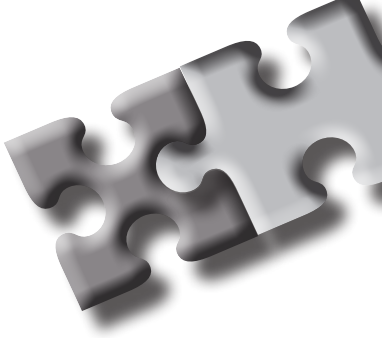

Summary

English

Don't Miss Out!
Incomplete data can contain valuable information

In epidemiological research, patient reported outcomes are often measured by a multi-item questionnaire. In a multi-item questionnaire a construct is measured by combining the scores on several items (i.e., questions). Often these questionnaires contain missing data because one or several items are not filled out by the respondent, or the entire questionnaire was not filled out. Missing item scores might require different missing data methods than missing total scores.

The underlying reasons for missing data can be differentiated in so called missing data mechanisms. Missing data can be missing completely at random (MCAR) when the missing part of the data is a completely random subsample of the data, for example when a questionnaire gets lost in the mail. However, when the probability of missing data is related to other measured variables in the data, data are missing at random (MAR). For example when physical activity scores are more often missing for the older people, then the missings are related to age. Missing data are missing not at random (MNAR) when the missing data are related to the missing values itself, for example when people with lower scores on physical activity have a missing physical activity score. The performance of the missing data methods is dependent on the underlying missing data mechanism. For that reason it is important to make a valid assumption about the most probable missing data mechanism by investigating the data and think about probable reasons for the missing data.

Missing data in epidemiological studies are most frequently handled by a complete-case analysis. Moreover, in manuals of multi-item questionnaires it is often advised to replace a missing item score with a single value (e.g., a mean score). However, these methods do not perform well and cause biased study results irrespective of the missing data mechanism or the amount of missing data. Advanced methods to handle missing data are multiple imputation and full information maximum likelihood estimation. These methods work well with MCAR and MAR data. In multiple imputation the missing values are replaced by multiple plausible values, resulting in multiple copies of the dataset with each time different imputed values. The plausible values are estimated from the observed data with regression techniques. Item scores in a multi-item questionnaire are often measured by a Likert scale. Consequently, items are ordinal and are not necessarily normally distributed. Accordingly, an imputation method based on linear regression might not always suffice. The predictive mean matching procedure is robust against the deviation from the normal distribution and imputes more realistic values compared with linear regression. Predictive mean matching randomly draws from the observed data values that are closest to the predicted estimate from the regression equation.

The imputed datasets are each analyzed according to the main analysis model (i.e., the analysis that would have been performed had the data been complete). The multiple sets of results are combined as the final analysis results. In full information maximum likelihood the population parameter values are obtained that would most likely produce the sample of data that is analyzed. In this method no values are imputed, but all observed data are used to obtain the estimates. Both these methods are considered as the state-of-the-art methods to handle missing data.

The best method to deal with missing data depends on the analysis method that is applied to analyze the dataset (i.e., longitudinal analysis or not), the type of variable in the main analysis model (i.e., predictor/covariate or outcome), the missing data mechanism (i.e., MCAR, MAR, MNAR), the overall percentage of subjects with missing data, and the level of missing data in the questionnaire (i.e., item score or total score missings).

Missing data in a multi-item questionnaire should be handled on the item level of the questionnaire. When the outcome of the study is measured at one time-point, multiple imputation of the items should be applied. This means that the item variables with missing values are imputed and after the multiple imputation process, the total scores for the questionnaires can be calculated and analyzed.

In studies where many questionnaires or extremely large questionnaires are used, the number of item variables will become too large to reliably estimate imputations. Passive imputation can be a solution to this problem. Passive imputation methods combine variables in the imputation model to reduce information. The item scores of one questionnaire are imputed, while the total scores of other questionnaires are used as predictors. These total scores may contain missing values caused by missing item scores as well, and will be imputed with the same method. The total scores will be updated after each imputation run (i.e., iteration) using the imputed item scores.

When the outcome is measured at multiple time-points, the analysis method should take the correlation between the multiple measurements into account. Longitudinal analysis methods often use full information maximum likelihood procedures to obtain the parameter estimated and these procedures handle the missing data in the analysis. Usually, a longitudinal analysis with a multi-item questionnaire outcome, uses only the total scores of the questionnaire in the analysis. However, when total scores are incomplete due to missing item scores, the missing data should be handled at the item level. The item-level information can be included as auxiliary variables in the analysis. That way the parameter estimates are more precise and accurate. The missing data in the predictors or covariates in a longitudinal analysis should be handled by multiple imputation.

The advice with respect to multi-item questionnaires can also be used for other purposes. For example cost-data where the total costs are used in a cost-effectiveness

analysis. These total costs can be missing due to missing sub-costs and accordingly the missing data handling is best handled at the sub-cost level. Furthermore, the distribution of cost data is almost never a normal distribution. Costs are constrained to be positive and often skewed to the right with an excess of zeroes. The imputation strategy can therefore be adapted by using predictive mean matching on the log-transformed costs to handle the extreme skewness. After the imputation, the data can be transformed back and analyzed.

The most important conclusions of this thesis are that the methods that are advised in manuals for multi-item questionnaires are often sub-optimal and should be ignored. Missing data in multi-item questionnaires should always be handled at the item level. Using the information from the observed item scores in the missing data handling method improves accuracy and precision of analysis results. Furthermore, including item information as auxiliary variables to handle missing data in longitudinal models that analyze the total scores improves precision and power of coefficient estimates. And applying passive imputation to impute the item scores when the number of items is extremely large is a valid method to handle missing item scores in large survey studies.

Overall, in handling missing data in multi-item questionnaires it is important to incorporate all available information from the item scores in order to obtain the optimal level of accuracy and precision in study estimates.