

Author's response to reviews

Title: Variable selection under multiple imputation using the bootstrap in a prognostic study

Authors:

Martijn W Heymans (mw.heyman@vumc.nl)

Stef van Buuren (s.vanbuuren@pg.tno.nl)

Dirk L Knol (d.knol@vumc.nl)

Willem van Mechelen (w.vanmechelen@vumc.nl)

Henrica CW de Vet (hcw.devet@vumc.nl)

Version: 2 Date: 22 December 2006

Author's response to reviews: see over

BMC-series journals
Tel: +44 (0)20 7631 9921
Facsimile: +44 (0)20 7631 9923
e-mail: editorial@biomedcentral.com
Web: <http://www.biomedcentral.com/>

22 December 2006

Dear editor,

We thank the reviewers for their interest and time in reviewing our manuscript, and appreciate the possibility to react on the reviewer's comments. We enclose our response to the remarks of the reviewers. We have incorporated our suggested revisions to the text in bold type.

Reviewer 1 (Royston)

Major Compulsory Revisions

Ad 1.

We agree with Reviewer 1 that one cannot validly conclude anything from method B. It was not our purpose to propose method B as a separate prognostic modelling method in data sets with missing values. But we needed this method to achieve the goal of the study, which was to examine two sources of variation which complicate prognostic modelling in data sets with missing values. The first one originates from variation induced by sampling, i.e. by applying the bootstrap, the other one from variation induced by the spread of multiply imputed values, i.e. by applying multiple imputation techniques. We presented both the B and MI method to identify the amount of variation generated by each method. By reporting the MI method and not reporting the B method, it would be impossible to explain where the major variation occurred in the MI + B100 and MI + B10 methods. Therefore we choose to report on the variation induced by each method. We have now added this argument to the text.

Ad 2 and 3.

The reviewer asks for more detailed information about how the MICE imputation was done, e.g. if non-normality of continuous variables during the imputation modeling and model building was accounted for. **MICE was done with the closest predictor option (“predictive mean matching”) as described in Rubin (Rubin DB. Multiple Imputation for Nonresponse in Surveys, Wiley, New York, 1987) and Van Buuren et al. (Van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. Stat Med. 1999;18(6):681-94). This method models a missing variable A as a linear combination of predictor variables B, finds the complete case whose A estimate is closest to that of the current incomplete case, and takes the observed A from the former as the imputed A value**

for the latter. Although applying a linear model may seem a rather simplistic choice here, note that the only function of the imputation model is to provide ranges of plausible values. Neither the form of the model nor the parameters estimates are particularly interesting. Additional advantage is that only a subset of the predictor values is used to find the complete case which makes this procedure robust against non-normal linear combinations. We propose to add the bold text as written above to the methods section of the manuscript.

Ad 4.

The reviewer wants more information about how reliable the imputation method is and more insight in the procedures we used for building the imputation model. According to advice given in Schafer (Schafer, JL. *Analysis of Incomplete Multivariate Data*, Chapman & Hall, London, 1997) and Van Buuren et al (Van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Stat Med*. 1999;18(6):681-94), we included the best, say, 15 variables into the imputation model. We first included all study variables in the imputation model, and used a correlation level of 0.2 to include any remaining predictive factors. We built our imputation model by following this strategy for each target variable. In this way we are quite confident that valid imputations were made. Furthermore, in our experience, any incompatibility that may arise out of such modeling specification is almost never a problem. For some limited cases, simulations have shown that using extremely incompatible imputation models did not affect the statistical validity of the final inferences (Van Buuren S, Brand JPL, Groothuis-Oudshoorn CGM, Rubin DB. Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation* 2006, 76: 1049-64). Of course, we can never be sure that such robustness will also hold in our present application, but the evidence obtained thus far strongly suggests that incompatibility is unlikely to be a major problem in general.

Ad 5.

We agree with this reviewer that higher P-values might be responsible for larger unstable models and that this eventually can be lowered by the correction for optimism. As this reviewer suggested we repeated the analyses with a p-values of 0.157 (Akaike's level) and compared these results with the analyses in our manuscript (see table next page). As expected by using a stricter inclusion criterion, the range of variable inclusion frequencies becomes larger, i.e. the separation between strong and weak variables becomes more pronounced. It is reassuring to see that for the strongest variables in the BMC article, i.e. Change in pain intensity, Pain at baseline, Level of functional status at 3 months and Physical activity (> 60% of inclusion frequency), results are comparable. For weaker variables results are less stable. This means that simplifying a prognostic model by our method can be obtained by either choosing a higher inclusion frequency or a more strict selection criterion. We expect that more parsimonious models will need less shrinkage and will show inferior discrimination (Steyerberg EW et al. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Stat Med* 2000; 19: 1059-1079).

	Inclusion frequencies with p = 0.157 (Akaike's level)	Rank	Rank Manuscript
Change in pain intensity	99.10	1	2
Pain at baseline	97.70	2	3
Level of functional status at 3 months	87.95	3	1
Whole body vibration	70.85	4	6
Physical activity	61.30	5	4
Change in functional status	55.90	6	26
Stooping	52.95	7	27
Job demands	52.20	8	7
Sitting	50.15	9	9
functional status at baseline	46.75	10	30
Treatment during enrolment	46.15	11	10
Bending and twisting of the trunk	45.85	12	17
Working with vibration tool	43.70	13	5
Lifting	41.75	14	25
Passive pain coping	41.10	15	11
BMI	39.15	16	14
Job satisfaction	37.65	17	22
Job control	36.85	18	13
Duration of complaints	35.85	19	8
Education	32.10	20	16
Pain radiation	31.60	21	12
Working with hands under knee level	31.15	22	15
Self predicted certainty	28.05	23	19
Social support	27.90	24	18
Fear avoidance	25.40	25	21
Quality of life	24.90	26	23
Gender	24.55	27	20
Active pain coping	23.45	28	28
Age	23.00	29	29
Kinesiophobia	22.55	30	24
Work absence	19.20	31	31

As recommended by this Reviewer we corrected the apparent c-index for optimism. The apparent and corrected estimated are presented in the table. We have also corrected the manuscript with respect to the c-index values and replaced Table 3 with our new Table.

Table 3. The performance of methods 1 to 4 at different levels of proportions of selected indicators.

	MI			B			MI100+BI			MI10+BI						
	n	c-index		slope	n	c-index		slope	n	c-index		slope	n	c-index		slope
		AP	BC			BC	AP			BC	BC			AP	BC	
90%	5	0.74	0.71	0.86	6	0.74	0.71	0.85	3	0.74	0.71	0.86	4	0.72	0.71	0.85
80%	8	0.76	0.71	0.85	8	0.76	0.72	0.83	5	0.74	0.71	0.84	5	0.74	0.71	0.85
70%	10	0.76	0.71	0.77	12	0.78	0.72	0.80	11	0.77	0.72	0.79	11	0.77	0.72	0.80
60%	13	0.77	0.72	0.70	18	0.79	0.71	0.75	27	0.79	0.70	0.67	26	0.79	0.70	0.64
0% (full model)	31	0.80	0.70	0.65	31	0.79	0.69	0.65	31	0.80	0.70	0.64	31	0.80	0.70	0.65

n: number of indicators selected in the multivariable models.

AP: apparent index

BC: bootstrap corrected index

Due to the bootstrap corrections of the apparent c-index in the new Table 3 we propose to remove Figure 1 from the manuscript because all relevant information is now presented in this Table.

Ad 6.

To respond on Reviewer 1 about what bootstrap inclusion fraction threshold is desirable. In the discussion section, paragraph 6 we have covered this topic of choosing an inclusion fraction threshold. This paragraph concludes with the following sentences **“On basis of this recommendation the c-index and the slope among the models in our study with the 70% threshold provides a reasonable trade-off. When a parsimonious model is more important a model that is chosen at a higher inclusion threshold, e.g. 90%, is a good alternative”**.

Minor Essential Revisions.

Ad 7.

In the first paragraph of the Results section at page 10 we stated that there are 354 chronic low back pain cases. We propose to correct this mistake by the sentence **“The overall number of chronic low back pain patients is 493 (of out 628), which is a prevalence of 79%.”** and also added the following information to the manuscript. **For the individual studies the prevalence rates were: trial 1: 111 (of out 134) patients (83%); trial 2: 139 (of out 195) patients (71%) and trial 3: 243 (of out 299) patients (81%) developed chronic low back pain.** Because prevalence rates were quite similar for all studies we did not have to adjust for them.

Reviewer 2 (Yang)

Major Compulsory Revisions

Ad 1.

We agree with the reviewer about adding more discussion regarding the missingness mechanism and MI. We propose to adapt paragraph 7 of the discussion section with the following paragraph (the references refer to the reference list of the revised manuscript):

We assumed that the data were missing at random (MAR). It is, by definition, not possible to test the MAR assumption. The prognostic variables that we have included in our study are fairly comprehensive with respect to their importance in low back pain studies. Using all these data in the imputation model makes the MAR assumption plausible, even if the data are not missing at random [6]. It is therefore reasonable to assume that although some variables might be not MAR, this is ignored by the inclusion of other variables in the imputation model when MI is applied [8]. Furthermore, if there are deviations from the MAR assumption in the data set the question is to what extent this affects the final results. Collins et al. [39] showed in a simulation study that an incorrect MAR assumption only had a minor effect on estimates and standard errors in combination with MI. Van Buuren et al. [40] reported in several strongly MAR incompatible models that the negative effects on estimates after MI were only minimal. On basis of these study results we are fairly confident that we have generated valid imputations and that we were able to make reliable inferences from our data. In our data set also some values with respect to the outcome variable were missing. We choose to impute these missing values within the MI algorithm.

For ref 6 and 8 see manuscript

39. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol Methods* 2001;6(4):330-51.

40. van Buuren S, Brand JPL, Groothuis-Oudshoorn CGM, Rubin DB (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation* 2006, 76: 1049-64.

Ad 2.

For a reaction about the validity of applying the MICE algorithm, we would like to refer to our response to Reviewer 1 (ad. 4) where we discussed the reliability in generating valid imputations by the MICE algorithm. Furthermore, to our opinion there are no studies published on convergence problems when the MICE algorithm is applied. On the other hand, as discussed in the response to Reviewer 1, also Ad. 4 several simulations have shown that MICE performs well and behaves flexible under different missing data patterns.

Ad 3.

We agree with this reviewer that more (simulation) studies have to be conducted in which the new method of applying the bootstrap with stepwise regression has to be compared with other sophisticated methods for the purpose of theoretical justification or simulation-based evaluation. However, this requires a completely new study. The goal of the current study was

to test theoretical insecurities concerning the bootstrap method. For the near future, we have planned some simulation studies to evaluate our procedure under different conditions. Some results are already available, where the performance of the bootstrap is evaluated with other criterion based methods as the BIC or AIC (Holländer N, Augustin NH, Sauerbrei W. Investigation on the Improvement of Prediction by Bootstrap Model Averaging. *Methods Inf Med* 2006; 45: 44–50; (Augustin, N.H., Sauerbrei, W. and Schumacher, M. The practical utility of incorporating model selection uncertainty into prognostic models for survival data. *Statistical Modeling* 2005, 5: 95-118). Conclusions of these studies were that the bootstrap yielded similar results which from a simulation-based perspective justifies the bootstrap procedure.

Minor Essential Revisions

Ad 1.

With regard to the missing values on the outcome variable, we stated in the discussion section (paragraph 7, last sentence) that we imputed these missing values within the MI algorithm.

Ad 2.

We thank this Reviewer for his suggestion, we will do that.

Discretionary Revisions

Ad 1.

We do not completely agree with this reviewer about the limitation of MICE when 30 prognostic indicators are included and with his alternative proposal of using Schafer's Data Augmentation algorithm. We are not aware of careful studies that have shown that MICE fails in the situation described by this reviewer. Moreover, "General location models" as proposed by Schafer often lack flexibility to specify the imputation model or to account for important data features. If the data contain derived variables, for example the calculation of BMI from somebody's height and weight, we cannot fully ensure that the imputation procedure is consistent between the constituent parts. Belin et al. explored the usefulness of the general location model in a mental health services study and concluded that this method was not suitable in combination with numerous variables and a complicated pattern of missing data. (Belin TR, Hu MY, Young AS, Grusky O. Performance of a general location model with an ignorable missing-data assumption in a multivariate mental health services study. *Stat Med*. 1999 Nov 30;18(22):3123-35). As we stated in our response to reviewer 1 (ad. 4), simulation studies suggest that MICE is quite robust.

Ad 2.

This reviewer worries about our justification of the MAR assumption because as he describes many missing values were not due to the design of the study. We agree that some of our missing values could be MNAR. In order to alleviate any such problems, we have included as many relevant predictors as possible in our imputation model. We further agree that more simulation studies are needed to generalize our method to other data sets. But as simulation studies are always somewhat artificial, we highly value the use of new methods on realistic data sets, as we did.

We hope we have convinced your reviewers and yourself with our arguments and that you will now be able to approve our manuscript for publication. Please, don't hesitate to contact me, if you feel that further debate is necessary. I am looking forward to your reaction.

Sincerely yours, on behalf of the co-authors,

Martijn W Heymans

VU University Medical Center
EMGO Institute
Van der Boechorststraat 7
1081 BT Amsterdam
The Netherlands
M: mw.heyman@vumc.nl