ELSEVIER

Journal of Clinical Epidemiology 60 (2007) 979

Journal of Clinical Epidemiology

LETTER TO THE EDITOR

Imputation is beneficial for handling missing data in predictive models

Recently a study was published on bias arising from missing data in predictive models [1]. The author concludes that three methods of handling missing data lead to bias in the estimated regression coefficients of a predictive model: complete case (CC) analysis, ‘‘missing assumed to be normal’’ (MAN) analysis, and imputation analysis. These conclusions are at most partly correct.

When do missing data lead to biased estimates of regression coefficients in predictive models? As the author correctly points out, three mechanisms of missingness are commonly considered: Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR). The author presents a simulation study where four scenarios are considered for fever and white blood cell count (WBC), which predict hospital admission of a patient. Scenario 1 is MCAR, and scenarios 2−4 are variants of MAR scenarios:

2) MAR on outcome (patients with hospital admission more likely to have a WBC),
3) MAR on another predictor (patients with fever more likely to have WBC), and
4) MAR on outcome and another predictor (patients with fever and/or admission more likely to have WBC).

Theoretically, we know that a CC analysis is unbiased in many scenarios: of course under MCAR, but also under MAR on another predictor, and also under a MNAR mechanism [2,3]. The reported bias in scenario 3 must hence be attributed to an error at some level of the simulation program. The CC analysis is only biased with MAR on outcome (scenarios 2 and 4). The MAR on outcome scenario may be less frequent than MAR on another predictor, because we measure predictors before the outcome is known in prediction research [3]. In sum, bias by CC analysis is not a general fact.

Next, the results from the ‘‘missing assumed to be normal’’ analysis are fully determined by the parameters that were set for the simulation. WBC was not actually observed; WBC values were generated in the NHAMCS dataset [1]. Physicians’ decisions may have a high sensitivity to order a WBC test, with WBC testing even if there is a low suspicion of a bacterial infection. If this is true, the MAN analysis would not lead to bias. Indeed, we should ideally have complete WBC measurements of at least a subset of patients to inform us on this missing value pattern [1].

Finally, the question is whether imputation methods can overcome biases of CC and MAN analyses. In the presented analysis, a simplistic conditional mean imputation was performed for a continuous version of WBC. WBC was subsequently dichotomized, which did not work out correctly, because the distribution of imputed WBC values was too narrow. This problem was already known before, as the author points out [2]. Although dichotomizing a predictor is anyway a bad idea in prediction research [4], we can easily think of better alternatives for imputation. We can, for example, use a logistic model to estimate the probability of elevated WBC, and use a variable with imputed probabilities in the predictive model. We could also do this in a multiple imputation procedure, with imputing 0 or 1 for normal vs. elevated WBC values in multiple copies of the database. Instead of logistic regression, we could use a linear regression model to estimate the continuous WBC values, with imputations drawn from the predicted distribution. Likely, imputation with any of these alternatives would have removed any bias in regression models.

In conclusion, this study rightly emphasizes that we should be careful in handling missing data in prediction models. From theory we know when bias may arise in a CC analysis, but that is only in specific situations. In addition to addressing bias, the main advantage of (multiple) imputation may be greater efficiency. A CC analysis is very inefficient when several predictors have a few missing values, but few patients have more than one missing value. In such a situation, imputation of relatively few values leads to an analysis with much more predictive information.

Ewout W. Steyerberg
Center for Medical Decision Making
Department of Public Health, AE 236, Erasmus MC
P.O. Box 2040, 3000 CA Rotterdam, The Netherlands
E-mail address: E.Steyerberg@ErasmusMC.nl (E. Steyerberg)

Mirjam van Veen
Department of Pediatrics, Erasmus MC

References

[1] Gorelick MH. Bias arising from missing data in predictive models. J Clin Epidemiol 2006;59(10):1115−23.

[2] Little RJA, Rubin DB. Statistical analysis with missing data. New York: Wiley; 1987. Wiley series in probability and mathematical statistics. Applied probability and statistics.

[3] Vach W, Blettner M. Missing data in epidemiologic studies. Encyclopedia of biostatistics. New York: John Wiley & Sons; 1998. 2641−54.

[4] Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med 2006;25(1):127−41.