Iranian J Publ Health, Vol. 41, No.5, May 2012, pp.110-115

Original Article

Assessment of Internal Validity of Prognostic Models through Bootstrapping and Multiple Imputation of Missing Data

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

1. *Research Center for Modeling in Health, Kerman University of Medical Sciences, Kerman, Iran*
2. *Shahid Faghihi Hospital, Shiraz University of Medical Sciences, Shiraz, Iran*

(Received 27 Oct 2011; accepted 11 Feb 2012)

Abstract

Background: Prognostic models have clinical appeal to aid therapeutic decision making. Two main practical challenges in development of such models are assessment of validity of models and imputation of missing data. In this study, importance of imputation of missing data and application of bootstrap technique in development, simplification, and assessment of internal validity of a prognostic model is highlighted.

Methods: Overall, 310 breast cancer patients were recruited. Missing data were imputed 10 times. Then to deal with sensitivity of the model due to small changes in the data (internal validity), 100 bootstrap samples were drawn from each of 10 imputed data sets leading to 1000 samples. A Cox regression model was fitted to each of 1000 samples. Only variables retained in more than 50% of samples were used in development of final model.

Results: Four variables retained significant in more than 50% (i.e. 500 samples) of bootstrap samples; tumour size (91%), tumour grade (64%), history of benign breast disease (77%), and age at diagnosis (59%). Tumour size was the strongest predictor with inclusion frequency exceeding 90%. Number of deliveries was correlated with age at diagnosis (r=0.35, P<0.001). These two variables together retained significant in more than 90% of samples.

Conclusion: We addressed two important methodological issues using a cohort of breast cancer patients. The algorithm combines multiple imputation of missing data and bootstrapping and has the potential to be applied in all kind of regression modelling exercises so as to address internal validity of models.

Keywords: Missing data, Multiple imputation, Bootstrap, Breast neoplasm, Internal validity

Introduction

Multifactorial regression models are frequently used in medicine to develop prediction tools. Development of multifactorial models needs careful considerations. Two main issues, which are of crucial importance, are two select variables for final multifactorial model and avoiding bias estimates due to missing data.

The first issue discussed here is selection of variables to be contributed to the multifactorial model. In development of regression models, researchers usually apply stepwise variable selection procedures such as Backward Elimination (B.E.) and Forward Selection (F.S.). However, such methods suffer lack of stability. This is because the inclusion or exclusion of a few cases can affect the variables selected for the model and resulting parameter estimates (1-3).

This issue has been addressed in the development of a prediction model for acute myocardial infraction mortality. In 1000 bootstrap samples, B.E.

***Corresponding Author:** Tel: 0098 913 442 39 48, E-mail address: m_baneshi@kmu.ac.ir

produced 940 unique models (2). Out of 29 variables, only three variables were significant in all the bootstrap samples, 18 variables were selected in fewer than half of the bootstrap samples, and six variables in less than 10%. This demonstrates the sensitivity of B.E. to small differences between bootstrap samples. It has therefore been recommended to use B.E. in conjunction with bootstrap procedure (4-6). That is, to apply B.E. to a number of bootstrap samples (typically 100) and then to check selection of variables across samples (known as inclusion frequency or percentage).

It has therefore been recommended to use B.E. in conjunction with bootstrap procedure (4-6). That is, to apply B.E. to a number of bootstrap samples (typically 100) and then to check selection of variables across samples (known as inclusion frequency or percentage).

The second issue is missing data. The simplest approach, to tackle missing data, is to exclude cases with missing data. Case-wise (list-wise) deletion means to omit all records that contain missing data for any variable. Pair-wise deletion method, on the other hand, uses a correlation matrix where correlation between each pair of variables is calculated from all cases that have valid data for those two variables. This method seems better than case-wise deletion but the problem is that the parameters of the model will be based on different sets of data, with different sample sizes and different standard errors. Therefore, the resulting correlation matrix may not be suitable for further analysis such as regression models. Disadvantages of C-C analyses are highlighted elsewhere (7, 8). Furthermore, this method does not work when the aim is to develop a multifactorial regression model to adjust effect of variables in presence of other covariates.

On the other hand, the Multivariable Imputation via Chained Equations (MICE) method can be used to impute the missing values (7, 9). Superiority of this method over other imputation algorithms has been addressed elsewhere (10).

Majority of studies illustrated usefulness of MICE to tackle missing data, and bootstrapping to address internal validity of regression models. However, very few studies combine them together to provide a larger picture of uncertainties that can happen (i.e. uncertainties due to imputation of missing data, and sampling variations). A recent study recommended the issue of combining these two methods in prognostic studies (11).

Using a breast cancer data set, we already illustrated different methodological issues (12). In a recent work, we addressed the process of the MICE method and its superiority over the Complete Case (C-C) analysis (9). In other words, we performed and reported results of majority of imputation methods, including the MICE model, elsewhere.

The main aim of this paper was to combine these two steps (MICE and bootstrapping) to develop a prognostic model.

Materials and Methods

Patients and outcome

Study sample comprised of 310 breast cancer patients in Shiraz (southern Iran) of which 56 cases died due to breast cancer. Data were collected from Hospital-based Cancer Registry of Nemazee Hospital affiliated to Shiraz University of Medical Sciences. To secure confidentiality of patients, the data set does not include personal information such as name, address, email, or phone number. Median follow-up time was 2.5 years.

Variables

Variables offered to the multifactorial models were 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), age at diagnosis <=48 versus > 48), treatment option (lumpectomy versus mastectomy), and number of deliveries.

Imputation of missing data to tackle missing data

For all candidate variables, we imputed missing data using the MICE method. We imputed 10 values for each missing value, thus creating 10 imputed data sets (13).

Selection of bootstrap samples to check internal validity

To circumvent the risk of an over-fitted model, and to check the internal validity of the model, we have employed bootstrap sampling to refine the models by excluding variables with unreliably included as necessary for prediction. We drew 100 bootstrap samples from each of 10 imputed data sets, giving 1000 data sets in total.

Multifactorial model

When modelling across bootstrap samples, the prognostic variables that truly are important should be retained in most models fitted. This is because each bootstrap replication is a random sample that should therefore reflect and mimic the underlying structure of the data, and it is this should drive the variables needed in the majority of models fitted (4-6). In other words, bootstrap technique can be used as a measurew of internal validity. Therefore, a measure of inclusion frequency can be used to screen for the selection of the variables (1, 14). It has been shown that the inclusion of a variable in the model at selection levels of 1% and 5% in the original data can be checked against a cut of value for the bootstrap inclusion fraction of 73% and 50% respectively (1). Therefore, only variables retained in more than 50% of samples (i.e. 500 out of 1000 samples) were selected to construct the final model.

It should be added that when independent variables are correlated, if the inclusion frequency of correlated variables together exceeds 90%, then the one with higher inclusion frequency should be offered to the model. Otherwise, both should be omitted (1, 14).

Aggregation of results

Using only the reliable variables identified in the previous step (i.e. those retained significant in >50% of samples), a final model was then fitted to each of the 1000 samples. Applying Rubin's rule, coefficients for these 1000 models were then averaged across models, and standard errors combined (8, 9). Final Hazard Ratios (HR) was then estimated applying exponential transformation to the aggregated coefficients.

Results

The inclusion frequency for all 6 variables offered to multifactorial models is given in Table 1. In total four variables were retained significant in more than 50% of samples: tumour stage, tumour grade, history of benign disease, and age at diagnosis time. The final multifactorial model for Breast Cancer Specific Death (BCSD) retained four variables (Table 1), together with aggregated hazard ratios (as described in Methods).

Tumour stage seems the most important predictor of breast cancer specific death. This variable significantly contributed to the model in more than 90% of replications.

History of benign disease was the second most important predictor. This variable retained in nearly three-forth of replications. The Hazard Ratio (HR) of breast cancer specific death for those with positive history of benign breast was 2.4 (95% C.I.: 1.25, 4.19) times higher than others.

Tumour grade was contributed to about 60% of samples, and therefore used to develop final model. However, if one wishes to use 0.01 as significance level, then this variable should not be offered to the final model, since its inclusion frequency is lower than 73%.

The inclusion frequency of age at diagnosis and number of deliveries were 59% and 32% respectively. The spearman correlation between these two variables was 35% (P<0.001). This means that umber of deliveries can be used as a surrogate approximation for age variable.

However, when both variables were offered to the models, age variable was retained in majority of samples (59% versus 32%). Following recommendations given in the methods section, we used the age variable in the final modelling.

Patients in the early stage underwent either lumpectomy or mastectomy but in other stages, only mastectomy was done. This variable retained significant in small proportion of samples.

Table 1: Multifactorial Breast Cancer Specific Death (BCSD) model; relative frequency of covariate inclusion (in 1000 bootstrap samples drawn from 10 imputed data sets) and estimated HR's

Variable	Inclusion frequency (%)	Level	HR (95% C.I.)	*P*
Stage	91	Early	1	
		Locally advanced	3.07 (1.79, 5.84)	<0.001
		advanced	2.78 (1.15, 6.29)	0.02
Grade	64	1	1	
		2	2.87 (1.25, 5.69)	0.01
		3	1.78 (0.87, 4.23)	0.30
History of benign breas disease	77	No	1	
		yes	2.49 (1.25, 4.19)	0.01
Age At diagnosis	59	<= 48	1	
		>48	1.84 (1.04, 3.52)	0.03
Treatment	13	Mastectomy		
		Lumpectomy		
Number of deliveries	32	One unit change		

HR: Hazard Ratio C.I.: Confidence Interval

Discussion

By imputing multiple data sets (to avoid attrition in sample size due to presence of missing data) followed by bootstrapping (to check the reliability of inclusion across models (know as internal validity)), we applied a methodology which has the potential for future application in all medical areas.

The approach applied has the advantages that takes into account both imputation and sampling variations in to account. We offered variables with inclusion frequency of at least 50%. This should be added that when the aim is to fit a parsimony model, then only variables with very high inclusion frequency should be retained. On the other hand, when adjustment for covariates is the aim, selection of variables with low inclusion frequency is necessary and therefore, a low value for percentage of inclusion frequency should be selected (1).

Here we combined MICE and bootstrap. However, we have not made any comparison with C-C model, and MICE model without bootstrap. This is because those analyses have been published elsewhere and are not reported here (12). Our previous results showed that C-C provides biased estimates. However, in terms of variables contributed to the final model, results of MICE model without bootstrap (presented in (9)) were the same as results presented here. This might partially be explained by the fact that here we estimated eight regression coefficients. Therefore, ratio of number of Event Per Variable (EPV) was an acceptable figure (56/8 or 7). It has been shown that the lower the EPV the more unstable the model.

Ultimately, the most important issue for a model is its external validity, the extent to which it provides good predictions for similar patients who were not involved in the development of the model. However, before external validity can be checked, it is a prerequisite that there is adequate internal validity. Internal validation refers to the performance in patients from a similar population to those comprising the sample on which the model was developed. Therefore, internal validity is in contrast to external validity, where different populations are used to develop and test the model (15).

Generally, internal validity can be investigated by splitting the data into training and test samples, doing cross-validation, or performing a bootstrap resampling procedure (5, 16).

The data-splitting approach allows the hypothesis tests to be confirmed in the test sample. However, this method leads to a lower sample size, and consequently lower power, in training and test samples.

The Cross-validation method randomly divides the data, several times, as training and test samples, and applies the results obtained in training test-to-test sets. The use of leave-one-out cross-validation allows the researcher to build the model on N-1 cases (almost all) and then test it on the case that was left out. This method allows developing the model without scarifying the sample size. However, it is often the case, that the criterion to compare the performance of validation techniques could not be calculated for one case (17).

Alternatively, one can do 10-fodl cross-validation. Similar to data-splitting method, this technique may not be accurate if the training or test set is too small (18). However, when total sample size is 310, then each of test samples only formed by 31 individuals. Inference based on such small number of cases may not be robust.

On the other had, the bootstrap technique does not scarify the sample size and allows the researcher to be able to extract as much information as possible (19). It has been shown that, to assess internal validity, the bootstrap would be the best approach (17).

Assessment of internal validity of a model is an important issue. As an example, Chen et al. developed a prognostic model. To check the internal validity of the model they drew 100 bootstrap samples but the original model was seen just in 2% of replications (20). However, bootstrap inclusion frequencies of five of the six variables from the original model were between 64% and 82%. In another study in node positive breast cancer patients, the final prognostic model has not been seen in any of 200 bootstrap replications (21). These two examples show importance of assessment of internal validity of a model, to deal with sampling variation.

Here we demonstrated application of bootstrap technique in assessment of internal validity of models. We should add that bootstrap method has other methodological usages as well. For example, one can apply this technique to provide confidence interval for any parameter. For example, assume that the aim is to estimate the mean of a continuous variable in the population. We can estimate the mean from a sample and then apply the normal approximation formula to provide the confidence interval. However, if the data violated the distributional assumption, then this approximation should be avoided. In that case, it is possible to draw bootstrap samples and calculate bootstrap mans from each sample. Then sorting values, percentiles on 2.5 and 97.5 can be used as lower and upper bounds on confidence interval.

It has been shown that internal validity may not be sufficient for the good performance of the model in case of future patients. External validation is essential before implementation of prediction models in clinical practice (22). It has been emphasized that usefulness of a model is determined by how well it works in practice and not by how many zeros there are in the associated *P values* in the multifactorial model (15, 23).

The importance of assessment of external validity is illustrated here by an example. A prediction model for the presence of serious bacterial infections in children with fever without source was derived (22). The discrimination (C-index) and predictive ability (R-square) of the model was 0.83 and 32% respectively. The model was then validated in an independent sample (n=179) giving discrimination ability of 0.57 (0.47–0.67) and R-square of 20%.

We combined multiple imputation and bootstrapping methods to develop a prognostic model, which predicts BCSD. This algorithm has been applied here as well (24). The MICE method applied has the advantage that takes into account the imputation variation. Furthermore, bootstrap approach allows checking the sensitivity of model to small changes in the sample. Although bootstrap method is a powerful technique to check the internal validity of a model, an independent and fresh data set is needed to investigate transportability of the model (15).

Ethical considerations

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

Acknowledgments

The data set analyzed in this project was collected under the direction of Professor Talei at Shiraz University of Medical Sciences. No funding was received for this study. The authors declare that there is no conflict of interests.

References

1. Sauerbrei W, Schumacher M (1992). A bootstrap resampling procedure for model building: application to the Cox regression model. *Stat Med*, 11 (16): 2093-109.
2. Austin PC, Tu JV (2004). Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. *J Clin Epidemiol*, 57 (11): 1138-46.
3. Derksen S, Keselman J (1992). Backward, forward, and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, 45 (2): 265-82.
4. Altman DG, Andersen PK (1989). Bootstrap investigation of the stability of a Cox regression model. *Stat Med*, 8 (7): 771-83.
5. Harrell FE, Lee KL, Mark DB (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*, 15 (4): 361-87.
6. Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KG (2003). Internal and external validation of predictive models: a simulation study of bias and precision in small samples. *J Clin Epidemiol*, 56 (5): 441-7.
7. Baneshi MR, Talei AR (2010). Impact of imputation of missing data on estimation of survival rates: an example in breast cancer. *Iranian Journal of Cancer Prevention*, 3 (3): 127-31.
8. Baneshi MR, Faramarzi H, Marzban M (2012). Prevention of disease complications through diagnostic models: how to tackle the problem of missing data? *Iranian J Publ Health*, 14 (1): 66-72.
9. Baneshi MR, Talei AR (2011). Multiple Imputation in Survival Models: Applied on Breast Cancer Data. *Iranian Journal of Cancer Prevention*, 13 (8): 547-52.
10. Baneshi MR, Talei AR (2012). Does the Missing Data Imputation Method Affect the Composition and Performance of Prognostic Models? *Iranian Red Crescent Medical Journal*, 14 (1): 31-6.
11. Heymans MW, Van Buuren S, Knol DL, Van Mechelen W, de Vet HC (2007). Variable selection under multiple imputation using the bootstrap in a prognostic study. *BMC Med Res Methodol*, 7:33.
12. Rajaeefard AR, Baneshi MR, Talei AR, Mehrabani D (2009). Survival Models in Breast Cancer. *Iranian Red Crescent Medical Journal*, 11 (3): 295-300.
13. Schafer JL (1999). Multiple imputation: a primer. *Stat Methods Med Res*, 8 (1): 3-15.
14. Austin PC, Tu JV (2004). Bootstrap methods for developing predictive models in cardiovascular research. *American Statisticians*, 58: 131-7.
15. Justice AC, Covinsky KE, Berlin JA (1999). Assessing the generalizability of prognostic information. *Ann Intern Med*, 130 (6): 515-24.
16. Harrell FE, Lee KL, Matchar DB, Reichert TA (1985). Regression models for prognostic prediction: advantages, problems, and suggested solutions. *Cancer Treat Rep*, 69 (10): 1071-7.
17. Steyerberg EW, Harrell FE, Jr., Borsboom GJ et al. (2001). Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol*, 54 (8): 774-81.
18. Azuaje F (2003). Genomic data sampling and its effect on classification performance ass-essment. *BMC Bioinformatics*, 28: 4-5.
19. Sauerbrei W, Royston P (2007). Modelling to extract more information from clinical trials data: On some roles for the bootstrap. *Stat Med*, 26 (27): 4989-5001.
20. Chen CH, George SL (1985). The bootstrap and identification of prognostic factors via Cox's proportional hazards regression model. *Stat Med*, 4 (1): 39-46.
21. Sauerbrei W, Royston P (1999). Building multivariate prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. *Journal of Royal Statistical Society*, 162 (1): 71-94.
22. Bleeker SE, Moll HA, Steyerberg EW et al. (2003). External validation is necessary in prediction research: a clinical example. *J Clin Epidemiol*, 56 (9): 826-32.
23. Altman DG, Royston P (2000). What do we mean by validating a prognostic model? *Stat Med*, 19 (4): 453-73.
24. Baneshi MR, Warner P, Anderson N, Edwards J, Cooke TG, Bartlett JMS (2010). Tamoxifen resistance in early breast cancer: statistical modelling of tissue markers to improve risk prediction. *Br J Cancer*, 102: 1503-10.