Iranian J Publ Health, Vol. 41, No.1, 2012, pp. 66-72

Original Article

Prevention of Disease Complications through Diagnostic Models: How to Tackle the Problem of Missing Data?

*MR Baneshi [1] , H Faramarzi [2] , *M Marzban [3,4]*

[1]Reserch Center for Modeling in Health, Kerman University of Medical Sciences, Kerman, Iran
[2]Shiraz HIV/AIDS Research Center, Shiraz University of Medical Sciences, Shiraz
[3]Research Center for Traditional Medicine and History of Medicine, Shiraz University of Medical Sciences, Shiraz, Iran
[4]Dept. of Biostatistics and Epidemiology, Kerman University of Medical Sciences, Kerman, Iran

(Received 10 Feb 2011; accepted 23 Nov 2011)

Abstract

Background: Diagnostic models are frequently used to assess the role of risk factors on disease complications, and therefore to avoid them. Missing data is an issue that challenges the model making. The aim of this study was to develop a diagnostic model to predict death in HIV/ AIDS patients when missing data exist.

Methods: HIV patients (n=1460) referred to Voluntary Consoling and Testing Center (VCT) of Shiraz southern Iran during 2004-2009 were recruited. Univariate association between variables and death was assessed. Only variables which had univariate *P*< 0.25 were selected to be offered to the Multifactorial models. First, patients with missing data on candidate variables were deleted (C-C model). Then, applying Multivariable Imputation via Chained Equations (MICE), missing data were imputed. Logistic regression was fitted to C-C and imputed data sets (MICE model). Models were compared in terms of number of variables retained in the final model, width of confidence intervals, and discrimination ability.

Result: About 22% of data were lost in C-C model. Number of variables retained in the C-C and MICE models was 2 and 6 respectively. Confidence Intervals (C.I.) corresponding to C-C model was wider than that of MICE. The MICE model showed greater discrimination ability than C-C model (70% versus 64%).

Conclusion: The C-C analysis resulted to loss of power and wide CI's. Once missing data were imputed, more variables reached significance level and C.I.'s were narrower. Therefore, we do recommend the application of the imputation method for handling missing data.

Keywords: HIV/AIDS, Missing Data, Imputation, MICE

Introduction

In epidemiological studies, researchers usually try to assess the associations between several predictors and an outcome variable using multivariable regression techniques. Once such predictors are identified, through diagnostic models, they will be combined to get a risk score. The estimated risk scores are then converted to estimate the risk of a disease related event. In other words, diagnostic models enable clinicians to estimate the risk of a disease-related event (such as death due to HIV/ AIDS) by combinations of multiple predictor values. This helps the management of future patients.

AIDS is a disease of the human immune system caused by the human immunodeficiency virus. AIDS is now a pandemic (1). In 2008, an estimated 35,000 people in the Middle East and North Africa became infected with HIV, and 20,000 AIDS-related deaths occurred (2). Until

***Corresponding Author:** Tel: +98 2123 37 589, E-mail address: marzbanh@gmail.com

this year 15% of people who diagnosis with HIV/AIDS in Iran died.

Therefore, identification of risk factors to be used so as to predict risk of death for patients is necessary. This makes physicians enable to identify high risk patients who need special care and treatment.

However, one of the issues that challenge the process of model development is missing data. Missing data is a common problem in clinical data sets (3). In the content of regression modeling, most software by default exclude every subject from the analysis with at least one missing value on any of the predictors or outcome analyzed (know as Complete Case (C-C) analysis), and offer those with available data to the regression model (4). However, usefulness of the C-C analysis depends on the rate and the mechanism of missing data (see rest of the text)**.** In addition these circumstances rarely occur in practice. Here we briefly reviewed the missing data mechanisms. Three main missing mechanism are Missing Completely At Random (MCAR), Missing Not At Random (MNAR), and Missing At Random (MAR) (4).

Missing data are categorized as MCAR when subjects with missing data are a random sample of data (5). For example, MCAR occurs when a blood sample tube is broken or when a questionnaire is accidentally lost. Categorization of missing data as MNAR applies when the probability that an observation is missing is related to unobserved information, such as the actual (missing) value of the variable (5). This can happen, for example, when a patient is so sick that a medical procedure cannot be applied to measure a study variable. However, missing data are usually neither MCAR nor MNAR but MAR (6). MAR applies when the probability that an observation is missing is related to other observed patient characteristics (4).

The C-C method seems reasonable under the MCAR assumption if less than five percent of entire data set is missing (7). However, when missing rate is high, even under MCAR mechanism, exclusion of missing data will diminish precision of estimates (4,8). Under other missing mechanisms C-C method leads to biased estimates (9).

Various methods have been proposed to deal with missing data. There are modern likelihood-based imputation methods (such as Multivariate Imputation via Chained Equations (MICE) which recover the missing data avoiding waste of information.

The aim of this study was to address the impact of exclusion of patients with missing data o model composition, and estimated Confidence Intervals (C.I.). To do this, we used a set of HIV data set as an example

Materials and Methods

Sample and outcome

Making confidential and voluntary counseling and testing (VCT) services available has proven to encourage people to determine their HIV status (2). In this study we used information on 1460 infected person with HIV during 2004-2009, which were referred to VCT center of Shiraz. The main outcome of study was death due to HIV.

Statistical methods

Variables selected

Information on a large umber of variables was available. However, EPV rule recommends that at least ten Events Per independent Variable being tested are required (10). Therefore, candidate variables for the multifactorial modeling were selected through a series of univariate logistic regression analysis. Only variables with univariate *P*-value less than 0.25 were selected to be offered to the multifactorial models (11).

Complete-Case model (C-C model)

In the C-C model, patients with missing data on any of variables selected in the screening round were excluded. A logistic regression model in conjunction with Backward Elimination (B.E.)

variable selection method was then fitted to the data.

Imputation model (MICE model)

Details of the Multivariable Imputation via Chained Equations (MICE) method are presented elsewhere (12). Briefly a series of regression models are run whereby missing value on each variable is predicted upon the other variables in the data. The MICE method takes into account the uncertainty regarding by replace each missing value by multiple imputed values, (usually 10) (6, 13).

It has been noted that by including enough variables in the imputation model (10 to 15) the MAR assumption would be more plausible (14). Therefore, all variables, regardless of their univariate ***P***-value, and patient's outcome were offered to the imputation model (15).

As candidate variables for multifactorial modeling was dichotomous, logistic regression model was used to impute missing data 10 times. Estimates derived from 10 imputed data sets (the coefficients and standard errors) were combined applying Rubin’s rule (16). Then the variable with the highest *P*-value (if >0.05) was excluded, and new coefficients and SE's were estimated for the rest of variables. This process was applied iteratively until all variables remain significant in all imputed data sets (analogous to B.E. variable selection method in C-C analysis) (14). Odds Ratios (OR) and corresponding 95% Confidence Intervals (C.I.) were calculated from regression coefficients and standard errors that have been imputed across multiply imputed data sets.

Comparison of C-C and MICE models

Models developed were compared in terms of variables contributed to the multifactorial models, estimated Standard Errors (S.E.), estimated Odds Ratio's (OR) and associated Confidence Intervals (C.I.). In addition the probability of death was estimated from logistic models developed. This was done for both C-C and imputed data sets. The estimated probabilities were then compared with the observed patients' outcome to compare the models in terms of the Area under ROC Curve (AUC).

Software

Complete-Case model was fitted using SPSS (Version 15) software. Missing data were imputed using MICE package which works under R software.

Results

Descriptive statistics for variables selected are presented in Table 1. Out of 1460 infected cases 283 deaths were observed, giving a death rate of 9.4%. Approximately forth-fifth of cases were from urban areas. Nearly 86% of sample was male. Furthermore, about 75% of subjects had history of injection, and 23% of patients have at least one symptom of HIV. Infection with HCV was as high as 65%**.**

In a screening process, univariate association between 9 variables and death was confirmed at a 0.25 significance level. These variables, which are selected to be offered to the multifactorial model, are listed in Table 1. For each variable, ***P***-value, Odds Ratio (OR) and corresponding CI in univariate logistic regression analysis is presented. In addition, number of patients with available data is reported.

The HCV status of nearly 15% of patients was not available (Table 1). Although missing rate for the rest of candidate variables was less than 1%, by exclusion of missing data, the sample size was reduced to 1134. This indicates loss of about 22% of data.

Out of 9 variables offered to the C-C model only 2 remained significant in the multifactorial analysis (Table 2). Those who had history of prison, relative to those who did no were at higher risk of death with OR= 3.55 (95% C.I: 2.02, 5.73). In addition OR of death for patients with HIV symptoms relative to those without were about 3.17 (95% C.I: 2.32, 4.33).

Once missing data were imputed, 4 more variables retained significant in the multifactorial model (6 in the MICE versus 2 in the C-C models). This was due to the recovery of information and gain in power. These 4 variables were having AIDS syndrome, place of residence, sex, and history of surgery. In particular we found that risk of death for those with AIDS syndromes was about 5 times higher than others (OR (95% C.I.): 4.71 (2.15, 10.30)). Furthermore, being female was associated with more than 55% reduction in the risk of death.

As models were developed in conjunction with B.E. variable selection method, estimated S.E.'s and *P*-values corresponding to the variables reached significance level in the MICE model but not in the C-C model are not reported in the table. This is because those variables were not retained in the final C-C model However, for the sake of comparison, estimated (SE, *P*-value) in the first step of C-C and last step of the MICE models for those variables are given; (0.6, 0.07) and (0.43, <0.001) for having AIDS syndrome; (0.21, 0.16) versus (0.18, 0.02) for the place of residence, (0.47, 0.10) versus (0.37, 0.02) for sex, and (0.23, 0.07) versus (0.21, 0.02) for history of surgery.

As expected, estimated SE's corresponding to the MICE model was always smaller than that of the C-C model. In addition, these four variables were far from being significant in the C-C model but retained in the final model once the MICE imputation method was applied.

We also compared models developed in terms of discrimination ability. AUC for C-C model was 0.64. Corresponding figures for MICE model were 0.70. This indicates nearly 10% improvement in discrimination ability of model after imputation of missing data.

Table 1: Descriptive statistics of variables selected for multifactorial analysis

Variables	category	Frequency (%)	% of cases with available data	Univariate OR (CI)	*P*-value
Sex	male female	86.3 13.2	99.9	0.35 (0.20, 0.59)	0.0001
History of Prison	NO Yes	22.9 77	99.9	2.44 (1.64, 3.62)	0.0001
History of injection	NO Yes	25.4 74.	99.9	1.65 (1.18 , 2.31)	0.003
AIDS Syndrome	NO Yes	97.7 1.8%	99.5	4.71 (2.15, 10.30)	0.0001
With or without HIV	WITH OUT WITH	73.3 22	96.7	2.95 (2.21, 3.91)	0.0001
tuberculosis	POSITIVE NEGATIVE	1.5 97.8	99.3	0.27 (1.52, 8.64)	0.004
HCV	POSITIVE Negative	65.5 18.6	84.1	1.99 (1.33, 3.01)	0.001
history of surgery	NO Yes	82.4 16.8	99.2	0.61 (0.41, 0.91)	0.016
Place of residence	Rural Urban	21.5 78.1	99.6	1.45 (1.04 , 2.04)	0.03

Table 2: Using two kind of model (C-C and MICE) for 5 variables selected

Variable	Category	C-C Model (n=1134)			MICE Model (n=1460)		
		OR (95% C.I.)	*P*-value	S.E.	OR (95% C.I.)	*P*-value	S.E.
With or without HIV symptoms	WITH OUT						
	WITH	3.17 (2.32, 4.33)	<0.001	0.16	3.09 (2.31, 4.15)	<0.0001	0.15
Prison	NO						
	Yes	3.55 (2. 02, 5.73)	<0.001	0.25	2.01 (1.12, 3.62)	0.02	0.27
AIDS Syndrome	NO						
	Yes				4.53 (1.95, 10.51)	<0.001	0.43
Place of residence	Rural						
	Urban				1.52 (1.07, 2.16)	0.02	0.18
Sex	male						
	female				0.43 (0.21, 0.89)	0.02	0.37
Surgery	NO						
	Yes				0.60 (0.40, 0.91)	0.02	0.21

Discussion

Diagnostic models provide valuable information regarding contribution of variables on clinical outcome (17). However, presence of missing data makes exercise of model building difficult. When missing data exist, and excluded, power would be lost. Consequently, variables with real predictive ability may not contribute to the multifactorial model. Therefore, imputation is necessary. Imputation of missing data, relative to C-C analysis, leads to gain in power.

Although understanding of the factors associated with HIV/ AIDS death is vital, it should be noted that our main goal was not to develop the best possible model to predict death due to HIV/ AIDS. We did not have information of some known and important risk factors such as CD4 level. The main goal of our study was to demonstrate the impact of exclusion of missing data on model composition and its performance. It has been emphasize that exclusion of cases with missing data resulted in the loss of precision and a wider confidence interval. This was the case in our analysis. In the C-C model only 2 variables were retained in the model. However, after imputation of missing data, and gain in power, 4 more variables were survived in the model (6 variables in total). We have seen that estimated S.E.'s in the MICE model were narrower which was due to recovery of information.

Lots of important risk factors of death due to HIV/ AIDS are known. Some studies find a strong association between AIDS death rates and positive history of primary and secondary syphilis rates among men (18). Concurrent HIV/AIDS was associated with more than twice the risk of HIV-related death within the 4 months after diagnosis (19). In addition, injection drug use was known to be associated with a substantially increased risk of morbidity and mortality (20, 21).

We found that patients who have symptom of HIV or AIDS syndrome are more at risk than others. Regarding the place of residence, although metropolitan become more infected with HIV but rustic patient are at higher risk of death, possibly due to lack of facilities in rural areas. We also found that those used methadone, were at lower risk of death. This might be due to the fact that use of methadone, in comparison with illegal drugs, can increase longevity and compliance them to referred VCT for more care and decrease high risk behavioral on addicted patients.

To address the impact of missing data on estimation of the treatment effect, a randomized clinical trial of antiretroviral therapy for HIV-infected individuals were carried out (22). It has been shown that exclusion of missing data resulted in the underestimation of the true treatment effect. This was due to selective dropout of participants with lower or decreasing CD4 counts.

Another approach frequently applied to tackle missing data is to perform missing indicator analysis. That is to put subjects with missing data into a separate category. This method has been applied analyzing data of health care workers who exposed with HIV-infected blood (23). Results of C-C and missing indicator were similar in terms of variables contributed to the models. However, estimated confidence intervals in the C-C model were wider. Results were not compared with that of the MICE model.

In another study to identify factors associated with non adherence during the maintenance phase of HAART significant difference between results of the C-C model and model in which missing data were imputed was seen (24). Although disadvantages on the C-C analysis are known, majority of studies prefer to exclude missing data (25), possibly due to its simplicity. The MICE model, on the other had, is a flexible approach to generate multivariate multiple imputations. We do recommend application of modern imputation methods before development of diagnostic models.

Ethical considerations

Ethical issues (Including plagiarism, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy, etc) have been completely observed by the authors. However, because of the anonymous of data we did not fill any inform consent or the Ethics Committee did not approve the study.

Acknowledgment

The data set analyzed in this project was collected from VCT Center of Shiraz under the direction of Shiraz University of Medical Sciences. We should thank their employee who facilitated our access to the data set. Also thank committee of research center of student in Kerman University of medical sciences. The authors declare that there is no conflict of interests.

References

1. Kallings LO (2008). The first postmodern pandemic: 25 years of HIV/AIDS. *J Intern Med* ,263 (3): 218–43.
2. Anonymous (2011).AIDS epidemic update. UNAIDS.Available from*: www.unaids.org*
3. Burton A, Altman DG (2004). Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines.*BJC*, 91(1): 4-8.
4. Altman DG, Bland JM (2007). Missing data. *BMJ*,334(7590): 424.
5. Donders A R, van der Heijden T, Gjmg Stijnen T, Moons K G M (2006). Review: a gentle introduction to imputation of missing values. *JCE*, 59(10): 1087-1091.
6. Schafer JL (1999). Multiple imputations: a primer. *Stat Methods Med Res*, 8(1): 3.
7. Fairclough DL (2004). Patient reported outcomes as endpoints in medical research. *Stat Methods Med Res*, 13(2): 115.
8. Joseph L, Bélisle P, Tamim H, Sampalis J (2004). Selection bias found in interpreting analyses with missing data for the pre hospital index for trauma. *JCE*, 57(2): 147-153.

9. Baneshi MR, Talei AR (2011). Impact of Imputation of Missing Data on Estimation of Survival Rates: An Example in Breast Cancer.*IJCP*,3(3): 127-31.
10. Peduzzi P,Concato J, Feinstein A R, Holford T R (1995). Importance of events per independent variable in proportional hazards regression analysis II. Accuracy and precision of regression estimates. *JCE*, 48(12): 1503-1510.
11. Baneshi MR (2011). Modeling of many skewed biomarkers and missing data: An Example in Breast Cancer.1st ed. *Lambert academic publishing* ,UK .
12. Baneshi MR, Talei AR .Multiple Imputation in Survival Models: Applied on Breast Cancer. *IRCMJ*, 13 (8): 544-549.
13. Schafer JL (1997). Analysis of incomplete multivariate data. Chapman & Hall/CRC.
14. Van Buuren S, Boshuizen HC, Knook DL(1999) .Multiple imputation of missing blood pressure covariates in survival analysis. *Stat Med*, 18(6): 681-694.
15. Moons KG, Donders M, Rart Stijnen T, Harrell F E (2006). Using the outcome for imputation of missing predictor values was preferred. *JCE*, 59(10): 1092-1101.
16. Rubin DB (2004). Multiple imputation for nonresponse in surveys. *Wiley-IEEE*.
17. Bacchetti P (1995). Historical assessment of some specific methods for projecting the AIDS epidemic. *Am J Epidemiol*, 141(8): 776.
18. Chesson HW, Dee TS, Aral SO (1990s). AIDS mortality may have contributed to the decline in syphilis rates in the United States in the.*Sex Transm Dis* , 30(5): 419.
19. Hanna D B, Pfeiffer MR, Torian L V, Sackoff J E (2008). Concurrent HIV/AIDS diagnosis increases the risk of short-term HIV-related death among persons newly diagnosed with AIDS 2002-2005. *AIDS Patients t Care STD*,22(1): 17-28.
20. Bewley TH , Ben-Arie O, Marks V (1968). Morbidity and mortality from heroin dependence. 3. Relation of hepatitis to self-injection techniques.*BMJ British medical journal*,1(5594): 730.
21. Tyndall M W, Craib K J P, Currie SLi K, O'Shaughnessy M V, Schechter M T (2001). Impact of HIV infection on mortality in a cohort of injection drug users. *JAIDS*, 28(4): 351.
22. Raboud J M , Montaner J S G, Thorne A, Singer J, Schechter M T(1996). Impact of missing data due to dropouts on estimates of the treatment effect in a randomized trial of antiretroviral therapy for HIV-infected individuals. *JAIDS J*, 12(1): 46.
23. Cardo D M, Culver DH, Ciesielski C A, Srivastava P U, Marcus R, Abiteboul D et al. (1997). A case-control study of HIV seroconversion in health care workers after percutaneous exposure. *NEJM*, 337(21): 1485.
24. Carrieri M P, Leport C, Protopopescu C, Cassuto J P, Bouvet E, Peyramond D, et al (2006). Factors associated with nonadherence to highly active antiretroviral therapy: a 5-year follow-up analysis with correction for the bias induced by missing data in the treatment maintenance phase. *JAIDS*, 41(4): 477.
25. Tokars J I, Marcus R, Culver D- H, Schable C A, McKibben P S, Bandea CI, et al. (1993) . Surveillance of HIV infection and zidovudine use among health care workers after occupational exposure to HIV-infected blood.*Ann Intern Med*, 118(12): 913.