Appendix 1: Explanation of bias when using the missing-indicator method (as supplied by the authors)

Suppose one wants to predict blood pressure as a function of age and sex. Further, suppose the covariate age is missing in a number of observations in the study population. When using the indicator method to handle missing covariate data, the missing values are set to 0 (or any other value as long as the same number is used for all missing observations), and a new variable (indicating missingness) is defined and set to 1 if age is missing and 0 otherwise. Then, not only age and sex but also the new variable (the missingness indicator) are included in a multivariable model:

$$\text{Blood pressure} = b_0 + b_1.\text{age} + b_2.\text{sex} + b_3.\text{Indicator}$$

For those subjects without missing data, the Indicator is zero and the model fitted to the data is:

$$\text{Blood pressure} = b_0 + b_1.\text{age} + b_2.\text{sex}.$$

For those subjects with missing data on the variable age, the Indicator is one and the model fitted to the data is:

$$\text{Blood pressure} = b_0 + b_1.\text{age} + b_2.\text{sex} + b_3.$$

Since all missing values on age are set to zero, this model is in fact:

$$\text{Blood pressure} = b_0 + b_2.\text{sex} + b_3.$$

Clearly, when estimating the regression coefficient of sex (b_2), the relation between age and sex (i.e., the confounding effect of age on the association between sex and blood pressure) is not taken into account in the latter model. This is no problem if age and sex are not related and missingness of age is not related to the outcome conditional on (or given) sex. Then, the regression coefficient of sex (b_2) is correctly estimated, even though age is not included (adjusted for) in the model.

However, when age and sex are related, the estimated regression coefficient of sex (b_2) will be biased, since the relation between sex and age is not adequately taken into account. Formally,

Appendix to: Groenwold RHH, White IR, Donders ART, et al. Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. *CMAJ* 2012. DOI:10.1503/cmaj.110977.

the indicator method will only provide unbiased estimates if either (i) there is no missing data or (ii) the association between (or covariance of) two predictors is zero, meaning that two predictors are not related, and missingness is conditionally independent of outcome.[1,2] In most (nonrandomized) studies this is seldom the case.

References

1. White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Stat Med* 2010;29:2920-31.

2. Jones MP. Indicator and stratification methods for missing explanatory variables in multiple linear regression. *J Am Stat Assoc* 1996;91:222-30.

Appendix to: Groenwold RHH, White IR, Donders ART, et al. Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. *CMAJ* 2012. DOI:10.1503/cmaj.110977.