Impact of Imputation of Missing Data on Estimation of Survival Rates: An Example in Breast Cancer

Baneshi MR[1], Talei AR[2]

Abstract

Background: Multifactorial regression models are frequently used in medicine to estimate survival rate of patients across risk groups. However, their results are not generalisable, if in the development of models assumptions required are not satisfied. Missing data is a common problem in pathology. The aim of this paper is to address the danger of exclusion of cases with missing data, and to highlight the importance of imputation of missing data before development of multifactorial models.

Methods: This study was performed on 310 breast cancer patients diagnosed in Shiraz (Southern Iran). Performing a complete-case Cox regression model, a prognostic index was calculated so as to categorise the patients into 3 risk groups. Then, applying the Multivariate Imputation via Chained Equations (MICE) method, missing data were imputed 10 times. Using imputed data sets, modelling was performed to assign patients into risk groups. Estimated actuarial Overal Survival (OS) rates corresponding to analysis of complete-case and imputed data sets were compared.

Results: Cases with at least one missing datum experienced a significantly better survival curve. Estimates derived analysing complete-case data, relative to imputed data sets, underestimated the OS rate in all risk groups. In addition confidence intervals were wider indicating loss in precision due to attrition in sample size and power.

Conclusion: Results obtained highlighted the danger of exclusion of missing data. Imputation of missing data avoids biased estimates, increases the precision of estimates, and improves genralisability of results to other similar populations.

Key words: Missing data; Multiple imputation; Breast neoplasm; Overall survival, Iran

Please cite this article as: 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; Vol3, No3, p.127-31.

1. Health School, Kerman University of Medical Sciences, Deptartment of Biostatistics and Epidemiology, Kerman, Iran
2. Shahid Faghihi Hospital, Shiraz University of Medical Sciences, Shiraz, Iran

Corresponding author:
Mohammad Reza Baneshi, PhD in Biostatistics
Tel: (+98) 913 442 39 48
Email: m_baneshi@ kmu.ac.ir

Received: 12 Jan, 2010
Accepted: 21 Jun, 2010
Iran J Cancer Prev 2010; 3: 127-31

Introduction

Cancer is one of the most major health problems worldwide. In 2002, a quarter of the 11 million new cases of cancer reported worldwide occurred in Europe. Among new cancer patients diagnosed in the UK, which is more than a quarter of a million per year, the most prevalent carcinomas (incidence rate) were breast (16%), lung (13%), bowel or colorectal (13%), and prostate (12%) [1]. Breast carcinoma, with one million newly diagnosed cases annually, is the most prevalent malignancy, comprising 18% of all female cancers [2].

In Iran, cancer is the third cause of deaths after cardiovascular diseases and accidents [3]. The breast cancer is the most lethal one among women. The prevalence of breast cancer was reported 25.4 and deaths due to breast cancer were 12.3 per 100,000 [3].

Clinical trials typically involve collection of patient data at entry and in so far as are possible these data will include variables of potential relevance to the likely cause of the disease under study. These data sets have been used in development of prognostic models, which provides a valuable resource in identifying important risk factors for disease course and hence also for risk stratification of patients.

However, if in development of prognostic models, one ignores model assumptions and limitations the models obtained might not be generalisable [4,5].

Presence of missing data is one issue which makes difficulties in model building. When missing data

present, researchers frequently drop out patients with missing data on any of variables under study from consideration. This ad hoc method is known as Complete-Case (C-C) analysis [6]. It has been emphasized that exclusion of missing data will diminish precision of estimates and can lead to biased estimates [7].

Survival rates are frequently reported in the literature to compare treatment options, and to inform the patients about their likely outcome [8]. Exclusion of missing data results in biased estimate of cohort survival rates, in particular when there is difference in survival curve of cases with available data with the remainder (who had at least one missing datum) [9].

As an example when cases with missing data, in comparison with those who had data available, exhibits lower survival curve, omission of missing data results in overestimation of survival rates [9]. Therefore, appropriate methods should be applied to impute missing data so as to avoid attrition in sample size.

The aim of this paper is to compare estimation of survival rates under two scenarios: in complete-case analysis, and after imputation of missing data. Methods were applied analysing a breast cancer data set.

Materials and Methods

Patients and outcome

From 1994 to 2003, the information of 310 breast cancer patients in Shiraz, southern Iran were collected from Hospital-based Cancer Registry of Nemazee Hospital affiliated to Shiraz University of Medical Sciences. Median follow-up time was 2.5 years. The main outcome of study was Overall Survival (OS). Survival was considered as the time period between diagnosis and death for patients who died, and from diagnosis to the last visit for censored patient. At the end of the study, there had been 56 deaths.

At the first step a multifactorial model was developed (see the rest of the text). The OS rates were estimated from risk groups derived (explained later). Variables offered to the multifactorial models were those showed to have univariate predictive ability [10] (tumour stage with 3 levels (early, locally advanced, and advanced), tumour grade with 3 levels (1, 2, and 3), history of benign breast disease (positive versus negative), and age at diagnosis). Prior to analysis, the age variable was dichotomised at 48 to be a surrogate for approximate menopausal status [11].

Multifactorial Models

At first a dummy variable was created which took a value of 0 if patient had available data on all variables under consideration and 1 otherwise. Survival curve of patients with and without missing data were compared plotting Kaplan-Meier curves ad performing Log-Rank test. Linear Cox model was then applied to develop the multifactorial regression models [12].

Complete-Case (C-C) Model

In the C-C model, patients with missing data on any of 4 candidate variables were excluded. Cox regression model in conjunction with ENTER variable selection method was then fitted. A final risk score was calculated by multiplying variables into the estimated regression coefficient. Tertiles of the risk score estimated were applied as cut off to categorise patients into low, intermediate, and high risk groups.

MICE Model

Multivariable Imputation via Chained Equations (MICE) method is then applied to impute missing data. The MICE method is a powerful tool to tackle the missing values. The MICE method replaces each missing value by multiple imputed values, typically 10, resulting in multiply imputed data sets [13,14].

Patients' outcome and set of 4 risk factors were used in the MICE algorithm [15]. Polytomous and logistic regression were used to impute missing data for categorical and binary data respectively.

The creation of 10 data sets means there is a requirement for 10 modelling analyses, one for each data set, and there will therefore be 10 different estimates for each parameter. A Cox regression model was fitted to each of 10 imputed data sets. In each of 10 data sets, multiplying data set specific estimates into the variables, a risk score was calculated (10 in total). Finally, for each patient a single averaged risk score was calculated by averaging her estimated risk scores from each of the 10 imputed data sets. Tertiles of the final risk score was applied as cut offs to divide patients into low, intermediate, and high risk groups.

Estimation of Overall Survival (OS) rates

To compare the OS rates in risk groups, actuarial 2, 4, and 5-year OS rates in the lowest, intermediate, and highest risk groups are reported. This was done analysing complete-case and imputed data sets.

Based on definition the survival function, say S (4), is the probability of being alive at least till 4th year of

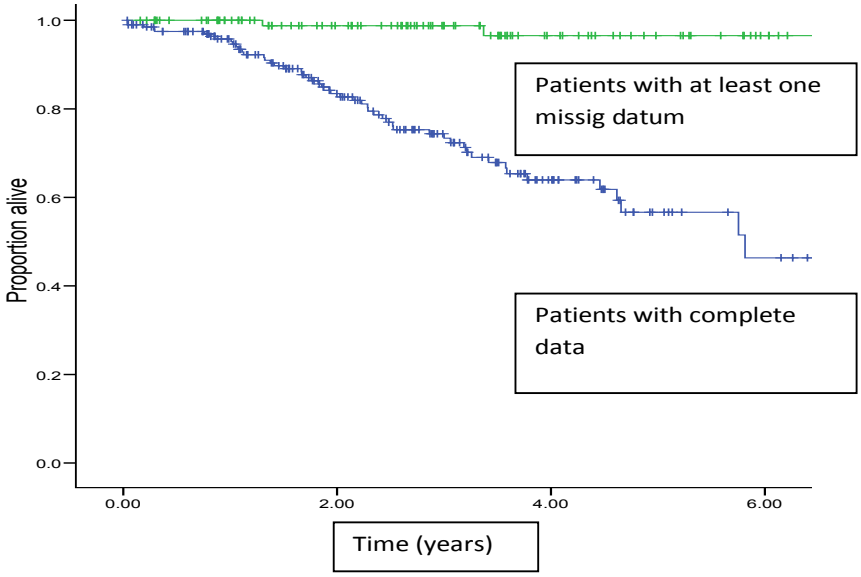

Figure 1. K-M curves for cases with available data and cases with at least one.

Table 1. Comparison of survival of patients with available data and with at least one missing datum

Group	# of patients	# of events	Log-Rank P-value
Cases with available data on all 4 variables	203	54	<0.0001
Cases with at least 1 missing datum	107	2	

Table 2. Comparison of estimated OS rates in the risk groups derived analysing complete case and imputed data sets

Model	Risk group	2-year OS (%) (95% C.I.)	4-year OS (%) (95% C.I.)	5-year OS (%) (95% C.I.)
Complete Case	Low	92 (84, 100)	84 (70, 96)	84 (70, 96)
	Intermediate	79 (67, 91)	67 (51, 83)	67 (51, 83)
	High	52 (38, 66)	28 (12, 44)	16 (0, 32)
Imputed data set	Low	95 (91, 99)	90 (82, 98)	90 (82, 98)
	Intermediate	88 (80, 96)	82 (70, 94)	82 (70, 94)
	High	64 (52, 76)	42 (28, 56)	32 (16, 48)

follow up. Therefore, survival at the 4th year depends on survival at first, second…and 4th year which implies that $S(4) = P(T \geq 4)$. In actuarial life-table procedure, the whole follow-up duration will be split to intervals (as an example to 1 year intervals (0, 1], (1, 2], (2, 3], (3, 4] respectively). If n_i and d_i show number of patients at risk just before the i-th interval and the number of events at i-th interval, then the probability of surviving to 4th year is given by $S(4) = \prod_{i=1}^{4} (1 - \frac{d_i}{n_i})$

Based on Greenwood's formula the variance of this estimator can be estimated by

$$\hat{V}(\hat{S}(t)) = \hat{S}^2(t) \sum_{i=1}^{4} \frac{d_i}{n_i(n_i - d_i)}$$

To address loss in precision of estimates, confidence intervals of OS rates, corresponding to analysis of C-C and imputed data sets, were estimated and compared.

Software

A series of packages which work under R software (version 2.5.1) were used [16]. Missing data were

imputed using MICE package [17]. Performance of models (discrimination and predictive ability) were assessed using Design [18] library. K-M curves are plotted using SPSS software.

Results

The numbers (percentages) of patients with missing value on node status, grade, and history of benign disease were 63 (20.3%), 64 (20.6%), and 47 (15.2%) respectively. In total, out of 310 patients, 203 cases (65%) had data available on all 4 variables of which 54 had died.

Table 1 reports the number of deaths for patients with complete data and the remainder with at least one missing datum. Corresponding K-M curves is plotted in Figure 1. Cases with complete data had much lower survival curve (Log-Rank P-value <0.0001). This indicates that exclusion of cases with missing data leads to underestimation of the true OS rates in the cohort analysed.

As explained in methods section a risk score was estimated for complete-case and imputed data sets. Using tertiles as cut off, patients were categorised into 3 risk groups (low, intermediate, and high). Estimated OS rates in risk groups derived are summarised in Table 2. Estimations derived analysing patients with available data, underestimated OS rates in all 3 risk groups. This was the case in all 3 risk groups, and time points. For example, estimated 2-year OS rate in lowest risk group for complete-case ad imputed data sets were 92% and 95% respectively. Corresponding rates at 4 years were 84% and 90% respectively.

Furthermore, C.I.'s corresponding to imputed data sets, relative to complete-case data, was tighter since attrition in sample size is avoided.

Discussion

We have seen that confidence intervals of OS rates corresponding to the imputed data sets were narrower indicating improvement in precision of estimates. Furthermore, comparing K-M curves of patients with available data with those with at least one missing datum suggested that exclusion of missing data leads to underestimation of OS rates. This was consistent with estimated we obtained which are summarised in Table 2.

To provide more accurate estimates, we imputed missing data 10 times. This was to protect against chance effects dues to imputation. This protection was to be felt worth the inconvenience of having to average risk scores across 10 final models. Easier imputation methods such as Expectation Maximum (E-M) algorithm are likelihood based and suitable approaches. However, E-M method replaces each missing data by a single value so does not take into account imputation uncertainty.

It has been noted that under the Missing Completely At Random (MCAR) assumption, subjects with complete data are a random sample of data [19]. It has been argued that under MCAR mechanism if missing rate is less than 5%, case deletion is a reasonable approach [20]. However, it should be emphasized that even when C-C analysis give results comparable to the MICE, a gold standard (MICE) is required to compare results from other simpler methods [21].

On the other hand, when missing rate is high, exclusion of missing data will diminish precision of estimates. Another issue is that even a low rate of missing data on each variable might cause serious problems in multivariate modelling when patients with missing data on are scattered across the data. That is because this might substantially reduce the number of complete cases available for analysis, and increase the chance of bias due to excluded cases.

There are lots of ad hoc (such as C-C, replacement by mean, and missing indicator approaches) and maximum likelihood methods (such as E-M algorithm, and multiple imputation technique) to deal with missing data [22]. Application and comparison of alternative imputation methods was beyond the scope of this paper and will be published elsewhere.

The ultimate consequence of complete-case analysis is power reduction. In addition, case-deletion might result in biased regression coefficients if the remaining cases are not the representative of the whole sample [7,23]. Results presented showed that exclusion of cases with missing data leads to bias and imprecise estimates. Therefore imputation of missing data should be a prime before any modelling practice.

Acknowledgment

We should thank staff of Motahhari Para clinic and Shahid Faghihi hospital who facilitated our access to patients' folder and information.

Conflict of Interest

There is no conflict of interest in this article.

Authors' Contribution

The data set analyzed in this project was collected under the direction of Professor TAR at Shiraz University of Medical Sciences. All analyses and writing of manuscript has been done by BMR. Both

authors have read and approved the final version of the manuscript.

References

1. Cancer Research UK. UK cancer incidence statistics. http://info cancerresearchuk org/ cancerstats/ incidence/ ?a=5441 2007 January [cited 2007 Feb 26];Available from: URL: http:// info. cancerresearchuk. org/ cancerstats/ incidence/?a=5441
2. McPherson K, Steel CM, Dixon JM. ABC of breast diseases. Breast cancer-epidemiology, risk factors, and genetics. BMJ 2000 Sep 9; 321(7261):624-8.
3. Naghavi M. Iranian annual of national death registration report. Iran ministry of health and medical education; 2005.
4. Concato J, Feinstein AR, Holford TR. The risk of determining risk with multivariable models. Ann Intern Med 1993 Feb 1; 118(3):201-10.
5. Wyatt JC, Altman DG. Prognostic models: clinically useful or simply forgotten. BRITISH MEDICAL JOURNAL 1995; 311:1539-41.
6. Burton A, Altman DG. Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. Br J Cancer 2004 Jul 5; 91(1):4-8.
7. Altman DG, Bland JM. Missing data. BMJ 2007 Feb 24; 334(7590):424.
8. 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.
9. Van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. Stat Med 1999 Mar 30; 18(6):681-94.
10. Rajaeefard AR, Baneshi MR, Talei AR, Mehrabani D. Survival Models in Breast Cancer. Iranian Red Crescent Medical Journal 2009; 11(3):295-300.
11. Ayatollahi SM GHASA. Menstrual-reproductive factors and age at natural menopause in Iran. International journal of gynaecology and obstetrics 2003; 80(3):311-3.
12. Cox DR. Regression models and life tables. Journal of royal statistical society 1972; 34:187-220.
13. Schafer JL. Analysis of Incomplete Multivariate Data. Florida: Chapman and Hall; 1997.
14. Schafer JL. Multiple imputations: a primer. Stat Methods Med Res 1999 Mar; 8(1):3-15.
15. Moons KG, Donders RA, Stijnen T, Harrell FE, Jr. Using the outcome for imputation of missing predictor values were preferred. J Clin Epidemiol 2006 Oct; 59(10):1092-101.
16. R: A language and environment for statistical computing [computer program]. 2007.
17. Mice: Multivariate Imputation by Chained Equations [computer program]. 2007.
18. Design: Design Package [computer program]. 2008.
19. Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 2006 Oct; 59(10):1087-91.
20. Fairclough DL. Patient reported outcomes as endpoints in medical research. Stat Methods Med Res 2004 Apr; 13(2):115-38.
21. Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol 1995 Dec 15; 142(12):1255-64.
22. 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.
23. Harrell FE. Regression modelling strategies with application to linear models, logistic regression, and survival analysis. New York: Springer-Verlag; 2001.