*: Imputation of missing data and selection of appropriate risk function are of importance . Sometimes a variable with continuous nature will be offered to the regression model as an ordinal variable. Our aim is to investigate whether to offer the continuous form of the variable to the imputation phase and its ordinal from to the modeling phase, or whether to offer the ordinal version to both phases.

Material and Methods: The outcome and main variable of interest was use of diet as a body change approach, and Body Mass Index (BMI). We randomly deleted 10%, 20%, and 40% of BMI values. In strategies 1 and 2, BMI was offered to the imputation phase as a continuous (BMIC) and ordinal variable (BMIO). Missing data were imputed using linear and polytomous regression respectively. In strategy 1, after imputation, BMIC was categorized (named BMICO) and offered to the modeling phase. In strategy 2, after imputation of BMIO values, this variable was offered to the logistic model (named BMIOO). We compared two strategies at Event Per Variables (EPV) of 75, 10, and 5.

Result: At EPVs of 75 and 10 no remarkable difference was seen. However, at EPV of 5, strategy 2 was superior. At 20% and 40% missing rates, strategy 1 was 2.21 and 3.67 times more likely to produce Severe Relative Bias. At high missing rate, power was higher in strategy2 (90% versus 83%).

Conclusions: When EPV is low and missing rate is high, categorizing of variable before imputation of missing data produces less SRB and leads to higher power.

Keyword: Missing data, risk function, transformation, Multiple Imputation.

INTRODUCTION

Missing data and appropriate form of risk function are two issues that challenge the process of model building. Missing data is a common problem in medical data sets [1, 2]. Impact of omission of missing data, and comparison of performance of imputation methods, has been addressed extensively in the literature [3-6]. While exclusion of subjects with missing data results in loss in power and can lead to biased estimates, imputation can recover the data so as to avoid such problems [5, 7-9].

Majority of published manuscripts compared different imputation methods in terms of estimation of survival rate [1, 5, 10], regression coefficients [7, 11, 12], or in terms of comparison of performance of diagnostic and prognostic models [8, 13, 14]. Research so far suggested the Multiple Imputation (MI) as the most appropriate approach as this method takes into account the imputation variation [15, 16]. This method uses linear regression (or predictive mean matching), logistic regression, and polytomous regression to impute missing data for continuous, binary, and ordinal variables [16, 17].

Regarding the shape of risk function, in medical applications, researchers often categorize the continuous covariates prior to modeling analyses. This makes data summarization more efficient, offers a simple risk classification, and allows for simple interpretation of results [18]. In the regression setting, for instance, interpretation of the impact of an ordinal covariate on outcome is easier than that for a change of 1 unit in a continuous covariate.

We already have explored the factors that contribute to body change activities (such as severe dieting) among Iranian populations, as an Asian culture. Body image has been defined as a person's feelings and thoughts about his or her body. We have shown that Body Mass Index (BMI) is a dominant factor which was positively associated with the chance of going for severe dieting [19]. It is possible to keep BMI in the continuous form, or to apply some cutoffs at 18.5, 25, and 30 to define thin, normal, overweight, and obese groups [20].

*Address correspondence to this author at the Research Center for Social Determinants of Health, Institute for Futures Studies in Health, Kerman University of Medical Sciences, Iran; Tel: +98 9133874358;
Fax: +983413205405; E-mail: sa.maghsoudi@gmail.com

E-ISSN: 1929-6029/14

When missing data exist, modeling process involve two steps: imputation of missing data, and development of appropriate diagnostic regression model (such as logistic regression model) to assess the significance of independent variables. Therefore, two regression modeling exercises are required. One issue which was of less concern was that in the case of missing data for a continuous variable with an ordinal form of risk function, whether to a) offer continuous from to imputation phase, b) categorize the variable after imputation, and c) offer the categorized variable to the logistic regression model, or whether to a) categorize the variable, b) offer the categorized from to imputation model, and then c) to the logistic regression model.

MATERIAL AND METHODS

We used data from a population based study carried out in south east of Iran. The main outcome of this study was the use of diet as a method body change (yes/ no question). Independent variables include demographic characteristics, socio-economic status, BMI, Perceived socio-cultural pressure, and body esteem score. Data were collected through a multistage household sampling. In each household, only one subject was interviewed. All of participants signed informed consent.

BMI values were available in the continuous form. Applying the cut offs at 18.5, 25, and 30 this variable was changed to an ordinal variable. Throughout this manuscript, we call BMI in the continuous and ordinal forms as BMIC and BMIO.

At the first step, we fitted a multifactorial logistic regression model in which BMIO was considered as the risk factor of interest. Effects of other independent variables were adjusted. Estimated coefficient and Odds Ratio (OR) was considered as the gold standard.

We then randomly deleted 10% of BMIC and BMIO values 100 times. Missing At Random indicates that the probability of being missing depends on the values of other covariates (say X_1, X_2,...,X_K) but not to the depend. It has been noted that including enough independent variables into the imputation model makes the MAR assumption plausible.

Multiple imputation approach was used to impute the missing data 10 times. This led to creation of 1000 data sets. Two strategies were followed. In strategy 1, missing data for BMIC were imputed using linear regression method. This variable was then categorized to generate BMICO (i.e. ordinal form of variable after imputation of continuous form). In strategy 2, missing values for BMIO were imputed using polytomous regression (named BMIOO). This was followed by fit of logistic regression model to each of 2000 data sets independently (1000 in each strategy).

Two strategies were compared in terms of proportion of data sets in which Severe Relative Bias (SRB) happened in estimation of OR, and power. Relative bias was defined as the difference between predicted OR minus real OR divided by real OR. Absolute values above 10% were defined as SRB. In addition, we counted number of data sets in which BMICO and BMIOO retained significant in the multifactorial model.

To address the level of overestimation and underestimation of the true OR, difference between estimated ORs and true OR was calculated. Mean of positive and negative values was considered as mean overestimation and underestimation respectively.

We also explored the impact of number of Events Per Variable (EPV) and missing rate on our results. In studies with a binary outcome, EPV has been defined as the minimum of number of subjects in the two groups. In our original data, sample size was 1204 of which 456 subjects experienced dieting for body change. Number of independent variables was 6. This gave EPV of about 75. We randomly selected a part of data set corresponded to EPV of 10 (N=150) and 5 (N=75). In addition, to address the impact of missing rate, 20% and 40% of data were deleted as explained above. The whole process was repeated to all scenarios. In other words, 9 compositions were studies changing EPV (75, 10, and 5) and missing rate (10%, 20%, and 40%). All analyses were done in R software.

RESULTS

In total 1204 subjects participated in this study of which 38% experienced severe dieting as a method of body change. About 16% of participants aged <20, and 67% formed the 20 to 40 years old group. Proportion of subject in four BMI groups was 10% (thin), 62% (normal), 24% (overweight), and 4% (obese) respectively.

EPV OF 75

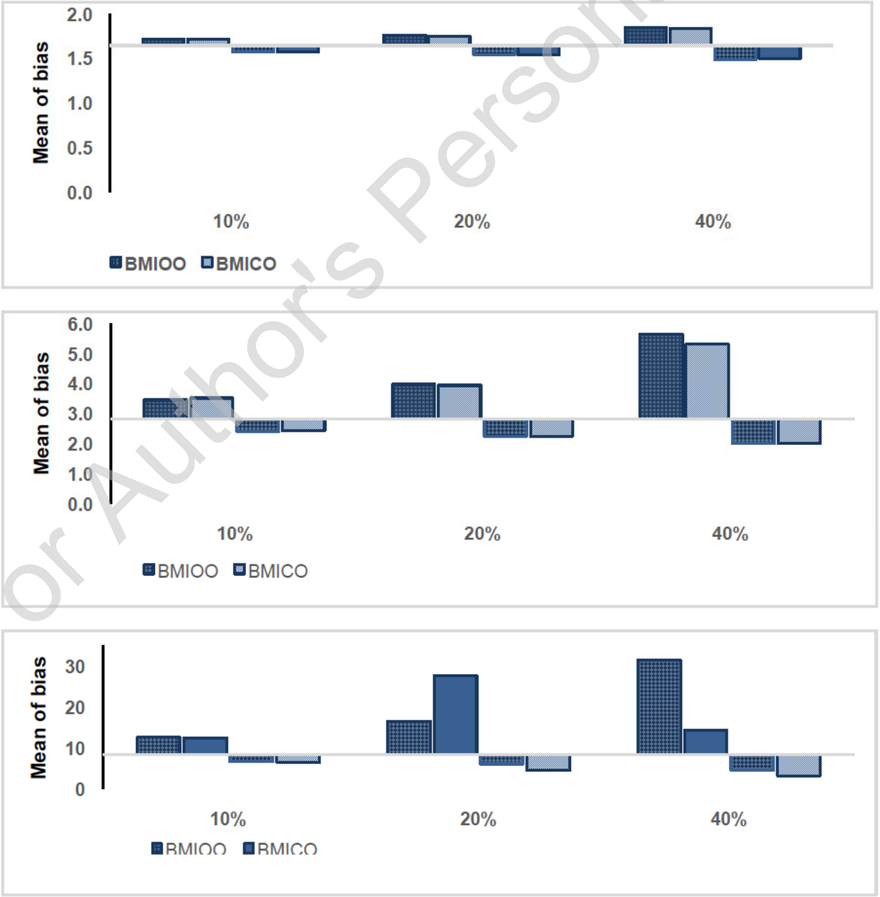
In all missing rates, performance of two strategies was almost the same in terms of power (nearly 100%),

Table 1: Comparison of Two Strategies in Terms of Power, and Proportion of Data sets at which Severe Relative Bias Observed

EPV	Strategy	Missing Rate					
		10%		20%		40%	
		SRB%	Power %	SRB%	Power %	SRB%	Power %
75	BMIOO	7.1	100	21.9	99.9	46.2	98.4
	BMICO	7.0	100	22.9	100	45.1	98.4
10	BMIOO	61.8	94.4	74.8	84.8	86.3	75.6
	BMICO	61.7	94.9	76.5	84.0	85.0	74.9
5	BMIOO	71.8	99.9	84.6	98.3	92.7	90.1
	BMICO	75.6	99.6	92.4	97.0	97.9	83.1

(Table **1**). At 40% missing rate slight reduction in power was seen in both strategies (98%).

The true OR for BMIO variable was 1.64. In all scenarios mean of ORs was fairly close to the true value (Figure **1**). In addition, level of overestimation and underestimation was the same. At 10% missing rate, in both strategies, mean of overestimation and underestimation was about 0.08 and 0.06. This figure reached to 0.20 and 0.15 at 40% missing rate.



Horizontal line shows true odds ratio
Upward columns show mean of overestimation
Downward columns show mean of underestimation
Top, middle, and bottom panels are correspond to EPVs of 75, 10, and 5

Figure 1: Estimation of mean of overestimation and underestimation of true OR at different EPVs, missing rates, and imputation strategy.

Table 2: Comparison of Two Strategies in Terms of Risk of Creation of Severe Relative Bias Odds Ratios

EPV	Strategy	Missing Rate		
		10%	20%	40%
		OR (CI 95%) P-value	OR (CI 95%) P-value	OR (CI 95%) P-value
75	BMIOO	Ref	Ref	Ref
	BMICO	0.99 (0.70,1.39) 0.93	1.06 (0.86,1.31) 0.59	0.96 (0.80,1.14) 0.62
10	BMIOO	Ref	Ref	Ref
	BMICO	1 (0.83,1.19) 0.96	1.09 (0.89,1.35) 0.38	0.90 (0.70,1.15) 0.41
5	BMIOO	Ref	Ref	Ref
	BMICO	1.22 (1,1.47) 0.05	2.21 (1.66,2.96) <0.001	3.67 (2.24,6.0) <0.001

At all missing rate, proportions of SRB in both strategies were fairly the same (7% at low, 22% at intermediate, and 45% at high missing rates respectively). Results of logistic regression modeling revealed that effect of strategy on creation of SRB estimates was not significant, regardless of missing rate (Table **2**).

EPV OF 10

No remarkable difference between two strategies was seen in terms of power. At low and high missing rate power was 0.95 and 0.75 respectively in both scenarios (Table **1**).

The true OR for BMIO variable was 2.82. Level of overestimation and underestimation, in both strategies, at 10% missing rate was about 0.65 and 0.40, which reached to 2.7 and 0.80 at high missing rate (Figure **1**).

Performances of models were also comparable in terms of proportion of SRB estimates (Table **1**). At low and high missing rate, in about 62% and 85% of data sets analysed using either of strategies, absolute relative bias was higher than 10%. Performance of strategy one was not poorer than strategy two.

EPV OF 5

In contrast to high and moderate EPVs, here power was higher in strategy2 at high missing rate. At 10% and 20% missing rate, in both strategies, powers were about 100% and 98% respectively. However, at high missing rate corresponding figures were 90% for strategy 2 versus 83% for strategy 1.

At this EPV, the true OR was 8.36. In both strategies at 10% missing rate, mean of overestimation and underestimation was about 4.2 and 1.7 respectively. In contrast to other EPVs, at moderate and high missing rates, mean of overestimation in two strategies were far from each other. At 20% missing rate, mean of overestimation in BMICO and BMIOO strategies were 20 and 8.5. Corresponding figures at 40% missing rate were 6 and 24 respectively. However, such big differences were not seen in terms of underestimation of true OR.

At 10% missing rate, in about 72% (strategy 2) and 76% (strategy 1) of data sets analysed, absolute relative bias was higher than 10%. We have seen that strategy 1 was 1.22 times more likely to produce SBR results (P=0.05). At 20% missing rate, corresponding figures were 85% versus 92%, giving OR of 2.21 (P-value<0.0001). At 50% missing rate, corresponding figures 93% versus 98%, giving OR of 3.67 (P-value<0.0001).

DISCUSSION

Our results showed that, at EPV of 75, performance of both strategies were fairly the same. No remarkable difference was seen in terms of neither mean ORs estimated nor power. Mean of estimated ORs in both strategies and at all missing rates were close to the true OR of 1.64. Mean bias was less than 0.10. Power at 10% and 40% missing rates were about 100% and 98% respectively.

We also observed that performance of MI was poor from SRB prospect in particular at high missing rate. When missing rate was 10%, in either of strategies, in about 7% of data sets SRB estimates derived. When missing rate increased to 40%, proportion of SRB estimates increased by a factor of around 6, and reached to 45%.

When EPV reduced to 10, both strategies at all missing rates, were able to provide mean ORs close to the true value of 2.82. However, increase in missing rate was associated with 20 percentage point decrease in power (94% at EPV of 75, versus 75% at EPV of 10). Interestingly, majority of estimated ORs were severely biased. SRB at 10% missing rate in both strategies was about 62%. Proportion of SRB estimates increased by a factor of about 1.5 and reached at 85%, when missing rate changed to 40%.

At low EPV of 5, difference between two strategies was remarkable. At 10% missing rate, mean ORs estimated were close to the true value of 8.36. However, at 40% missing rate, BMIOO and BMICO strategies led to huge overestimation and underestimation of true OR respectively. Furthermore, at 40% missing rates, power in BMIOO and BMICO was about 90% and 83% respectively. Proportion of SRB estimates in BMICO was significantly higher than BMIOO (76% versus 72% at 10% rate; 98% versus 93% at 40% rate). At 10% missing rate, BMICO strategy was 1.22 times more likely to create SRD results than BMIOO. This figure increased to 3.67 at high missing rate.

We also observed that increase in EPV was associated with decrease in SRB results. Furthermore, reduction at low missing rate was much faster that high missing rates. For example for BMIOO strategy, at 10% missing rate, SRB at EPV of 5 was about 10 times higher than EPV of 75 (71.8 versus 7.1). Corresponding ratio at 40 % missing rate was about 2 (92.7% versus 46.2%).

Research so far suggested that, in regression settings, at least 10 events per independent variable are required to get reliable estimates [21]. Furthermore, ability of MI to impute plausible values, and to provide unbiased estimates, has been confirmed in many studies [1, 11, 22, 23]. For example, Knol *et al.* assessed the risk factors of depression analyzing 1075 subjects [22]. They created missing data under different rates, and compared performance of different imputation methods. They have shown that performance of MI was satisfying even at 30% missing rate. However, we have seen that at EPV of 10 under 10% missing rate the probability of getting SRB results was about 60%.

The difference between our results and previous studies might be partially justified as below. Some authors admired the ability of MI, used one data set with no missing data and fitted a model. They considered results obtained, for example regression coefficients or ORs, as gold standard [1, 14, 22, 24]. Then they randomly deleted a proportion of the data, followed by imputation of missing data applying different imputation methods. Results obtained after imputation were compared with gold standard estimates. This approach does not take into account the sampling variation.

Some other authors generated the missing data several times [11, 12, 23]. However, they considered the average of estimates derived across data sets as the final estimate. In our experience, averaging might lead to wrong conclusion. To clarify this issue suppose the true OR is 3, and suppose we estimate it as 1 in half and 5 in the other half of data sets. In this hypothetical example, the mean of ORs converge to the true value. However, all of our 100 estimates are severely biased.

Another aspect of our work was to address when to apply the appropriate form of risk function; before or after imputation process. Our results showed that at high or intermediate EPV none of strategies was superior to other. On the other hand, at low EPV results suggests that application of appropriate form of risk function before imputation procedure is superior.

In our literature review, we could not find any manuscript that considers the process of imputation and risk function together. We only found one manuscript which addressed the imputation of a ratio variable. The variable of interest was BMI, which is the ratio of weight over square of height. Authors considered two strategies. In strategy 1, they offered BMI and some other independent variables, but not height and weight, to the imputation process. In strategy 2, they offered height and weight plus rest of independent variables to the imputation model, and computed BMI afterwards. No remarkable difference was seen between these two strategies at 20% and 40% missing rates. The EPV at Morris *et al*. work was about 30. However, authors did not compare two strategies at different EPVs. They performed extensive simulations in which the impact of coefficient of variation, R square, and missing mechanism was addressed [25].

Our study had several limitations. Firstly, we only considered ordinal form of risk function. Therefore, impact of complex non-linear effects remains to be addressed. Secondly, in our data set BMIC exhibited a normal distribution, and we applied linear regression in imputation phase. Future studies with extensive simulations are required to explore the performance of these strategies in the case of non-normal variables, and to compare different imputation schemes (such as linear regression versus predictive mean matching).

Besides these limitations, to our knowledge, our manuscript in one of the first studies that highlighted the problem of shape of risk function when missing data exists. We compared performance of two strategies at different EPVs and missing rates (18 combinations in total with 1000 sample at each). Our results demonstrate that, at high EPV, performance of both strategies were the same. However, at low EPV, BMIOO provides had better performance, in particular at high missing rates.

APPENDIX

BMI	=	Body Mass Index
BMIC	=	BMI in the continuous form
BMIO	=	BMI in the ordinal form
EPV	=	Events Per Variable
MI	=	Multiple Imputation
OR	=	Odds Ratio
SRB	=	Severe Relative Bias

REFERENCES

[1] Barzi F, Woodward M. Imputations of missing values in practice: results from imputations of serum cholesterol in 28 cohort studies. Am J Epidemiol 2004; 160(1): 34-45. http://dx.doi.org/10.1093/aje/kwh175

[2] Donders ART, *et al*. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 2006; 59(10): 1087-1091. http://dx.doi.org/10.1016/j.jclinepi.2006.01.014

[3] Acock AC. Working with missing values. J Marriage Family 2005; 67(4): 1012-1028. http://dx.doi.org/10.1111/j.1741-3737.2005.00191.x

[4] Arnold AM, Kronmal RA. Multiple imputation of baseline data in the cardiovascular health study. Am J Epidemiol 2003; 157(1): 74. http://dx.doi.org/10.1093/aje/kwf156

[5] Baneshi MR, Talei AR. Impact of imputation of missing data on estimation of survival rates: an example in breast cancer. Iran J Cancer Prevent 2010; 3(3): 127-31.

[6] Bono C, *et al*. Missing data on the Center for Epidemiologic Studies Depression Scale: a comparison of 4 imputation techniques. Res Soc Admin Pharm 2007; 3(1): 1-27. http://dx.doi.org/10.1016/j.sapharm.2006.04.001

[7] Baneshi MR. Prevention of Disease Complications through Diagnostic Models: How to Tackle the Problem of Missing Data? Iran J Public Health 2012; 41(1).

[8] Baneshi MR, Talei AR. Does the Missing Data Imputation Method Affect the Composition and Performance of Prognostic Models? Iran Red Cresce Med J 2012; 14(1): 51-6.

[9] Zhang X. A Study of Methods for Missing data problems in Epidemiologic Studies with Historical Exposures 2009; University of Southern California.

[10] Van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. Statist Med 1999; 18(6): 681-694. http://dx.doi.org/10.1002/(SICI)1097-0258(19990330)18:6<681::AID-SIM71>3.0.CO;2-R

[11] Janssen KJM, *et al*. Missing covariate data in medical research: to impute is better than to ignore. J Clin Epidemiol 2010; 63(7): 721-727. http://dx.doi.org/10.1016/j.jclinepi.2009.12.008

[12] Marshall A, *et al*. Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. BMC Med Res Methodol 2010; 10(1): 7. http://dx.doi.org/10.1186/1471-2288-10-7

[13] Gorelick MH. Bias arising from missing data in predictive models. J Clin Epidemiol 2006; 59(10): 1115-1123. http://dx.doi.org/10.1016/j.jclinepi.2004.11.029

[14] Schenker N, *et al*. Multiple imputation of missing income data in the National Health Interview Survey. J Am Statist Assoc 2006; 101(475): 924-933. http://dx.doi.org/10.1198/016214505000001375

[15] Azur MJ, *et al*. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res 2011; 20(1): 40-49. http://dx.doi.org/10.1002/mpr.329

[16] White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Statist Med 2010; 30(4): 377-399. http://dx.doi.org/10.1002/sim.4067

[17] Wayman JC. Multiple imputation for missing data: What is it and how can I use it 2003.

[18] Baneshi M, Talei A. Dichotomisation of continuous data: review of methods, advantages, and disadvantages. Iran J Cancer Prevent 2011; 4(1): 26-32.

[19] Garrusi B, Garousi S, Baneshi MR. Body image and body change: predictive factors in an Iranian population. Int J Prevent Med 2013; 4(8): 940.

[20] Al-Sendi A, Shetty P, Musaiger A. Prevalence of overweight and obesity among Bahraini adolescents: a comparison between three different sets of criteria. Eur J Clin Nutr 2003; 57(3): 471-474. http://dx.doi.org/10.1038/sj.ejcn.1601560

[21] Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. Am J Epidemiol 2007; 165(6): 710. http://dx.doi.org/10.1093/aje/kwk052

[22] Knol MJ, *et al*. Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example. J Clin Epidemiol 2010; 63: 728-736. http://dx.doi.org/10.1016/j.jclinepi.2009.08.028

[23] Langkamp DL, Lehman A, Lemeshow S. Techniques for handling missing data in secondary analyses of large surveys. Acad Pediatr 2010; 10(3): 205-210. http://dx.doi.org/10.1016/j.acap.2010.01.005

[24] Guan NC, Yusoff MSB. Missing values in data analysis: Ignore or Impute? 2011.

[25] Morris TP, *et al*. Multiple imputation for an incomplete covariate that is a ratio. Statist Med 2014; 33(1): 88-104. http://dx.doi.org/10.1002/sim.5935

Received on 25-07-2014 Accepted on 19-10-2014 Published on 06-11-2014

http://dx.doi.org/10.6000/1929-6029.2014.03.04.6