JP Journal of Biostatistics
© 2014 Pushpa Publishing House, Allahabad, India
Published Online: March 2015
Available online at http://www.pphmj.com/journals/jpb.htm
Volume 11, Number 2, 2014, Pages 117-134

IMPACT OF MISSING RATE AND METHOD OF IMPUTATION OF MISSING DATA ON THE SIZE ESTIMATION OF HIDDEN GROUPS USING NETWORK SCALE-UP

Nasim Dehdashti[1], **Aliakbar Haghdoost**[2] **and**
Mohammad Reza Baneshi[2,*]

[1]Regional Knowledge Hub for HIV/AIDS Surveillance
Institute for Futures Studies in Health
Kerman University of Medical Sciences
Kerman, Iran

[2]Research Center for Modeling in Health
Institute for Futures Studies in Health
Kerman University of Medical Sciences
Kerman, Iran
e-mail: m_baneshi@kmu.ac.ir
rbaneshi@yahoo.com

Abstract

Network scale-up is a standard tool in size estimation of hidden groups. Our aim is to address the impact of missing data and imputation methods on its results. Recruiting 997 Iranian from general population, the prevalence of misuse of ten drugs was calculated. Then 10%, 30% and 50% of data were deleted 200 times. Sizes of groups were predicted analyzing complete case (CC), and after imputation by

Received: May 21, 2014; Accepted: August 14, 2014
Keywords and phrases: AIDS, hidden group, missing data, network scale-up, size estimation.
This work has been supported by Kerman University of Medical Sciences (Grant no 92/189).

[*]Corresponding author

median replacement (MED), linear and negative binomial regression (NB), and expectation maximum (EM). For positive relative biases (RB), values > 10% were defined as severe relative bias (SRB+). For negative RBs, SRB– was defined as values < –10%. At 10% and 30% missing rates, differences between contribution of MED and EM to create SRBs were 35% (41% versus 6%) and 10% (25% versus 10%). For MED, majority of SRBs happened was SRB–. However, majority of SRBs seen in linear and NB regression was SRB+. At 10%, relative to EM, all methods were more likely to produce SRB. By increase in missing rate, superiority of EM over other methods reduced. MED and linear regression imputations were the poorest methods. At 10% missing, EM partially reduced bias. However, at moderate missing rate, performance of no method was satisfying.

Abbreviations

CC	:	Complete case
EM	:	Expectation maximum
MED	:	Median substitution
NB	:	Negative binomial
OR	:	Odds ratio
RB	:	Relative bias
SRB	:	Severe relative bias
SRB+	:	Positive severe relative bias
SRB–	:	Negative severe relative bias

Introduction

To control spread of HIV, estimation of size of the most at risk populations such as injected drug users is vital. This is because the prevalence of AIDS among such groups is much higher than the general population. For example, while the prevalence of HIV among Iranian general population is 0.1% [11], corresponding figure among injected drug users is much higher, at 15% [12].

In the case of sensitive issues, direct size estimation methods are not recommended [16, 19]. The network scale-up is an indirect approach in which we ask a random sample of respondents about prevalence of different risky behaviors in their network [10, 14]. This method assumes that prevalence of risky behaviors among network of respondents is proportional to that of the population [13, 15].

As explained, in network scale-up approach, people reply on behalf of their network rather than themselves. This decreases the level of sensitivity and encourages respondents to participate in the study. While this method paves the way for estimation of size of several hidden groups in one single study, depending on the cultural issues, still some respondents might refuse to answer some parts of questionnaires. This causes the problem of incomplete questionnaires.

Problems caused by missing data are well addressed in the literature. Analysis of cases with available data (complete case (CC) analysis) is the most frequently applied method. However, it is known that omission of missing data results in power reduction and imprecise estimates. Other frequently applied methods include replacement with a fixed value such as median, regression imputation, expectation algorithm (EM), and multiple imputation [3, 8, 17, 18].

Almost all papers published address problems of missing data in setting of standard statistical methods such as longitudinal data, regression modeling, or survival analysis [1, 4, 20]. In other words, attention was mostly on magnitude of bias in assessment of correlation between variables, estimation of regression coefficients and its precision (standard error), or performance of prediction models.

Less attention, if any at all, was paid to the problem of incomplete questionnaires in the context of size estimation. In this manuscript, we addressed the impact of missing data on estimation of size of hidden groups. We then explored whether imputation methods can recover the data to provide unbiased estimates.

Methods

In response to the request of the Iranian Ministry of Health and Medical Education, a national-wide study was implemented to estimate the size and prevalence of misuse of different types of drugs in Iran. In street-based gender-match interviews, we approached pedestrians who walked alone and explained the aims of the study to them. Only those who verbally consented to participate in the study were recruited. Questions about misuse of 10 different types of drug were asked. For example, we asked respondents ‘how many people you know who misused opium at least once in the last year?’ Details and process of data collection have been addressed elsewhere [13]. Following the network scale-up methodology, the prevalence at national and province level is estimated and submitted to the Iranian Ministry of Health and Medical Education (confidential data). This study was approved in research chancellor of Kerman University of Medical Sciences, with code of ethical of 163/90/KA.

Data preparation

In this manuscript, we only used data collected in one area, with sample size around 1000. For each drug, in turn, we randomly deleted 10% of the replies. To take into account the sampling variation, this has been repeated 20 times; thus creating 200 data sets. To address the impact of missing rate, these steps were repeated at 30% and 50% nonresponse rates.

Imputation methods

To explore the behavior of different imputation methods, five methods were applied: two ad hoc (CC and median replacement (MED)) and three likelihood based methods (regression imputation, negative binomial regression (NB), and EM). NB was applied as responses to sensitive questions usually exhibit a skewed distribution. In the CC method, questionnaires included missing data were ignored. In the MED data sets, missing data for each variable were replaced by median value of response to the same question. In three likelihood based methods, variable with missing data was considered as dependent, and the other nine variables plus demographic characteristics [9] were served as predictors.

Therefore, fifteen compositions were studied by changing missing rate (3 rates) and imputation method (five methods), with 200 data sets at each (3000 data sets in total). This was followed by prediction of frequency of misuse of drugs in all data sets.

Comparison of methods

Results submitted to the Iranian Ministry of Health and Medical Education were served as real values. The relative bias was defined as the difference between predicted and real values divided by the real values. The averages of positive and negative relative biases were considered as average overestimation and underestimation, respectively. In addition, we defined two binary variables named positive severe relative bias (SRB+) and negative severe relative bias (SRB–). To define SRB+, we recode relative biases between 0 and 10% to 0 and higher values to 1. To create SRB–, relative biases between 0 to –10% were coded to 0, and lower values to 1.

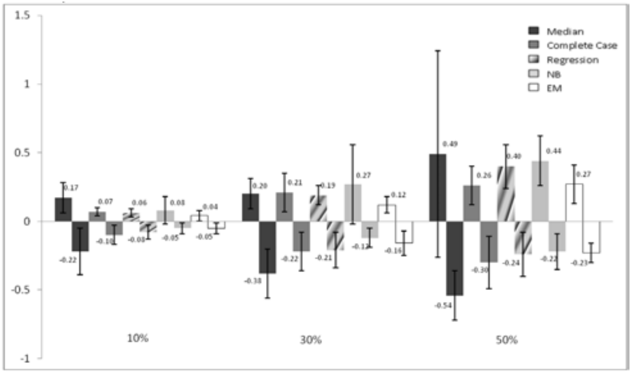
At each of the fifteen different compositions, we calculated average of overestimation and underestimation of relative bias, and prevalence of SRB. Finally, impact of missing rate and imputation method on creation of SRB+ and SRB– were evaluated though multifactorial logistic regression model. *P*-value < 5% was considered as being significance. All *P*-values were two sided. All analyses were applied in STATA and Excel software packages.

Results

Comparison of overestimation and underestimation of average relative bias

In 38 out of 3000 data sets, relative bias was exactly zero. Overestimation and underestimation were occurred in 1497 and 1465 data sets. Frequency of overestimation ranged 27 (MED at 50% rate) to 138 (NB at 50% rate). Corresponding figures for underestimation were 61 (NB at 50% rate) and 173 (MED 50%).

Overestimation at 10% *missing rate*. The MED exhibited the poorest performance with overestimation of 17%, which was at least two times higher than other methods (Figure 1). The EM had the lowest overestimation at 4%. Marginal difference was seen between regression, CC and NB.



Figures show mean of average relative bias. Arrows show standard deviation. Up and down bars show mean of overestimation and underestimation of relative bias, respectively.

Figure 1. Comparison of mean (standard deviation) of relative bias at different missing rates and imputation methods.

Underestimation at 10% *missing rate*. Performance of EM and NB was better than CC (–5% versus –10%), (Figure 1). The MED produced poor results; where in average –22% underestimation was occurred.

Overestimation at 30% *missing rate*. Performance of CC, MED and regression imputation methods was almost the same (about 20% overestimation). EM method was still better than other approaches (Figure 1). NB produced the highest overestimation at 27%.

Underestimation at 30% *missing rate*. The MED revealed the worst estimate at –38%. EM and NB provided the lowest underestimation (–16% and –12%). Regression imputation and CC were almost the same with around –20% underestimation.

Overestimation at 50% *missing rate*. At 50% missing rate, absolute relative bias across all imputation methods was higher than 20%.

Overestimation in MED and NB was higher than other methods. Lowest overestimation observed applying either CC or EM (26% and 27%, respectively).

Underestimation at 50% *missing rate*. MED exhibited the poorest performance with underestimation level around two times higher than other approaches. No remarkable difference was seen between NB, EM and regression.

Comparison of Severe Relative Bias (SRB)

For MED, regardless of missing rate, majority of SRBs happened was SRB–. However, majority of SRBs seen in regression and NB was SRB+ (Figure 2 top panel). In the case of CC and EM, contribution of SRB+ and SRB– in formation of SRB was equal. The EM was prone to SRB+ only at 10% missing rate (Figure 2 top panel).

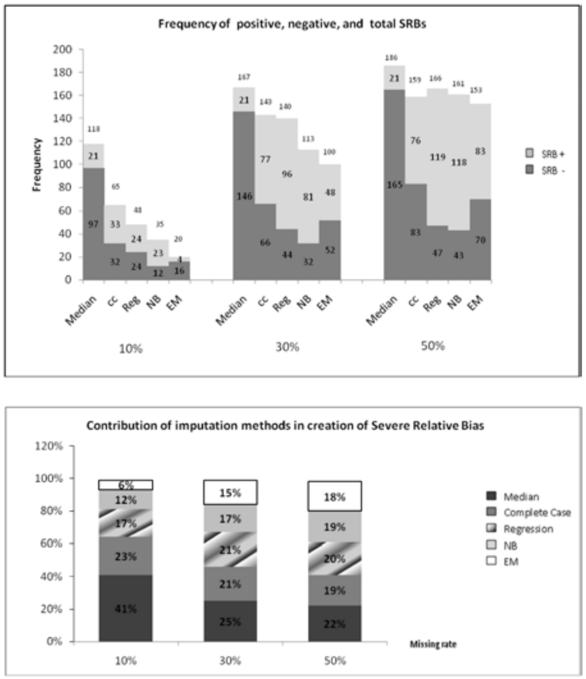
Missing rate 10%. We have seen SRB in 118 and 65 out of 200 data sets analyzed by MED and CC models (Figure 2, top panel). Furthermore, about one forth of samples (48 out of 200) analyses by regression led to SRB. Corresponding figure for EM and NB was about one third and half that of CC (20 and 35 versus 65).

The SRB was occurred in 286 of 1000 data sets analyzed. Contributions of MED and CC were 41% and 23%, respectively (Figure 2 bottom panel). Contribution of NB was twice that of EM (12% versus 6%).

Missing rate 30%. Here 167 and 100 data sets analyzed by MED and EM analysis led to SRBs (corresponded to 84% and 50% of data sets). Performance of regression and CC was the same. Performance of NB was slightly poorer than EM.

In 66% of data sets SRB was happened where largest and smallest contributions were corresponded to MED and EM, respectively (25% versus 15%). Difference between NB and EM was only 2 percentage points (17% versus 15%).

Missing rate 50%. The frequency of SRB observed was almost the same in all methods, but not MED (Figure 2 top panel). In total, 825 scenarios resulted in SRB. Contribution of different imputation methods was around 20% (Figure 2 bottom panel). Highest percentage point difference was seen between MED and EM at 4% (22% versus 18%).

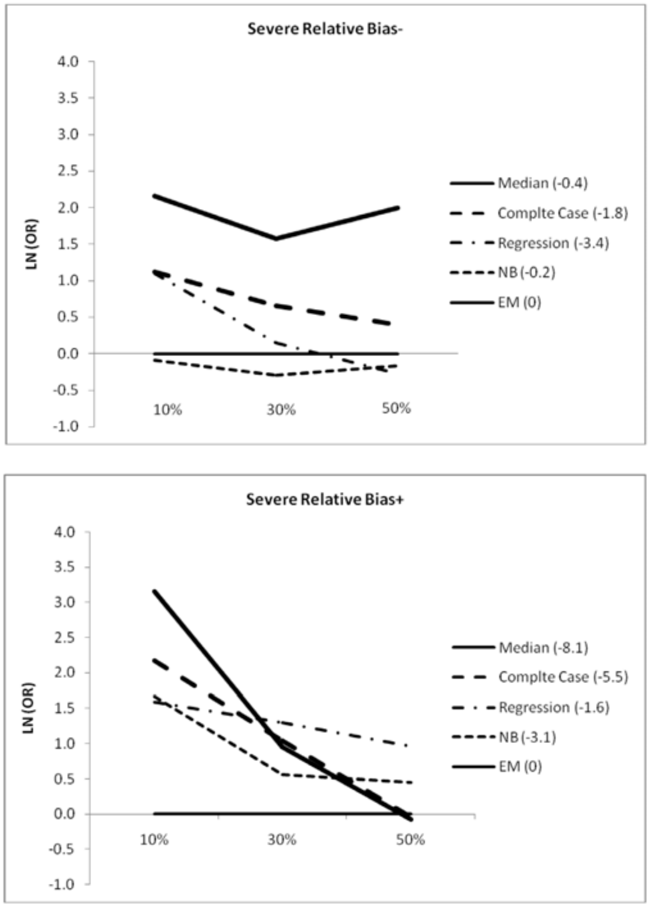


In the bottom panel, bars show percentage of SRBs by the method of imputation at different missing rate. For example at 10% missing rate for median approach $0.41 = 118/(118 + 65 + 48 + 35 + 20)$.

Figure 2. Comparison of frequency of severe relative bias (top panel) and contribution of each method to create it (bottom panel).

Logistic regression modeling

In the multifactorial logistic regression model, significant interaction between missing rate and method of imputation was seen (Figure 3). EM was considered as the reference group. In terms of SRB+, the patterns of changes in logarithm of OR, across all imputation methods were similar. Regression of logarithm of OR versus missing rate revealed that the regression imputation and MED methods exhibited smallest and largest slopes, respectively (–1.6 versus –8.1). This indicates that, by increase in missing rate from, superiority of EM over MED reduced sharper than other methods. The slopes for CC and NB were –5.5 and –3.1, respectively.



Figures in parenthesis show the slope of the regression line of logarithm of odds ratio (OR) versus missing rate.

Figure 3. Assessment of presence of interaction between missing rate and imputation method in creation of negative or positive severe relative biases (SRB).

With respect to SRB–, change in logarithm of OR in MED and NB methods was different with the other three methods where a V shape pattern was seen. Largest and smallest slopes were observed in the case of regression (–3.4) and NB (–0.2). We observed that increase in missing rate, does not change OR of NB, but remarkably decreases OR of regression imputation.

SRB+. At 10% missing rate, the performance of CC and NB methods was poorer than EM $(P < 0.01)$, (Table 1). The CC and NB methods were, respectively, 8.7 (95% confidence interval: 3.0, 25.7) and 5.3 (95% confidence interval: 1.8, 15.9) times more likely to produce SRB+ estimates. Poorest and best performances were seen in the case of MED $(\text{OR} = 23.4)$ and regression $(\text{OR} = 4.9)$.

At 30% missing rate, poorest and best performances were seen in the case of regression $(\text{OR} = 3.7)$ and NB $(\text{OR} = 1.8)$. Relative to 10% results, the superiority of EM imputation over MED, CC and NB reduced dramatically. The EM still kept its superiority over all methods. The NB was about 75% more likely to produce SRB+ estimates which was of marginal significant $(P = 0.04)$.

At 50% missing rate, only performance of regression was poorer than EM $(\text{OR} = 2.6, P = 0.01)$. Performances of CC, MED and NB were not significantly different with EM.

SRB–. Here NB and MED showed the best and worst performances, regardless of missing rate (Table 2). In all missing rates, NB was less likely to produce SRB– but this was not of statistical significance. MED was always poor. However, while in other methods increase in missing rate was associated with decrease in OR, in the case of MED sharp decrease followed by a sharp increase was seen. Regression was poorer than EM only in 10% missing rate $(\text{OR} = 3.0, P = 0.004)$. The CC was poorer than EM in 10% $(\text{OR} = 3.1, P = 0.002)$ and 30% $(\text{OR} = 1.9, P = 0.03)$ missing rates.

Table 1. Assessment of impact of imputation method and missing rate on the risk of observing positive severe relative bias

Missing rate	Method	*N*	OR	Confidence interval	*P*-value
10%	Median	21	23.4	7.2, 75.6	< 0.001
	Complete case	33	8.7	3.0, 25.7	< 0.001
	Regression	24	4.9	1.6, 14.6	0.004
	NB	23	5.3	1.8, 15.9	0.003
	EM	4	REF		
30%	Median	21	2.6	1.1, 6.2	0.03
	Complete case	77	2.8	1.6, 5.0	< 0.001
	Regression	96	3.7	2.1, 6.5	< 0.001
	NB	81	1.8	1.0, 3.0	0.04
	EM	48	REF		
50%	Median	21	0.9	0.3, 2.6	0.89
	Complete case	76	1.0	0.5, 1.9	0.90
	Regression	119	2.6	1.2, 5.6	0.01
	NB	118	1.6	0.8, 3.1	0.19
	EM	83	REF		

N shows frequency of SRB

Abbreviations:

OR: Odds ratio

NB: Negative binomial regression

EM: Expectation maximum algorithm

Table 2. Assessment of impact of imputation method and missing rate on the risk of observing negative severe relative bias

Missing rate	Method	*N*	OR	Confidence interval	*P*-value
10%	Median	97	8.6	4.6, 16.2	< 0.001
	Complete case	32	3.1	1.5, 6.1	0.002
	Regression	24	3.0	1.4, 6.3	0.004
	NB	12	0.9	0.4, 2.1	0.83
	EM	16	REF		
30%	Median	146	4.8	2.6, 8.7	< 0.001
	Complete case	66	1.9	1.1, 3.5	0.03
	Regression	44	1.2	0.6, 2.2	0.65
	NB	32	0.7	0.4, 1.4	0.36
	EM	52	REF		
50%	Median	165	7.4	3.2, 17.1	< 0.001
	Complete case	83	1.5	0.8, 2.9	0.25
	Regression	47	0.8	0.4, 1.5	0.44
	NB	43	0.9	0.4, 1.7	0.66
	EM	70	REF		

N shows frequency of SRB

Abbreviations:

OR: Odds ratio

NB: Negative binomial regression

EM: Expectation maximum algorithm

Discussion

Estimation of size of some hidden populations such as those who misuse drug or alcohol is important for policy making and management of resources. Two main sources of bias in estimation of size of MARP groups are visibility and popularity [15]. Approaches to calculate correction factors to tackle these biases are proposed [16]. However, one issue which was of less concern was incomplete questionnaires. Our results demonstrate level of bias in size estimation in low, moderate and high missing rates. In addition, we showed whether imputation of missing data can recover the data to avoid biased estimates.

We observed that at 10% missing rate, relative bias in regression and CC models was about 1.6 and 2 times higher than EM. Corresponding figure for MED method was more than 4. This ratio for NB was 1 and 2 in the case of underestimation and overestimation.

In addition, ratio of number of times MED and CC suffered SRB was 6 and 3 times higher than EM (118 and 65 versus 20). This ratio for NB was 1.75.

When we randomly deleted 30% of data, we observed that the overestimation and underestimation were much higher than that of 10% missing rate. However, the rates of increase for different imputation methods were not the same. The MED was affected much lower than other methods. For example, in all methods, these statistics were increased by a factor of between 2.5 to 3.5. This ratio in MED is 1.73 for underestimation and 1.18 for overestimation.

Ratio of frequency of SRB at 30% missing rate over 10%, in EM was around 1.5 times higher than regression and NB (5 in EM, 2.92 in regression and 3.23 in NB). Corresponding ratios for CC and MED were about 2.3 and 3.5, respectively. All these evidences suggest that EM was still the best imputation method. However, its superiority over other methods was diminished.

At 50% missing rate, in terms of relative bias, MED was remarkably poorer than other three methods. However, contribution of all methods in creation of SRB was about 20%.

In the multifactorial regression analysis, comparing with EM, in terms on SRB–, NB and MED were always best and worst options. However, in terms of SRB+, the picture was vague. For example, regression showed the best performance at 10% missing rate but the worst at higher missing rates. The opposite was true in the case of MED.

A search of literature reveals huge number of papers that address the impact of imputation methods at different missing rates. Results of papers that compared performance of imputation methods at different missing rates are controversial.

Some studies recommended use of ad hoc methods at low missing rate [4, 20]. For example, Asia Pacific Cohort Studies Collaboration aimed to determine the risk factors for coronary heart disease. The ability of imputation methods and CC analysis to handle the missing data on a single variable (cholesterol) was assessed. In 22 studies in which the missing rate was about 10%, the CC and imputation methods gave similar results [5].

As another example, data for 398 cases with suspected pulmonary embolism were available of which 246 participants (62%) had complete information on all 26 variables studied. Rates of missing values were as follows: 0% for 12 variables, < 10% for 11 variables, 14% for 1 variable and 21% for 2 variables. Results of multiple imputation were comparable with mean replacement in terms of magnitude of estimated coefficients, direction of association, and discrimination ability [21].

On the other hand, there are published papers where at low missing rate performance of CC or median replacement is not comparable to modern imputation methods [2, 20]. To address this issue, missing data are generated on one variable at 10%, 30% and 50% rates. The sample size was 100. At 10% missing rate, the MSE corresponded to EM analysis was minimum 246 times out of 500 iterations. Corresponding figures for regression and CC analysis are 202 and 52, respectively. Results were also the same at higher missing rates [20].

As another example, in a simulation study, true odds ratio (OR) between a diagnostic test and disease status was generated at 3. Then diagnostic test for 20% of diseased and non-diseased subjects was omitted. Replacement of missing data by overall mean led to OR of 1.73 [6].

Generally speaking, ad hoc methods should not be applied when missing rate is high. However, their performance is usually satisfying at low missing rate. On the other hand, EM algorithm has established its role as one of the best imputation methods [4]. Superiority of EM method over easier methods such as regression imputation, median replacement, or CC method has been addressed extensively [3, 8, 20].

Our results indicate that at, in low and moderate missing rates and in the case of SRB+, EM imputation is superior to other methods and can partially reduce the level of bias in size estimation practice. In addition, at high missing rate, performance of none of methods was satisfying.

When it comes to SRB–, MED was always the worst option. However, the superiority of EM over other methods was vanished at moderate missing rate.

These together suggest the usefulness of EM when missing rate is low, and un-usefulness of none of them at high missing rate. This was in contrast with lessons so far in the domain of missing data, which in general appreciated performance of ad hoc methods at low missing rate, and recommended application of modern methods at moderate and high missing rates.

Prevalence of misuse of different types of drugs is low. One of the weaknesses of our study we did not compare behavior of methods for rare to common prevalence. Therefore, the issue of impact of prevalence remains to be addressed. We are conducting simulation studies to investigate the behavior of imputation methods at different missing rates, for rare to common prevalence. Another limitation of our study was that all size estimation analyses (both real estimates and predictions) were based on frequency approach. In NSU practice, we asked respondents about number of people they know with risky behaviors. Therefore, the answer is a count.

Another approach would be to ask whether respondents know at least one person with risky behavior (yes, no question). Basic formulas for size estimation based on count or yes/no questions are different.

Another issue is that we deleted the data randomly. However, in the case of sensitive issues, nonresponse necessarily might not be random. In addition, our results showed behavior of our models in terms of SRB+ and SRB– was not comparable. Therefore, future work is needed to compare performance of frequency versus probability size estimation methods, in terms of SRB+ and SRB–, and to address the impact of missing mechanism.

Besides these limitations, to our best knowledge, this was the first study that highlights problems of incomplete questionnaires in setting of size estimation. In addition to that, we compared performance of frequently used imputation methods at low, medium and high missing rates. At each composition, we analyzed 200 data sets to take into account the sampling variation. We provided results that pave the way for future exploration in this field.

Generally speaking, results from literature suggest that, at low missing rate choice of imputation method is not of great concern [7]. On the other hand, modern imputation methods find room to show their ability when missing rate increases [2, 3]. However, our results present a different picture. We observed that median substitution was the poorest method. At 10% missing rate, EM imputation was superior and partially reduced bias. At moderate missing rate, performance of none of methods was satisfying. These results suggest that deep considerations should be taken into account to avoid incomplete questionnaires. Even low nonresponse can lead to biased estimates, which can partially be tackled by implementation of modern imputation method such as EM.

Acknowledgment

MRB and AAH designed the study. All authors contributed in analysis plan. Analyses are performed by ND and MRB. All authors have read and approved the final manuscript.

References

[1] M. Baneshi and A. Talei, Multiple imputation in survival models: applied on breast cancer data, Iranian Red Crescent Medical Journal 13(8) (2011), 547-552.

[2] M. R. Baneshi, Prevention of disease complications through diagnostic models: how to tackle the problem of missing data? Iranian Journal of Public Health 41(1) (2012), 66-72.

[3] M. R. Baneshi and A. Talei, Does the missing data imputation method affect the composition and performance of prognostic models? Iranian Red Crescent Medical Journal 14(1) (2012), 31-36.

[4] M. R. Baneshi and A. Talei, Impact of imputation of missing data on estimation of survival rates: an example in breast cancer, Iranian Journal of Cancer Prevention 3(3) (2012), 127-131.

[5] F. Barzi and M. Woodward, Imputations of missing values in practice: results from imputations of serum cholesterol in 28 cohort studies, American Journal of Epidemiology 160(1) (2004), 34-45.

[6] A. R. T. Donders, G. J. van der Heijden, T. Stijnen and K. G. Moons, Review: a gentle introduction to imputation of missing values, Journal of Clinical Epidemiology 59(10) (2006), 1087-1091.

[7] D. L. Fairclough, Patient reported outcomes as endpoints in medical research, Stat. Methods Med. Res. 13(2) (2004), 115-138.

[8] S. Haji-Maghsoudi, A. A. Haghdoost, A. Rastegari and M. R. Baneshi, Influence of pattern of missing data on performance of imputation methods: an example using national data on drug injection in prisons, International Journal of Health Policy and Management 1 (2013), 1-11.

[9] M. Jalali, A. Nikfarjam, A. A. Haghdoost, N. Memaryan, T. Tarjoman and M. R. Baneshi, Social hidden groups size analyzing: application of count regression models for excess zeros, Journal of Research in Health Sciences 13(2) (2013), 143-150.

[10] C. McCarty, P. D. Killworth, H. R. Bernard, E. C. Johnsen and G. A. Shelley, Comparing two methods for estimating network size, Human Organization 60(1) (2001), 28-39.

[11] M. Nasirian, F. Doroudi, M. M. Gooya, A. Sedaghat and A. A. Haghdoost, Modeling of human immunodeficiency virus modes of transmission in Iran, Journal of Research in Health Sciences 12(2) (2012), 81-87.

[12] S. Navadeh, A. Mirzazadeh, L. Mousavi, A. A. Haghdoost, N. Fahimfar and A. Sedaghat, HIV, HSV2 and Syphilis prevalence in female sex workers in Kerman, South-East Iran; using respondent-driven sampling, Iranian Journal of Public Health 41(12) (2012), 60-65.

[13] A. Rastegari, S. Haji-Maghsoudi, A. Haghdoost, M. Shatti, T. Tarjoman and M. R. Baneshi, The estimation of active social network size of the Iranian population, Glob. J. Health Sci. 5(4) (2013), 217-227.

[14] H. Russell Bernard, E. C. Johnsen, P. D. Killworth and S. Robinson, Estimating the size of an average personal network and of an event subpopulation: some empirical results, Social Science Research 20(2) (1991), 109-121.

[15] M. J. Salganik, D. Fazito, N. Bertoni, A. H. Abdo, M. B. Mello and F. I. Bastos, Assessing network scale-up estimates for groups most at risk of HIV/AIDS: evidence from a multiple-method study of heavy drug users in Curitiba, Brazil, American Journal of Epidemiology 174(10) (2011), 1190-1196.

[16] M. J. Salganik, M. B. Mello, A. H. Abdo, N. Bertoni, D. Fazito and F. I. Bastos, The game of contacts: estimating the social visibility of groups, Social Networks 33(1) (2011), 70-78.

[17] J. Schafer, Multiple imputation: a primer, Stat. Methods Med. Res. 8(1) (1999), 3-15.

[18] J. L. Schafer and M. K. Olsen, Multiple imputation for multivariate missing-data problems: a data analyst's perspective, Multivariate Behavioral Research 33(4) (1998), 545-571.

[19] S. Snidero, F. Zobec, P. Berchialla, R. Corradetti and D. Gregori, Question order and interviewer effects in CATI scale-up surveys, Sociol. Methods Res. 38(2) (2009), 287-305.

[20] S. R. C. Suraphee, J. Busaba, C. Chaisorn and W. Nakornthai, A comparison of estimation methods for missing data in multiple linear regression with two independent variables, Thail. Stat. 4 (2006), 13-26.

[21] G. J. van der Heijden, A. R. T. Donders, T. Stijnen and K. G. Moons, Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example, Journal of Clinical Epidemiology 59(10) (2006), 1102-1109.