Iranian Red Crescent Medical Journal

ORIGINAL ARTICLE

Multiple Imputation in Survival Models: Applied on Breast Cancer Data

MR Baneshi [1*], A Talei [2]

[1] Research Center for Modeling in Health, Kerman University of Medical Sciences, Kerman, IR Iran
[2] Shahid Faghihi Hospital, Shiraz University of Medical Sciences, Shiraz, IR Iran

▸ *Please cite this paper as:*
Baneshi MR, Talei A. Multiple Imputation in Survival Models: Applied on Breast Cancer Data. *Iran Red Crescent Med J*. 2011;**13**(8):547-52.

Abstract

Background: Missing data is a common problem in cancer research. Although simple methods, such as complete-case (C-C) analysis, are commonly employed to deal with this problem, several studies have shown that such methods lead to biased estimates. The aim of this study was to address the issues encountered in the development of a prognostic model when missing data exist.

Patients and Methods: A total of 310 breast cancer patients were recruited. Initially, the patients with missing data for any of the 4 candidate variables were excluded. Then, the missing data were imputed 10 times. Cox regression model was fitted to the C-C and imputed data. The results were compared in terms of the variables retained in the model, discrimination ability, and goodness of fit.

Results: In the C-C analysis, some variables lost their significance because of a loss in power, but after imputation of the missing data, these variables reached significant level. The discrimination ability and goodness of fit of the imputed data sets model was higher than those of the C-C model (C-index, 76 % versus 72 %; likelihood ratio test result, 51.19 versus 32.44).

Conclusions: The results indicate the inappropriateness of an ad hoc C-C analysis. This approach leads to loss in power of the variables and imprecise estimates. Application of multiple imputation techniques is recommended for avoiding such problems.

Keywords: Prognostic model; Missing data; Multiple imputation; Breast cancer

Introduction

Prognostic models combine key patient characteristics (risk factors) to predict clinical outcomes such as recurrence of cancer. These models are excellent tools for investigating the contribution of variables to the course of a disease and for selecting the appropriate treatment approach (1). However, if the model assumptions are ignored during its development, the results may be misleading (2,3). One of the challenges in modeling practice is incomplete data. In survival analysis, a problem occurs when data on risk factors are missing (4). The traditional response to this problem is to exclude the individuals with incomplete data for any prognostic factors from the analysis (such an analysis is known as complete-case analysis [C-C analysis]) (4). However, exclusion of missing data leads to reduction in the sample size, which reduces the precision of estimates and can lead to biased estimates (5,6). Therefore, appropriate methods should be applied for imputing missing data. Methodological developments in the analysis of missing data offer a lot to modeling. Advanced likelihood-based methods can be applied to use partial data to predict the missing variables. This approach prevents reduction in sample size and helps avoid biased estimates. Many methods can help tackle the problem of missing data.
The main aim of this study was to highlight the issues encountered in the development of a prognostic model with missing data. Here, we have focused on the Multivariable Imputation via Chained Equations (MICE) method. MICE is a flexible method that has the capability to deal with all forms of variables (continuous, categorical, and binary), and it can be used in regression settings. Te MICE method was applied to analyze a breast cancer data set. To show the information recovery power of the MICE method, prognostic models were developed using complete data as well.

* *Corresponding author at:* Mohammad Reza Baneshi, (PhD), Research Center for Modeling in Health, Kerman University of Medical Sciences, Kerman, IR Iran. Tel: +98-9134423948, Fax: +98-3413205127, *e-mail:* m_baneshi@kmu.ac.ir
Received: 10 January 2011 Accepted: 09 May 2011

Patients and Methods

Patients and outcome

From 1994 to 2003, we collected information of 310 breast cancer patients in Shiraz (located in southern Iran), who had undergone a median follow-up of 2.5 years. The information was collected from the Hospital-based Cancer Registry of Nemazee Hospital (affiliated to Shiraz University of Medical Sciences). The end point of the study was death. At the end of the study, there were 56 deaths.

Variables

Variables included in the multifactorial models were those with univariate predictive ability (tumor stage (early, locally advanced, and advanced), tumor grade (1, 2, and 3), history of benign breast disease (positive versus negative), and age at diagnosis ($\leq$ 47 years versus $>$ 47 years) (7). This data set did not include personal information such as name, address, or phone number of the patients.

Multifactorial models

For data analysis, Kaplan-Meier (KM) analysis and log-rank tests were used to compare the survival curves. The linear Cox model was then applied to develop multifactorial regression models and to estimate Hazard Ratios (HR) (8). Two models were developed using C-C data and imputed data sets. The MICE method was applied to impute the missing data.

C-C model

In the C-C model, patients with missing data on any of the selected 4 variables were excluded. Cox regression model along with the ENTER variable selection method was then applied to patients for whom data on all 4 candidate risk factors was available. The final risk score was calculated by multiplying the variables with the estimated regression coefficient. Tertiles of the estimated risk score were applied as cutoffs to categorize patients into the following risk groups: low (L), intermediate (I), and high (H).

MICE model

The MICE method is a probabilistic approach. The usual practice is to replace each missing value by 10 values, leading to 10 imputed data sets, so as to reflect the uncertainty about the true values of the missing data (9,10). The process of the MICE method is described below.

Identification of the missingness mechanism

To identify the missingness mechanism, we adopted indicator variables for the variables with missing data. For example, the indicator variable for cancer stage variable was used to determine whether information on stage was available. A value of 1 was assigned for patients for whom stage variable was known, whereas a 0 was assigned for those whose stage was unknown. The association between the indicator variables, which reflected missingness, and the rest of the variables was assessed by applying the Chi-square test. When the missingness depends on observed variables mechanism is called Missing At Random (MAR).

Selection of variables for the MICE algorithm

For the best imputation, the outcome variable should be included in the imputation model (11). Therefore, the patients' outcome and a set of 4 risk factors were used in the MICE algorithm.

Specifications of the imputation model

Polytomous and logistic regression analyses were used to impute missing data for categorical (stage and grade) and binary data (age and history of benign disease), respectively.

Imputation process

The MICE method does not involve distributional assumption and can be used to impute missing data for continuous, categorical, and binary variables. To impute missing values for a variable that includes missing data, say x_j, a regression model relates x_j to other variables in the imputation model. This regression model is then used to generate imputation values from the posterior predictive distribution. Each predictor with missing values is considered in turn by using the current imputed value for each of the other predictors (12). The iterative process ends when all variables have been updated (technical details are given in the Appendix) (13). The entire

Table 1. Investigation of the association between missingness indicator variables and the present variables

Missing indicator	Status	Stage	Grade	Benign disease	Age
Stage	+ [a]	-	+ [a]	+ [a]	-
Grade	+	- [b]	-	+ [a]	-
Benign disease	+	- [b]	- [b]	-	+

[a]: Association between the missing variables and the present variables
[b]: Lack of association between the missing variables and the present variables

process was repeated and the imputed values, which were generated in the fifth round, were used as the first imputed data set. The whole process was repeated 10 times to replace each missing data by 10 values, thus creating 10 data sets (12). The standard algorithm imputes each incomplete column in the data set from left to right. The order of the variables is not relevant to the results.

Aggregation of estimates across imputed data sets

The creation of 10 data sets means that 10 modeling analyses are required, 1 for each data set, and therefore, there will be 10 different estimates for each parameter. The estimates derived from the imputed data sets (the coefficients and standard errors) therefore need to be combined; this was achieved by applying Rubin's rule (14). The final regression coefficient is simply the average of the coefficients across imputed data sets. While estimating standard errors, both between and within imputation variations should be taken into account (details are given in the Appendix).

Calculation of hazard ratios and confidence intervals

HR and corresponding 95 % Confidence Intervals (C.I.) were calculated on the basis of the regression coefficients and standard errors that were imputed across multiple imputed data sets.

Calculation of a risk score

A risk score was calculated for each of the 10 imputed data sets. For each patient, a single averaged risk score was calculated by averaging the estimated risk scores from each of the 10 imputed data sets.

Comparison of the Performance of the Models

Discrimination ability

In risk stratification studies, it is important to create risk groups such that the patients in all groups are equally likely to develop the outcome (15). Discrimination is measured using Harrell's C-index (concordance index), which is a generalization of Area Under Curve (AUC) (16,17). The C-index is a measure of correct ordering. It shows the proportion of times, when comparing risk predictions for 2 patients, that the calculated risk for the patient who develops the disease is higher than the calculated risk for the patient who does not develop the disease. The statistical value of the C-index varies between 0.5 and 1, and values near 1 indicate high discrimination power. However, if the performance is assessed using the same sample as that used for model development, then the performance will be overestimated. Therefore, bootstrap procedure was applied to determine bias-corrected C-indices (18).

Goodness of fit (Likelihood ratio test)

For all models, we will report the results of the likelihood ratio test (LRT), which indicates how well the model fits the data.

Table 2. Comparison of estimated hazard ratios 95% confidence intervals in the analysis of complete case and imputed data sets

	Complete-case model (n = 203, D [a] = 54)		**Imputed data sets model** (n = 310, D [a] = 56)	
	HR [b] (95% CI [c])	***P* value**	**HR [b]** (95% CI [c])	***P* value**
Stage				
1	1	-	1	-
2	2.89 (1.52, 5.51)	0.001	3.13 (1.64, 5.97)	< 0.001
3	1.94 (0.81, 4.63)	0.13	2.53 (1.05, 6.12)	0.03
Grade				
1	1	-	1	1
2	2.46 (1.61, 5.23)	0.02	2.46 (1.15, 5.24)	0.02
3	1.33 (0.58, 3.04)	0.50	1.52 (0.65, 3.60)	0.34
Age				
< 48 years	1	-	1	1
≥ 48 years	1.75 (0.91, 3.38)	0.10	1.92 (1.01, 3.65)	0.04
Benign disease				
No	1	-	1	-
Yes	1.91 (1.04, 3.49)	0.04	2.32 (1.24, 4.33)	0.01
Performance of models				
C-index , %	72	-	76	-
Likelihood ratio test	32.44	-	51.19	-

[a] D: Number of deaths
[b] HR: Hazard ratio
[c] CI: Confidence interval

Software

A series of packages that work using the R software (version 2.5.1) were used (19). The KM curves were plotted by using the SPSS software.

Results

Information for the age variable was available for all patients. The variables nodal status and grade have a missing rate of about 20 % (20.3 % and 20.6 %, respectively). The missing rate for the history of benign disease was 15.2 %. However, after exclusion of the missing data for all 4 variables, 35 % of the data were lost. In total, data on all 4 variables was available for 203 cases (65 %). Almost all patients with missing data were those who survived. Out of the 56 cases of deaths, only 2 were lost in the C-C analysis. First, we examined the missing data mechanism *(Table 1)*. Patient's disease status, grade, and history of benign disease can help predict missingness on the stage variable. Patient status and history of benign disease were predictors of missingness on the grade variable. Furthermore, patient status and age were predictors of missingness on the history of benign disease variable. These findings confirm that the data had an MAR mechanism.

The estimated HR's and 95 % C.I. values corresponding to the C-C and imputed data sets are given in *Table 2*. Age at diagnosis was not significant in the C-C model Furthermore, compared to the patients with stage 1 disease, patients with stage 3 disease did not have a significant risk of death. After imputing the missing data, age at diagnosis was retained in the model. In addition, compared to patients with stage 1 disease, patients with stage 3 disease had significant HR's. On comparing the performances of the models, we found that the imputation of missing data improved the discrimination ability of the MICE model (MICE, 76 % versus C-C, 72 %). Furthermore, improvement was seen in the goodness of fit of the model (before imputation of missing data, 32.44 versus after imputation of missing data, 51.19).

Discussion

Missing data are a common problem in medical and epidemiological data sets. Exclusion of missing data leads to loss of power of the model. In this study, some variables lost their statistical significance in the C-C analysis. For example, stage of disease, which is one of the most important prognostic variables (20, 21), did not reach the significance level in the C-C model. We imputed 10 data sets in order to protect against chance effects of imputation. We considered this protection to be worth the inconvenience of having to average risk scores across the final 10 models. Once missing data were imputed, the power increased and the variables that had lost their significance in the C-C model (such as stage of disease) reached the significance level.

We also showed that our data had an MAR mechanism, indicating that the missing data depends on other characteristics of patients; therefore, the missing data can be imputed using multiple imputation methods. Our main goal was to illustrate the process of development of a prognostic model for cases with missing data, and we used a breast cancer data set to illustrate the process. The discussion of the risk factors for breast cancer is beyond the scope of this paper. The risk factors for breast cancer have been discussed previously (20,21). When the missing rate is low, results of the C-C model, in terms of the variables retained in the final model, may be similar to those of the MICE model. Asia Pacific Cohort Studies Collaboration (APCSC) collects data to determine the risk factors for coronary heart disease (CHD). The ability of multiple imputation and C-C analysis to handle the missing data on a single variable (cholesterol) in 26 studies was compared (22). The missing rate for cholesterol varied from 0 % to 69 %. In 22 studies in which the cholesterol data were unavailable for about 10 % of the subjects, the C-C and MICE methods gave similar results. However, in the remaining 4 studies in which the missing rate for cholesterol data was between 10 % and 60 %, a clear difference was seen between the results of the 2 models.

However, we believe that even a low rate of missing data on each variable may cause serious problems in multivariate modeling when patients with missing data on different variables are not the same. Furthermore, the missing data may substantially reduce the number of complete cases available for analysis and in turn increase the chance of bias because of the exclusion of cases. We developed the multifactorial models in conjunction with the ENTER variable selection method. When backward elimination (BE) variable selection is adopted, a series of iterative steps are required to exclude variables that do not contribute significantly to the model. If a single multifactorial model is developed, then application of BE is straightforward. However, when there are 10 imputed data sets, direct application of BE will not be feasible. In an iterative process, the results are aggregated across the 10 data sets at each step, and the variable with the highest *P value* (exceeding 0.05) is removed. Another set of 10 models are fitted to the remaining variables, the results are aggregated, and the *P value* for the variable to be excluded in determined (variable is excluded if $P > 0.05$). The whole process is continued as long as the variables remain statistically significant (12,13). Before developing multifactorial models, we dichotomized the age variable at a median age of 48 years since we have previously shown that, compared to the continuous variable, the dichotomized variable improves the quality of the model (23). Therefore, in this study, only information on 2 binary and 2 categorical variables was analyzed. When continuous data are available, the Predictive Mean Matching (PMM) technique can be employed. In the PMM method, the complete case with a value closest to the imputed value is chosen by taking the observation from the complete case as the imputed value.

Our study has several limitations. We used a data set containing only 4 variables. Therefore, the impact of the number of variables on the multifactorial model was not investigated. Furthermore, we only compared the performances of the C-C model and MICE at a missing rate of 35 % and under the MAR mechanism. The performance of the models depends on the mechanism of missing data, rate of missing data, method of imputation of missing data, and sample size to a great extent (24,25,26,27). Our work was simply a case study to explain the issues encountered in the application of the MICE method and the skill involved in the recovery of information. Therefore, further studies are required to compare the performance of imputation models in different situations (different sample size, missingness mechanism, missing rate, and method of imputation).
We have already shown that the C-C model decreases the power of the variables and the MICE method recovers data. However, at this stage, because of the limitations listed above, we cannot provide a specific guideline on how the problem of missing data can be best tackled because there are many approaches to deal with missing data (28). Under special circumstances, alternative methods with easier techniques (such as replacement of missing data by the mean observed value) may provide comparable estimates. Application and comparison of alternative imputation methods are beyond the scope of this paper and will be published elsewhere. Our results show how exclusion of missing data affects the composition of the model. Application of ad hoc methods such as C-C analysis is hugely criticized (29,30). Even when C-C analysis gives results comparable to those of the MICE method, a gold standard such as the MICE method is required to compare results with those of other simpler methods. Therefore, the application of MICE-like methods is highly recommended.

Financial support

This work, with no financial support, was done using an available data set.

Conflict of interest

None declared.

Acknowledgment

We would like to thank the staff of Motahhari Para Clinic and Shahid Faghihi Hospital for facilitating our access to patients' folders and information.

Author Contribution

The data set analyzed in this project was collected under the direction of Prof. TAR at Shiraz University of Medical Sciences. BMR performed all analyses and wrote the manuscript.

Appendix A

1) Technical details of the MICE method

If $X = (X_1, X_2, ..., X_k)$ are k random variables where each variable contains missing data and t represents iteration number, missing data are imputed from the following sequence of Gibbs sampler iterations (31):

For X_1 draw imputations from X_1^{t+1} from $P(X_1 | X_2^t, X_3^t, ..., X_k^t)$
For X_2 draw imputations from X_2^{t+1} from $P(X_2 | X_1^{t+1}, X_3^t, ..., X_k^t)$
For X_k draw imputations from X_k^{t+1} from $P(X_k | X_1^{t+1}, X_2^{t+1}, ..., X_{k-1}^{t+1})$

2) Aggregation of estimates across imputed data sets

If $\hat{\beta}_i$ is the estimated regression coefficient in the i^{th} data set, then the final regression coefficients, say $\hat{\beta}$, and its variance can be estimated by applying the following formulas. Here, M shows the number of data sets imputed.

$$\hat{\beta} = \frac{1}{M}\sum_{i=1}^{M}\hat{\beta}_i \qquad \hat{Var}(\hat{\beta}) = \frac{1}{M}\sum_{i=1}^{M}\hat{Var}(\hat{\beta}_i) + (1+\frac{1}{M})(\frac{1}{M-1})\sum_{i=1}^{M}(\hat{\beta}_i - \hat{\beta})^2$$

References

1. Altman DG, Lyman GH. Methodological challenges in the evaluation of prognostic factors in breast cancer. *Breast Cancer Res Treat*. 1998;**52**(1-3):289-303.
2. Concato J, Feinstein AR, Holford TR. The risk of determining risk with multivariable models. *Ann Intern Med*. 1993;**118**(3):201-10.
3. Wyatt JC, Altman DG. Prognostic models: clinically useful or simply forgotten. *British Med J*. 1995;**311**:1539-41.
4. Burton A, Altman DG. Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. *Br J Cancer*. 2004;**91**(1):4-8.
5. Altman DG, Bland JM. Missing data. *BMJ*. 2007;**334**(7590):424.
6. Baneshi MR, Talei AR. Impact of imputation of missing data on estimation of survival rates: an example in breast cancer. *Iran J Cancer Prev*. 2010;**3**(3):127-31.
7. Rajaeefard AR, Baneshi MR, Talei AR, Mehrabani D. Survival Models in Breast Cancer. *Iran Red Crescent Med J*. 2009;**11**(3):295-300.
8. Cox DR. Regression models and life tables. *J R Stat Soc* 1972;**34**:187-220.
9. Schafer JL. Analysis of Incomplete Multivariate Data. Florida: *Chapman and Hall*; 1997.
10. Schafer JL. Multiple imputation: a primer. *Stat Methods Med Res*. 1999;**8**(1):3-15.
11. Moons KG, Donders RA, Stijnen T, Harrell FE, Jr. Using the outcome for imputation of missing predictor values was preferred. *J Clin Epidemiol*. 2006;**59**(10):1092-101.
12. Van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Stat Med*. 1999;**18**(6):681-94.
13. Clark TG, Altman DG. Developing a prognostic model in the presence of missing data: an ovarian cancer case study. *J Clin Epidemiol*. 2003;**56**(1):28-37.
14. Rubin DB. Multiple imputation for non response in surveys. 1978.
15. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med*. 1999;**130**(6):515-24.
16. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;**143**(1):29-36.
17. Pencina MJ, D'Agostino RB. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Stat Med*. 2004;**23**(13):2109-23.
18. Steyerberg EW, Harrell FE, Jr., Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression

analysis. *J Clin Epidemiol.* 2001;**54**(8):774-81.
19. R: A language and environment for [statistical computing computer program]. 2007.
20. Baneshi MR, Warner P, Anderson N, Tovey S, Edwards J, Bartlett JM. Can biomarkers improve ability of NPI in risk prediction? a decision tree model analysis. *Iran J Can Prev.* 2010;**2**:62-74.
21. Baneshi MR, Warner P, Anderson N, Bartlett JSM. Tamoxifen resistance in early breast cancer: statistical modelling of tissue markers to improve risk prediction. *British J Can.* 2010;**102**:1503-10.
22. 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.
23. Baneshi MR, Talei AR. Dichotomisation of continuous data: review of methods, advantages, and disadvantages. Iran J Cancer Prev. 2010;**4**(1):26-32.
24. Ambler G, Omar RZ, Royston P. A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome. *Stat Methods Med Res.* 2007;**16**(3):277-98.
25. Bono C, Ried LD, Kimberlin C, Vogel B. Missing data on the Center for Epidemiologic Studies Depression Scale: a comparison of 4 imputation techniques. *Res Social Adm Pharm.* 2007;**3**(1):1-27.
26. Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. *J Clin Epidemiol.* 2006;**59**(10):1087-91.
27. Musil CM, Warner CB, Yobas PK, Jones SL. A comparison of imputation techniques for handling missing data. *West J Nurs Res.* 2002;**24**(7):815-29.
28. Baneshi MR. Statistical Models in Prognostic Modelling of Many Skewed Variables and Missing Data: A Case Study in Breast Cancer (PhD thesis submitted at Edinburgh University) 2009.
29. Croy CD, Novins DK. Methods for addressing missing data in psychiatric and developmental research. *J Am Acad Child Adolesc Psychiatry*. 2005;**44**(12):1230-40.
30. Van Der Heijden GJ, Donders AR, Stijnen T, Moons KG. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. *J Clin Epidemiol.* 2006;**59**(10):1102-9.