Iran J Public Health, Vol. 50, No.7, Jul 2021, pp.1372-1380

Original Article

Evaluation of Multiple Imputation with Large Proportions of Missing Data: How Much Is Too Much?

Jin Hyuk Lee [1], J. Charles Huber Jr. [2]

1. *Graduate School of Social Welfare, Yonsei University, Seoul, Republic of Korea*
2. *Stata Corp, College Station, TX, USA*

***Corresponding Author:** Email: gene2you@gmail.com

(Received 12 Jun 2020; accepted 09 Sep 2020)

Abstract

Background: Multiple Imputation (MI) is known as an effective method for handling missing data in public health research. However, it is not clear that the method will be effective when the data contain a high percentage of missing observations on a variable.

Methods: Using data from "Predictive Study of Coronary Heart Disease" study, this study examined the effectiveness of multiple imputation in data with 20% missing to 80% missing observations using absolute bias (|bias|) and Root Mean Square Error (RMSE) of MI measured under Missing Completely at Random (MCAR), Missing at Random (MAR), and Not Missing at Random (NMAR) assumptions.

Results: The |bias| and RMSE of MI was much smaller than of the results of CCA under all missing mechanisms, especially with a high percentage of missing. In addition, the |bias| and RMSE of MI were consistent regardless of increasing imputation numbers from M=10 to M=50. Moreover, when comparing imputation mechanisms, MCMC method had universally smaller |bias| and RMSE than those of Regression method and Predictive Mean Matching method under all missing mechanisms.

Conclusion: As missing percentages become higher, using MI is recommended, because MI produced less biased estimates under all missing mechanisms. However, when large proportions of data are missing, other things need to be considered such as the number of imputations, imputation mechanisms, and missing data mechanisms for proper imputation.

Keywords: Public health research; Multiple imputation; Large proportions of missing data; Coronary heart disease

Introduction

Missing values are a common and complex problem in public health (1). If a researcher improperly treats missing values, this affects many, if not most, data analyses and can cause problems ranging from minor underestimation of variance estimates to severely biased parameter estimates (2). Bias occurs when observed cases are not representative of the complete sample and may be substantial when the percentage of missing is sufficiently large (3). Even in absence of bias, when data are analyzed ignoring missing values, the power of the test is decreased as the sample size is reduced (4).

In order to analyze data with missing values properly, it is necessary to understand different missing data mechanisms (5). Of the missing

mechanisms, Missing Completely At Random (MCAR) requires the strongest assumption (6). The assumption is that the probability of missingness is dependent neither on missing value itself nor any observed values in the data set (7). When the data are Missing At Random (MAR), the weaker assumption is made (6). When the observed values are given, the probability of missingness does not depend on the missing values themselves, but it might depend on the observed variables (7). If the MAR assumption is not valid, one can say that data are Not Missing at Random (NMAR) (6). That is, the probability of missingness is dependent on the unobserved value of missing variables itself (7).

Of the methods to handle missing data, Complete Case Analysis (CCA), which discards all observations with missing values, is relatively easy to use and it is most frequently used by researchers (6). Under MCAR, the subsample of cases with complete data is not different from the simple random sample from the original data. So, CCA does not introduce a bias for estimation (6). Under MAR, CCA is not recommended usually because it yields biased parameter estimates. If missing values are independent of the observed values but dependent on responses, estimation will not introduce bias with CCA even under MAR (5). Although the estimated standard error is unbiased, there is some loss of power (4).

An alternative method is the Multiple Imputation (MI), which was first proposed by Rubin (8). Generally, MI has 3 steps: imputation step, analysis step, and combination step. In the imputation step, based on the distribution of given set of data, by an imputation mechanism, multiple imputation methods replace missing values with plausible substitutes that correctly represent the uncertainty. In analysis step, after repeating the replacement procedure several times and creating M>1datasets, the multiply imputed data sets are analyzed separately and independently by standard procedures (6). In the combination step, the estimates are combined, and confidence intervals are acquired using Rubin's combination rules (7). MI is more efficient than CCA, because MI uses information in the incomplete cases (5). In addition, MI corrects the bias under MAR (6).

Even though proportion of missing data affects significantly statistical inference, there is no established guidelines about an acceptable percentage of missing data which MI will has benefits. In a literature, when more than 10% of data are missing, estimates are likely to be biased (9). Another paper mentioned that 5% of missing rate has been suggested as a lower cutoff point below which MI provides insignificant benefit (10). However, those cutoff points have a limited evidence to support them. A small number of studies have investigated bias and efficiency by increasing percentages of missing data. This has been done with a maximum of 50% missing data in study that showed increasing inconsistency of effect estimates with increased missingness (11). Where more than 50% missingness has been investigated (12), the study sample size was very small, thus limiting the applicability of results to larger public health researches. Where both more than 50% missingness and large sample size have been used (3), the study has been only examined under MAR. These findings relate to small percent of missing data or small sample size and not to the situation with the huge missing percent on large sample size data under different missing mechanisms, where such issue may exist in public health research (13).

As MI has been used in many fields increasingly (14), the method has been recommended to use in public health research and the effectiveness of the MI is tested in different settings (1,15-16). However, the usefulness and validity of MI still need to be examined in various settings including a heart disease data with high percent of missing on a continuous variable. So, this study will answer this question; does the MI attains accuracy and efficiency even in high percent of missing data on a public health setting?

The objective of this study was to explore how much data could be lost and successfully imputed using MI under a variety of scenarios. The first aim was to show the biasness of different methods according to the percentage of missing values in the data under MCAR, MAR, and NMAR as-

sumptions. The second aim was to suggest different optimal set-ups of MI (the number of M imputed data sets and imputation mechanisms) according to different missing mechanisms and percentages of missing data.

Materials and Methods

Data

We used the data which came from the study, "A Predictive Study of Coronary Heart Disease"(17). Subjects were collected from eleven business organizations in San Francisco and two in Los Angeles. In 1960, 3,524 males, aged 39 to 59 years, participated in interviews and medical examinations. However, in order to have complete data, completely observed 3,101 subjects were used in the analysis.

We used the CHD study to investigate the issues arising with imputing high percentages of missing data. The systolic blood pressure (SBP) variable was assigned to a missing variable. The mean of this variable was considered as a parameter of interest. The variables for imputation model were chosen from predictor variables in the dataset such as diastolic blood pressure, height, weight, age, BMI and cholesterol.

In this study, to compare various imputation mechanisms for Multiple Imputation, different conditions of data were manipulated using the complete data of the CHD study. A univariate pattern of missing data was generated according to previous studies (18). Some entities for a variable (SBP) in the dataset were deleted while all other variables were retained. Then the different amounts of entities for the variable (SBP) were deleted at random causing MCAR mechanism, which had 0%, 20%, 40%, 60%, and 80% missing data. MAR data was simulated by sorting according to one of the completely observed variables (age) and deleting the lower values of the cases of SBP by different percentages of the missing values to give the MAR mechanism. For the NMAR mechanism, the complete data was sorted by the missing variable itself and the values of SBP were deleted by five different rates. It was done separately and independently to create incomplete datasets for three missing mechanisms.

Methods

In this study, MI was used mainly to assess its effectiveness and bias. In order to compare the effectiveness of the two methods, such as CCA and MI, the bias was measured by absolute bias and root mean squared error. The parameter of interest was the mean of a variable, which was partially missing. We also studied how different methods work when 20%, 40%, 60%, and 80% of the data were missing, the complete data as a reference. Furthermore, the biases of the methods were compared under different missing mechanisms--MCAR, MAR, and NMAR assumptions.

One of the focuses of the study was finding an optimal setting for the MI under different conditions, such as different missing mechanisms (MCAR, MAR, and NMAR) and different percentages of missing data. The varying number, M, of imputed dataset and different imputation mechanisms were considered. Specially, the number of M imputed dataset raised from 10 to 50. Also, the used imputation mechanisms of MI included regression method, predictive mean matching method, and MCMC method. 500 repetitions were done for each result in order to reduce the random variability of imputed values.

Results

As shown in Table 1, MI had a lower |bias| and RMSE than CCA under all missing mechanisms. The |bias| and RMSE were obtained using a true parameter estimate (128.63), the mean of SBP at 0% missing. In other words, with MAR and NMAR the |bias| of MI was smaller than that of CCA. Moreover, with MCAR the MI's |bias| was smaller too.

Table 1: Comparison of CCA and MI

Missing mechanism	*Missing percent*	*CCA*			*MI*		
		Mean of estimate	Mean of |bias|	Mean of RMSE	Mean of estimate	Mean of |bias|	Mean of RMSE
MCAR	0%	128.63	-	-	128.63	-	-
	20%	128.94	0.3	0.43	128.75	0.12	0.31
	40%	129.22	0.59	0.69	128.89	0.25	0.4
	60%	129.23	0.6	0.74	128.92	0.29	0.45
	80%	129.88	1.25	1.4	129.2	0.57	0.74
MAR	0%	128.63	-	-	128.63	-	-
	20%	127.78	0.86	0.9	128.2	0.43	0.51
	40%	126.88	1.75	1.78	127.83	0.81	0.85
	60%	126.06	2.58	2.6	127.62	1.01	1.06
	80%	125.13	3.51	3.54	127.17	1.46	1.51
NMAR	0%	128.63	-	-	128.63	-	-
	20%	122.86	5.77	5.77	125.15	3.48	3.48
	40%	119.09	9.54	9.54	121.93	6.7	6.7
	60%	115.5	13.13	13.13	118.41	10.23	10.23
	80%	111.23	17.4	17.4	113.55	15.09	15.09

In addition, with the MCAR, MAR, NMAR assumptions, the RMSE of the MI was smaller than that of the CCA. With the missing mechanisms, as the amount of missing values increased, the |bias| and the RMSE became larger on both the CCA and MI. However, when the missing percentage is high, the estimates of CCA were more seriously biased than that of MI. For example, under MAR the difference between the |bias| of MI and the |bias| of CCA at 20% missing is 0.43, but the difference at 80% missing was 2.05. Moreover, as the difference between the RMSE of MI and the RMSE of CCA increased, the missing percentage also increased.

As shown in Table 2, when comparing different imputation numbers (M=10,20,30,40, and 50) of MI, the increased imputation numbers had no effect on |bias| and RMSE under different proportion of missing data and missing mechanisms. The difference between the RMSE with imputation numbers=10 and the RMSE with imputation numbers=50 was not more than 0.008. So, although the imputation numbers increased, the difference by imputation numbers in this dataset was not able to be seen.

Table 3 compared the imputation mechanisms--regression, PMM, and MCMC methods. With MCAR, MCMC produced much better estimates than other methods regarding |bias|. However, with MAR, it was not easy to determine if one was better than the others. That is, |bias| of the regression method was slightly smaller or larger than the others at some percent of missing values. With MAR and NMAR, the |bias| of the MCMC was almost the same as those of the regression method. With NMAR, MCMC and regression methods produced less biased imputed values than the PMM method. The variance of the estimate was so small that the results of the RMSE were the same as |bias|. In other words, under all missing mechanisms the MCMC method had a universally smaller RMSE (Fig. 1).

Table 2: Comparison of 10, 20, 30, 40, 50 imputation numbers of MI

Missing percent	*Imputation number*	*MCAR*			*MAR*			*NMAR*		
		Mean of estimate	Mean |bias|	Mean RMSE	Mean of estimate	Mean |bias|	Mean RMSE	Mean of estimate	Mean |bias|	Mean RMSE
0%	-	128.63	-	-	128.63	-	-	128.63	-	-
20%	10	128.75	0.12	0.31	128.20	0.43	0.51	125.15	3.48	3.48
40%	10	128.89	0.25	0.40	127.83	0.81	0.85	121.93	6.70	6.70
60%	10	128.92	0.29	0.45	127.62	1.01	1.06	118.41	10.23	10.23
80%	10	129.20	0.57	0.74	127.17	1.46	1.51	113.55	15.09	15.09
0%	-	128.63	-	-	128.63	-	-	128.63	-	-
20%	20	128.75	0.12	0.31	128.20	0.43	0.51	125.15	3.48	3.49
40%	20	128.88	0.25	0.40	127.82	0.81	0.86	121.93	6.70	6.71
60%	20	128.92	0.29	0.45	127.62	1.01	1.06	118.41	10.23	10.23
80%	20	129.20	0.57	0.74	127.17	1.46	1.51	113.54	15.09	15.09
0%	-	128.63	-	-	128.63	-	-	128.63	-	-
20%	30	128.75	0.12	0.31	128.20	0.43	0.51	125.15	3.48	3.49
40%	30	128.88	0.25	0.40	127.82	0.81	0.86	121.93	6.70	6.70
60%	30	128.92	0.29	0.45	127.62	1.01	1.06	118.41	10.23	10.23
80%	30	129.20	0.57	0.73	127.17	1.46	1.51	113.54	15.09	15.09
0%	-	128.63	-	-	128.63	-	-	128.63	-	-
20%	40	128.75	0.12	0.31	128.20	0.43	0.51	125.15	3.48	3.48
40%	40	128.88	0.25	0.40	127.82	0.81	0.85	121.93	6.70	6.70
60%	40	128.92	0.29	0.45	127.62	1.01	1.06	118.41	10.23	10.23
80%	40	129.20	0.57	0.74	127.17	1.46	1.51	113.55	15.09	15.09
0%	-	128.63	-	-	128.63	-	-	128.63	-	-
20%	50	128.75	0.12	0.31	128.20	0.43	0.51	125.15	3.48	3.48
40%	50	128.88	0.25	0.40	127.82	0.81	0.86	121.93	6.70	6.71
60%	50	128.92	0.29	0.45	127.62	1.01	1.06	118.41	10.23	10.23
80%	50	129.20	0.57	0.74	127.17	1.46	1.51	113.54	15.09	15.09

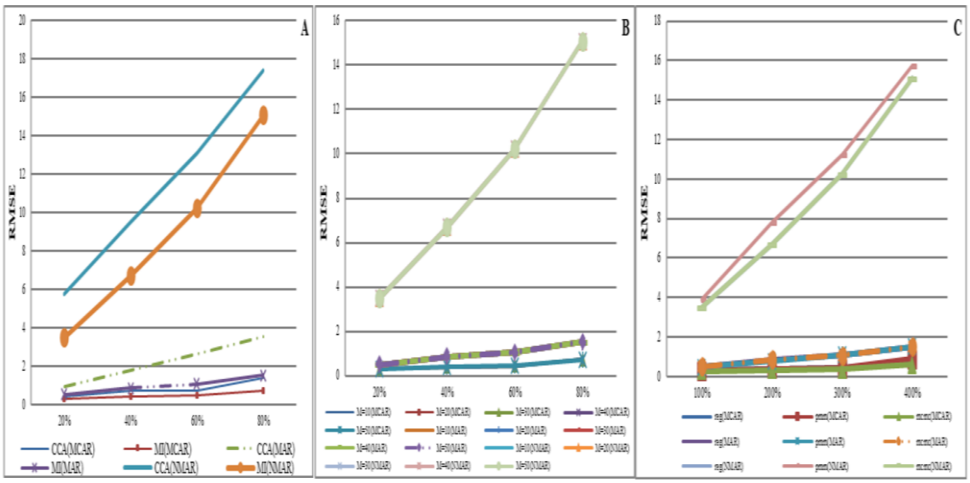

Fig. 1: The RMSE for CCA and MI(A), different imputation numbers of MI(B), and different imputation mechanisms under MCAR, MAR, and NMAR. (MI=Multiple Imputation, CCA= Complete case analysis, RMSE: Root Mean Square Error)

Table 3: Comparison of Regression Method, PMM and MCMC as Imputation Mechanism of MI

Missing mechanism	***Missing percent***	***Regression***			***PMM***			***MCMC***		
		Mean of estimate	Mean |bias|	Mean RMSE	Mean of estimate	Mean |bias|	Mean RMSE	Mean of estimate	Mean |bias|	Mean RMSE
MCAR	0%	128.63	-	-	128.63	-	-	128.63	-	-
	20%	128.75	0.12	0.31	128.76	0.12	0.31	128.64	0.01	0.28
	40%	128.88	0.25	0.4	128.89	0.26	0.4	128.53	0.1	0.32
	60%	128.92	0.29	0.45	128.89	0.26	0.41	128.73	0.1	0.37
	80%	129.2	0.57	0.74	129.44	0.81	0.89	128.24	0.39	0.59
MAR	0%	128.63	-	-	128.63	-	-	128.63	-	-
	20%	128.2	0.43	0.51	128.22	0.41	0.49	128.2	0.43	0.51
	40%	127.82	0.81	0.86	127.92	0.71	0.77	127.82	0.81	0.86
	60%	127.62	1.01	1.06	127.57	1.06	1.1	127.62	1.01	1.06
	80%	127.17	1.46	1.51	127.16	1.47	1.51	127.18	1.46	1.51
NMAR	0%	128.63	-	-	128.63	-	-	128.63	-	-
	20%	125.15	3.48	3.48	124.72	3.91	3.92	125.15	3.48	3.48
	40%	121.93	6.7	6.71	120.85	7.79	7.79	121.93	6.7	6.71
	60%	118.41	10.23	10.23	117.4	11.23	11.23	118.41	10.23	10.23
	80%	113.54	15.09	15.09	112.92	15.71	15.71	113.55	15.09	15.09

Discussion

The purpose of the analyses presented in this paper has been to highlight the importance of missing data and the potential implications of this problem with regard to the evaluation of theories and the making of parameter estimates. Based on literatures (6-7), MI is known as an effective method to deal with missing data problems, according to excellent parameter estimation, variance estimation, and increased power. However, this study investigated whether the method will still be accurate with high percentages of missing values and focused on how to increase the efficiency and accuracy of MI with changing conditions and options of the huge missing values in the data.

When CCA was employed for the data, the absolute bias and root mean squared error of the CCA was noticeably larger than those of MI. In other words, in addition to the decreased statistical power in CCA, the estimates of CCA was more biased than those of MI. Like other researches (5-6,15), the results in the study presented that with MAR and NMAR, the estimation of MI was more accurate than CCA. However, theoretically with MCAR the |bias| of CCA should be the same as that of MI (2), but MI produced more similar estimates to the true values in the study. Under MAR of this dataset, MI produced better estimates than CCA. Even though a research reported that the estimates of CCA were less biased than them of MI under their MAR simulations (19), these results were possibly due to the mis-modelling of an imputation model (20). In this study, the imputation model for MI were constructed by including all important variables, so the better performance of MI may be attained than CCA. In addition, the more data that were missing, the more the RMSE and the |bias| of both methods increased. Moreover, as the amount of missing values increased, the difference of the |bias| between MI and CCA increased. Thus, it is obvious that percentages of missing values had significant influences on the |bias| of both the MI and the CCA methods, but the |bias| of CCA was more seriously affected by the increased percentages of missing data. In other words, with all missing mechanisms, MI did better estimations than CCA in these data,

when there was a high percentage of missing data. However, we cannot say that MI provided an excellent estimation with NMAR. That is, given that the RMSE of MI was 1.5 at 80% missing value with MAR, with the NMAR assumption the RMSE of the MI was 3.5 at 20% missing value and the RMSE of the MI was 15.1 at 80% missing value. Thus, if the data were under MCAR or MAR, MI produced reliably accurate estimates even in large proportion of missing data in this dataset. However, under NMAR, MI provided biased estimates in relatively small proportion of missing data. When missing values were NMAR under certain settings, MI produced better estimates than CCA (5, 15), but neither MI nor CCA may be entirely appropriate in these data. It indicates that the performances of the MI under NMAR are likely to vary in different data or conditions (21). So, missing data analysis requires to conduct sensitivity analyses or to apply other strategies if NMAR is suspected (16, 20, 22). According to the results of this study, the MI produced relatively unbiased estimates than CCA under different conditions. However, we cannot say that MI is always preferable to CCA for any missing case even in similar scenarios if the MI is improperly used without careful consideration and appropriate examination.

In order to see whether increasing imputation numbers influence the efficiency of MI, the imputation numbers were increased from 10 to 50 and compared with the |bias| and the RMSE at different percentages of missing values. In agreement of a previous study (15), the RMSE and |bias| among different imputation numbers were almost the same on the data under MCAR, MAR, and NMAR. This is because 500 repetitions on each imputation number were done and averaged, which creates an estimate of 500×each imputation number (M=10, 20, 30, 40, and 50). When comparing the estimate of 5000 imputation numbers (500×10 imputation numbers) and the estimate of 25000 imputation numbers (50× 50 imputation numbers), RMSE and |bias| of them were not different. So, it did not improve the accuracy of estimation to increase the imputation numbers by a large amount. Thus, the imputation numbers may not have much effect on bias with the characteristics of this data.

For the study's data, which had continuous and univariate missing values, I compared regression method, predictive mean matching method, and MCMC method using different imputation mechanisms in order to determine which methods perform better for the considered data set. The results were similar to a previous study (16). With MCAR, the MCMC method produced a significantly lower RMSE and |bias| than the other methods as missing percent increased. With MAR, it was hard to tell which methods provided a better estimation for this data. With the NMAR assumption, the MCMC method and the regression method produced less biased estimates. Because the MCMC method produced overall unbiased estimates for missing values under all missing mechanisms, the MCMC method was the better imputation mechanism not only for continuous and multivariate missing variables (18), but also for continuous and univariate variable of large proportion of missing values.

This study's results have important implication for public health researchers, for conducting analysis on incomplete data. These results imply that researchers should not give up analysis even if the data has large proportion of missing in a variable. When MI is used with proper conditions, there are possibilities to correct bias and improve efficiency even with high percentages of missing data. This paper tested accuracy and efficiency of MI on various scenarios with 500 repetitions and used data with relatively large sample size which are similar to public health researches. However, there are some limitations including simple analysis model and missingness in one continuous variable. So, future studies are warranted to further investigate the effectiveness of MI with large proportion of missing data on more complex conditions.

Conclusion

The MI is not the best way to deal with missing data issues, even though the estimates of MI were relatively accurate and efficient in this study setting. In order to attain that effectiveness of MI even in a high percentage of missing data, many conditions need to be considered such as imputation numbers, imputation mechanisms, and missing mechanisms.

Ethical considerations

Ethical issues (Including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy, etc.) have been completely observed by the authors.

Acknowledgements

This study is based on a part of an author's M.S.P.H. dissertation completed at the Texas A&M University.

Conflict of interest

The authors declare that there is no conflict of interests.

References

1. Zhou XH, Eckert GJ, Tierney WM (2001). Multiple imputation in public health research. *Stat Med*, 20(9-10): 1541-9.
2. Deng Y, Chang C, Ido MS, Long Q (2016). Multiple imputation for general missing data patterns in the presence of high-dimensional data. *Sci Rep*, 6: 21689.
3. Lee KJ, Carlin JB (2012). Recovery of information from multiple imputation: a simulation study. *Emerg Themes Epidemiol*, 9(1):3.
4. Little RJA (1992). Regression with missing X's: a review. *J Am Stat Assoc,* 87:1227-1237.
5. Janssen KJ, Donders AR, Harrell FE Jr, et al (2010). Missing covariate data in medical research: to impute is better than to ignore. *J Clin Epidemiol*, 63(7): 721-7.
6. Allison PD (2002). Missing data. Sage publications. Thousand Oaks. doi: 10.4135/9781412985079.
7. Little RJ, Rubin DB (2019). *Statistical analysis with missing data*, John Wiley & Sons. New York. doi: 10.1002/9781119482260.
8. Rubin DB (1976). Inference and missing data. *Biometrika*, 63(3):581-92.
9. Bennett DA (2001). How can I deal with missing data in my study?. *Aust N Z J Public Health*, 25(5):464-9.
10. Schafer JL (1999). Multiple imputation: a primer. *Stat Methods Med Res*, 8(1):3-15.
11. Mishra S, Khare D (2014). On comparative performance of multiple imputation methods for moderate to large proportions of missing data in clinical trials: a simulation study. *J Med Stat Inform*, 2(1):9.
12. Hardt J, Herke M, Brian T, Laubach W (2013). Multiple imputation of missing data: a simulation study on a binary response. *Open Journal of Statistics*, 3(05):370.
13. Emdin CA, Rothwell PM, Salimi-Khorshidi G, et al (2016). Blood pressure and risk of vascular dementia: evidence from a primary care registry and a cohort study of transient ischemic attack and stroke. *Stroke*, *47*(6): 1429-35.
14. Kenward MG, Carpenter J (2007). Multiple imputation: current perspectives. *Stat Methods Med Res,* 16(3): 199-218.
15. Mirmohammadkhani M, Foroushani AR, Davatchi F, et al (2012). Multiple Imputation to Deal with Missing Clinical Data in Rheumatologic Surveys: an Application in the WHO-ILAR COPCORD Study in Iran. *Iran J Public Health*, 41(1): 87.
16. Miri HH, Hassanzadeh J, Rajaeefard A, et al (2016). Multiple Imputation to Correct for Nonresponse Bias: Application in Non-Communicable Disease Risk Factors Survey. *Glob J Health Sci*, 8(1): 133-42.
17. Rosenman RH, Friedman M, Straus R, et al (1964). A predictive study of coronary heart disease: The Western Collaborative Group Study. *JAMA*, 189(1):15-22.
18. Scheffer J (2002). Dealing with missing data. *Res Lett Inf Math Sci*, 3(1):153–60. Available from:

https://mro.massey.ac.nz/handle/10179/4355

19. Giorgi R, Belot A, Gaudart J, Launoy G (2008). the French Network of Cancer Registries FRANCIM. The performance of multiple imputation for missing covariate data within the context of regression relative survival analysis. *Stat Med,* 27:6310-31.
20. Horton NJ, White IR, Carpenter J (2010). The performance of multiple imputation for missing covariates relative to complete case analysis. *Stat Med*, 29(12): 1357.
21. Marshall A, Altman DG, Royston P, et al (2010). Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. *BMC Med Res Methodol*, 10: 7.
22. Galimard JE, Chevret S, Protopopescu C, et al (2016). A multiple imputation approach for MNAR mechanisms compatible with Heckman's model. *Stat Med*, 35(17):2907-20.