Tidsskriftet Tidsskrift for Den norske legeforening

Multiple imputation of missing data

MEDICINE AND NUMBERS

STIAN LYDERSEN

stian.lydersen@ntnu.no
Stian Lydersen, PhD, professor of medical statistics at the Regional Centre for Child and Youth Mental Health and Child Welfare, Department of Mental Health, Norwegian University of Science and Technology.
The author has completed the ICMJE form and declares no conflicts of interest.

Most statistical methods of analysis require complete data sets, but in nearly all studies some values are missing. Multiple imputation can be used to handle this.

Let us consider a data set where each row in the spreadsheet contains data from a study participant. It is common that some values in the data set are missing, here marked as X in Figure 1. For example, some participants may have data missing for weight, others for physical activity and others for marital status. All these are included as independent variables in the *analysis model*, in this case a regression model with blood pressure as dependent variable. If we analyse the data set without taking account of missing data, most analysis methods will include only participants with complete data, which is known as a *complete case analysis*. This is scarcely a problem if data are missing for only a few of the participants, for example fewer than 5 % or 10 %, but else, the statistical power will be substantially reduced because of the smaller sample size. A more serious problem, however, is that a complete case analysis will be unbiased only if data are *missing completely at random* (MCAR). An analysis based on multiple imputation, on the other hand, will be unbiased also if data are *missing at random* (MAR) (1).

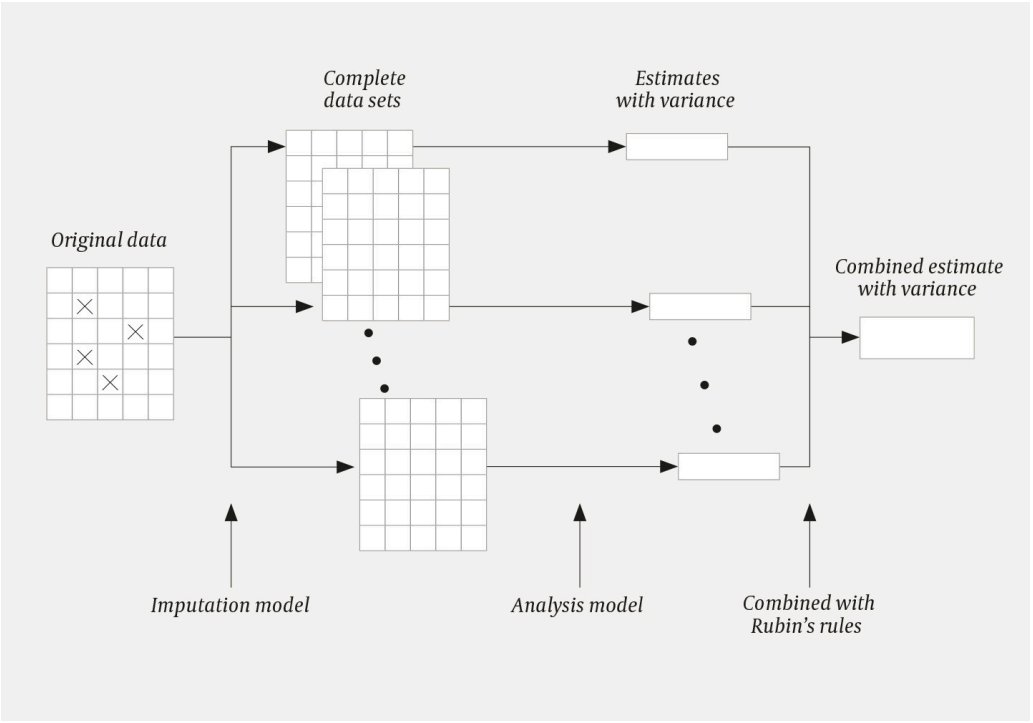

Figure 1 Procedure for multiple imputation.

The imputation model

Imputation of data means inserting estimates for missing values in the data set. These values are estimated based on other variables in the analysis model. This is normally done using linear regression models for continuous variables, such as weight, and with logistic regression models for dichotomous variables, such as marital status. Together, these regression models are referred to as the *imputation model*. A number of complete data sets are produced, which take account of the uncertainty of the imputed values by letting them vary between the data sets. It is generally recommended to produce from 20 to 100 complete data sets (2). The number is strongly dependent on the extent of the missing data. To be on the safe side, one may well use 100 imputed data sets, unless this requires excessive computation time.

The analysis model

After imputation, each of the complete data sets are analysed using the analysis model. Each analysis yields an estimate with an associated variance for the quantity or quantities of interest, for example the regression coefficient for physical activity. Finally, the estimates and variances are combined by applying specific rules, the so-called 'Rubin's rules' (1). The combined estimate is equal to the mean of the estimates. The combined variance is equal to the mean of the variances, plus a term that accounts for the variation between the imputed data sets. Thereafter, the confidence interval and *p*-value can be estimated. The procedure is illustrated in Figure 1. Many quantities can be combined using

Rubin's rules, such as the mean, standard deviation, proportion and regression coefficient. Odds ratios and hazard ratios should be logarithmically transformed before they are combined.

Fictitious data?

The imputation model must be thoroughly planned. All the variables to be included in the analysis model, including the dependent variable, must be included in the imputation model. Additional, so-called auxiliary variables can also be included. This is relevant if there are variables available that are not included in the analysis model, but are associated with variables with missing data. If the analysis model contains non-linear functions or interactions, these need to be handled in special ways (3). If the interaction contains a dichotomous covariate, such as sex, the imputation can be done in a separate file for each sex and the two imputed files subsequently merged (2). In general, the imputation model requires considerably more thinking and computation time than the analysis model itself.

Does multiple imputation imply that non-existent data are simply made up? Quite the reverse; all the participants are included in the analysis, including those for whom data are missing in one or more of the variables. Moreover, by using multiple imputation we achieve higher statistical power and less bias in the estimates than in a complete case analysis.

REFERENCES

1. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med 2011; 30: 377–99. [PubMed][CrossRef]

2. van Buuren S. Flexible imputation of missing data. 2 utg. Boca Raton, FL: CRC Press, 2018: 175-6.

3. Seaman SR, Bartlett JW, White IR. Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. BMC Med Res Methodol 2012; 12: 46. [PubMed][CrossRef]

Publisert: 26 January 2022. Tidsskr Nor Legeforen. DOI: 10.4045/tidsskr.21.0772