Iris Eekhout

Don't Miss Out!

Incomplete data can contain valuable information

Don’t Miss Out!

Incomplete data can contain valuable information

The studies described in this theses were performed at the EMGO+ Institute for Health and Care Research and the Department of Epidemiology and Biostatistics at the VU University medical center in Amsterdam, the Netherlands.

www.missingdata.nl

Cover design: Elise Eekhout, Studio 2 MAAL EE, www.2maalee.nl
Printed by: Ridderprint BV, Ridderkerk

ISBN: 978-90-5335-964-8

VRIJE UNIVERSITEIT

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

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. F.A. van der Duyn Schouten,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Geneeskunde
op dinsdag 6 januari 2015 om 15.45 uur
in de aula van de universiteit,
De Boelelaan 1105

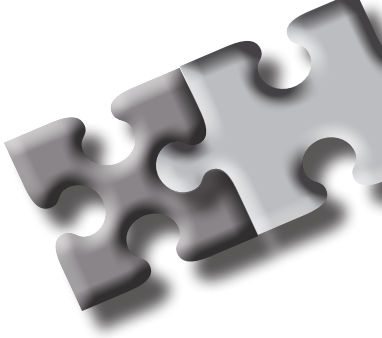
door

Iris Eekhout

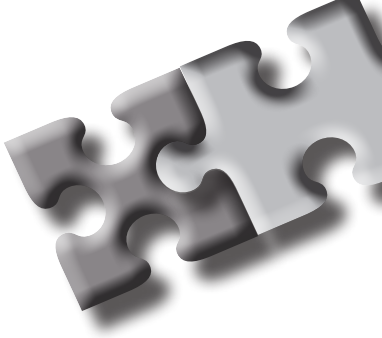
geboren te Wateringen

promotoren:	prof.dr. J.W.R. Twisk
	prof.dr.ir. H.C.W. de Vet
copromotoren:	dr. M.W. Heymans
	dr. M.R. de Boer

voor mama



Contents



Chapter 1

Introduction

Under review as introduction to a review article: Eekhout, I., de Vet, H.C.W., de Boer, M.R., Twisk, J.W.R., Heymans, M.W. Missing data in multi-item questionnaires: analyze carefully and don't waste available information. International Journal of Epidemiology.

Many empirical studies encounter missing data problems. Missing data occurs when a data value is unavailable and can occur in many stages of research and data situations. Missing data can take place on one or more of the measured variables that are used as a predictor, covariate or outcome. In the case that participants in a longitudinal study do not show up at repeated measurement occasions, the missing data are often referred to as loss to follow up or intermittent missing data. Missing data can also occur in a multi-item questionnaire due to questions that have not been filled out by the participant. In that case some items can be missing or the entire questionnaire might not be filled out. These examples of missing data can have different underlying causes and require different solutions.

Study designs and missing data

In the field of epidemiology many different sorts of studies are performed using different designs (Rothman, 2012). One way to distinguish study designs is by the outcome measurement, which can be assessed at one or at multiple time-points. In a cross-sectional study the outcome variable is measured at the same time as the covariate. The relations in these studies are usually analyzed in a regression model or with other simple statistical tests as t-test or analysis of variance. Another study design that is often applied in epidemiology and medical studies is the randomized controlled trial (RCT). In RCTs the sample is randomly divided over treatment groups. Prior to the treatment a measurement is often performed to register the baseline status of the study participants. Post treatment a second measurement is performed to measure the effect of the treatment, which is the study outcome. Usually a regression analysis is performed using the post treatment measurement as outcome, predicted by the treatment group which can be corrected for the baseline measurement.

RCTs often contain multiple follow-up measurements, hence the outcome is measured multiple times, in which case the study is longitudinal. In these studies the long-term effect of a treatment or intervention can be analyzed, as well as the change over time related to the treatment group or other covariates in the study. A longitudinal study can also be observational, where the change over time is related to baseline characteristics or predictors. These longitudinal studies require analysis techniques that take the correlation between the time-points into account (Twisk, 2013).

In the study designs mentioned above, missing data can occur in the predictors, the covariates and/or in the outcomes. In the studies that have the outcome measured just once, the consequences for the missing data in either type of variable in the main analysis is similar. However, in longitudinal analysis, missing data in the predictors or covariates might require different solutions than missings in the outcomes of the study. Furthermore, patient-reported data are often collected by

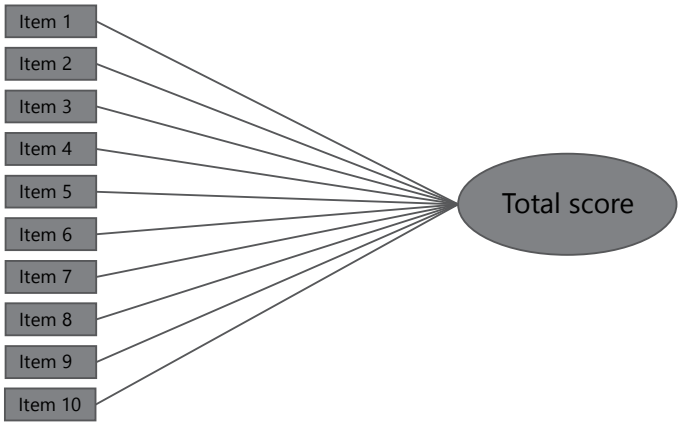


Figure 1.1. Example of a multi-item questionnaire with 10 items that result in a total score.

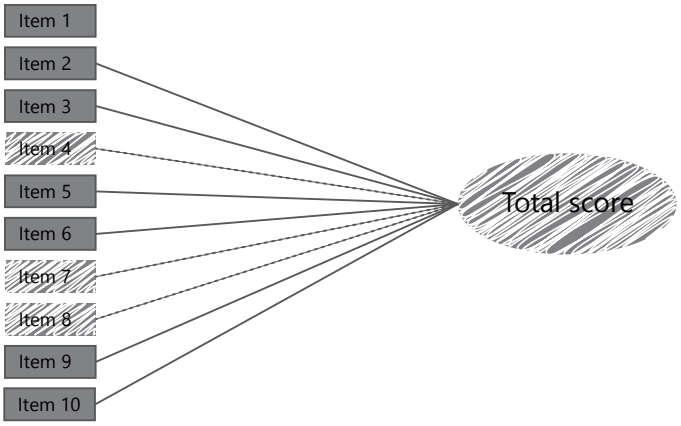


Figure 1.2. Example of a multi-item questionnaire with 3 out of 10 item scores missing that result in an incomplete total score.

multi-item questionnaires where the entire questionnaires data can be missing or only a part of the questionnaire. In the latter case some of the information is still available. Most missing data research is focused on missing data methods applied to total values; not many studies have focused on missing data methods for multi-item questionnaires.

Missing data in multi-item Questionnaires

Multi-item questionnaires often measure one underlying unobservable construct by several observable characteristics (i.e., items). Accordingly, the items are reflections

of the construct. The scores on the items are combined (e.g., by summing the item scores) into one total or scale score that represents the construct as presented in the example in Figure 1.1. This relationship between the unobservable construct and the items is called a reflective model (de Vet, Terwee, Mokkink, & Knol, 2011). These multi-item questionnaires are often used in epidemiological studies to measure patient-reported outcomes. Examples of such outcomes are physical functioning, measured by a subscale of the SF-36 (Ware, Kosinski, & Keller, 1994) or pain coping, measured by the pain coping inventory (PCI; Kraaimaat & Evers, 2003). Patient-reported outcomes are used as study outcomes, but also as covariates or predictors in studies.

In multi-item questionnaires, missing data can occur at two levels. These are the total score level, when respondents do not fill out the entire questionnaire, or the item level when respondents skip some questions (i.e., items) of the multi-item questionnaire. The missing data at the item level can result in missing total score data, because the missing item scores hamper the total score calculation as presented in Figure 1.2. In that situation, when one or more item scores are missing, the total score is missing as well. In most empirical studies that use multi-item questionnaires both kinds of missing data occur. Researchers usually do not distinguish between these two kinds of missing data in multi-item questionnaires when they use a method to handle the missing data (Eekhout, de Boer, Twisk, de Vet, & Heymans, 2012).

Manuals of multi-item questionnaires often contain an advice on how to handle missing item scores on that particular questionnaire. Mostly these advices are aimed at replacing the missing value with simple handling methods. For example the manual of the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) instructs to replace the missing item score with the mean subscale score when three or less items are missing and when four or more items are incomplete to leave the total score incomplete (Bellamy, 2000). A similar recommendation is stated in the manual for the Symptoms Checklist (SCL-90) where a missing item score is to be replaced with the average over the completed items by the criterion of replacing only one missing for every five complete items in the subscale (Hardt, Gerbershagen, & Franke, 2000). The SF-36 manual advises to calculate the average over the available items for the total scores, thus imputing the mean score over the completed items (Ware, et al., 1994). Other questionnaires advise to leave the total score missing when one or more items are incomplete (e.g., EuroQol-5D; The EuroQol Group, 1990).

Missing data mechanisms

The underlying reasons for missing data can be differentiated in so called missing data mechanisms. Rubin (1976) formulated three possible missing data mechanisms. 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 questionnaires are lost in the mail. Another possibility is that the data are missing at random (MAR). In that mechanism the probability of missing data is related to other measured variables in the dataset. For example when data are missing for physical functioning and the data are mostly missing for the older people in the dataset. In that case the probability of missing data on physical functioning is related to age. A third mechanism is missing not at random (MNAR). When data are MNAR the probability of missing data is related to the value of the missing data itself. For example, when only the lower physical functioning scores are missing, then the probability of missing data on physical functioning is related to the physical functioning score itself.

It is important to have a good understanding of the missing data mechanism, because the performance of missing data handling methods depends on assumptions about the missing data mechanism. There are two main ways that can help to make an assumption about the missing data mechanism: common sense and statistical approaches. The first and most important one is 'common sense'. Most researchers have an idea about the reasons for the missing data, by what is known about the data collection process and the data in general. Furthermore, it is advisable to collect as much information as possible about the reasons why data are missing (Curran, Bacchi, Schmitz, Molenberghs, & Sylvester, 1998). It is very important to take this knowledge into account when making an assumption about the missing data mechanism. The second possibility is to compare the characteristics of the data related to missingness with a statistical analysis. For example, by comparing the characteristics of the group with missing values on a certain variable, to the characteristics of the group with observed values on that variable using a t-test. When missing data are not MCAR these groups will have different mean values. Another example is to use an indicator for missing data (i.e., a dichotomous variable) as outcome in a logistic regression model to find variables related to the probability of missing data, which may be an indication that the data are not MCAR (Ridout, 1991). It is important to note that these methods can only distinguish between MCAR or not-MCAR mechanisms. However, it is not possible to test whether the missing data are MAR or MNAR, because there is no information about the missing data itself available. Furthermore, the validity of significance tests by using a t-test or a logistic regression model highly depends on the sample size. Therefore these tests are only indicative of the assumed missing data mechanism but can never be conclusive about that. For that reason it is recommended to combine statistical testing of the missing data mechanism with additional collected information about the underlying reasons that caused the missing data.

Missing data methods

Traditional methods

The method that is most often applied in epidemiological studies to deal with missing data is a complete-case analysis (CCA) (Eekhout, et al., 2012). In a CCA the subjects with completely observed data are included in the analysis; the subjects who have some data missing are simply not used. This method is easy to apply and is still the default method in many statistical packages (e.g., SPSS; SPSS Inc., 2008). Results from a CCA are only unbiased when the missing data are MCAR (Rubin, 1976). However, in any case the sample size is reduced in a CCA, so statistical power will be suboptimal.

In order to retain the original sample size it is possible to impute the missing values. That way the missing data entries are replaced with a value that is usually estimated from the observed data. In multi-item questionnaire data, imputation strategies can be applied to either the item scores or the total scores. When the imputation strategy is applied to the item scores, the missing item scores are imputed first and after that imputation, the total scores are calculated. These total scores are then used in the data analysis. When the imputation method is directly applied to the total score the total scores are first calculated for the persons without missing item scores, then the missing total scores are replaced with an imputed value and these imputed total scores are used in the analysis.

One of the most frequently observed single imputation methods is to replace the missing values with a mean score. When this imputation method is applied to the item scores the imputed values can be the average score that is observed for each particular item in the study sample. This is called item mean imputation (Hawthorne & Elliott, 2005). Another way is to impute the average score on all observed items for each subject in the data, i.e., the average over the available items. This is known as person mean imputation (Bernaards & Sijtsma, 2000; Fayers, Curran, & Machin, 1998; Hawthorne & Elliott, 2005). A method that combines both of these strategies is two-way imputation (van Ginkel, Sijtsma, van der Ark, & Vermunt, 2010). In that method the item and person means are added, and then, the overall mean score on the questionnaire is subtracted. Instead of applying the mean imputation method to the items, the total score can also be imputed directly by the average observed total score in the sample. Imputing the mean score via any of these strategies decreases the variability in the data and will ultimately cause biased results for any of the missing data mechanisms and is therefore not recommended to use (Eekhout et al., 2014; Schafer & Graham, 2002).

A single imputation strategy that restores the variability in the data is stochastic regression imputation (SRI). In this method the imputed value is estimated via a

regression equation from the observed data. Subsequently, a random error term that is drawn from a normal distribution around the estimated value is added to the estimated value (Roth, Switzer, & Switzer, 1999). SRI can also be applied to the item scores or directly to the total scores. This method is the only single imputation method that performs reasonably well in a MAR mechanism (Eekhout, et al., 2014; Enders, 2010).

However, in none of the single imputation methods the uncertainty around the missing data is included (Gold & Bentler, 2000). In single imputation it is assumed that the single imputed value is the correct one (i.e., the true values that are missing) and the precision is overstated. However, there can never be absolute certainty about validity of the imputed values and therefore uncertainty around these imputed values has to be incorporated in the missing data method (Little & Rubin, 1989).

Advanced methods

Multiple imputation

A well-known advanced method that incorporates the uncertainty around the imputed values is multiple imputation. In multiple imputation multiple plausible values are imputed resulting in multiple datasets with different imputed values in each set. The analyses are performed in each of these completed datasets and the analysis results are pooled to obtain the final data results (Rubin, 1987; Schafer, 1999; van Buuren & Groothuis-Oudshoorn, 2011). Accordingly, multiple imputation is performed in three phases. In the first phase, the imputation phase, the missing values are replaced with multiple plausible values. These values are estimated from the observed data by a multivariable model, which is called the imputation model. The specific imputation method that is used to estimate the imputed values can be adjusted to the distribution of the variable that needs to be imputed. Accordingly, continuous variables can be imputed by using a linear regression algorithm, dichotomous variables by a logistic regression algorithm, and ordinal variables by a proportional odds model. Frequently, continuous empirical data are not normally distributed. A method that handles deviations from normal distributions well is predictive mean matching. In this method the imputed values are sampled from the observed values. The individuals with observed values that are closest to the predicted values from the imputation model are identified and the imputed value is randomly drawn from these individuals. The advantage is that the imputed values are close to the values of the observed data (Little, 1988). Predictive mean matching is the default method for multiple imputation in the mice function in R statistical software (van Buuren & Groothuis-Oudshoorn, 2011).

The process of estimating plausible values is performed sequentially for each

variable with missing values in the dataset using a so called chain of regression equations. So for the missing values the plausible values are estimated from these regression equations. This process is performed sequentially for each variable that contains missing values within one chain (i.e., iteration). Generally, this iteration process is repeated multiple times, while each time using the imputed values from the previous run. After the specified number of iterations are performed the first imputed dataset is set aside. This whole procedure is then repeated for the next imputed dataset, until the specified number of imputed datasets are created. This algorithm for multiple imputation is called multivariate imputation by chained equations (MICE) (van Buuren, 2012; White, Royston, & Wood, 2011)

The imputation model has to contain all variables that are of interest in the main analysis. The main analysis is here the analysis that would have been performed had the data been complete, so all relevant predictors, covariates and the outcome should be included. Additionally, other variables can be relevant to the missing data (Meng, 1994). These variables are also referred to as auxiliary variables (Collins, Schafer, & Kam, 2001). Auxiliary variables are variables that are related to the incomplete variables or to the probability of missing values in a variable. Auxiliary variables can help improve the prediction of missing data and therefore they can mitigate bias and improve power. In the example where the older people in the sample have more missing values on their physical functioning score, the variable age is related to missingness and might therefore be a relevant auxiliary variable when the physical functioning scores are imputed. Including auxiliary variables in the missing data handling procedure is nearly always beneficial (Collins, et al., 2001).

In the analysis phase of multiple imputation, each imputed dataset is analyzed separately by the main analysis model. The performed main analysis is the same analysis that would have been applied had the data been complete. This results in multiple sets of results, which differ because the imputed datasets differ from each other. After the analysis phase the results are combined in the pooling phase by Rubins Rules (Rubin, 1987). For parameter estimates (e.g., regression coefficients), the combined estimate θ is the average of the estimates in each imputed dataset:

$$\theta = \frac{\sum_{j=1}^{m} \theta_j}{m}$$

The number of imputed datasets is denoted by m. The standard error of the parameter estimates is combined by using the within-imputation variance and the between-imputation variance. The within imputation variance $Var(\theta)_{within}$ is the average variance from the imputed data analyses which estimates the sample variability:

$$Var(\theta)_{within} = \frac{\sum_{j=1}^{m} Var(\theta)}{m}$$

The between imputation variance is the variance between the estimates from the imputed datasets, which represents the additional sampling error that results from the missing data. The between imputation variance $Var(\theta)_{\text{between}}$ is calculated by the sum of the squared deviation of the parameter estimate obtained in each imputed dataset from the pooled parameter estimate weighted by 1 over the number of imputations minus one:

$$Var(\theta)_{between} = \frac{\sum_{j=1}^{m}\left(\theta_j - \bar{\theta}\right)^2}{m-1}$$

The standard error of the parameter estimates is then calculated by combining the within and between variance as follows:

$$SE(\theta) = \sqrt{Var(\theta)_{within} + \left(1 + \frac{1}{m}\right) Var(\theta)_{between}}$$

Full Information Maximum Likelihood

As previously mentioned, in longitudinal data situations, analysis methods are needed that take the design of repeated measures within a person into account. The estimation methods in these kinds of methods are often based on full information maximum likelihood (FIML). FIML estimation is used to obtain the population parameter values that would most likely produce the sample of data that is analyzed. This is done by an iterative process that repeatedly tests different parameter values until the fit to the data is most optimal. In case of missing data no values are imputed, but the estimation process to obtain parameter values is done with all of the observed data (Enders, 2010; Little & Rubin, 2002; Schafer, 1997). FIML estimation produces unbiased estimates under a MAR mechanism and is also better than traditional methods in MCAR situations (e.g., complete-case analysis), because power is maximized by using all available information in the data (Schafer & Graham, 2002). Analysis methods that can use FIML are mixed models and structural equation models. Both procedures can be used to analyze repeated measures data (Kwok et al., 2008).

When multi-item questionnaires are used as the outcome in a longitudinal analysis, however, only the total scores will be used, ergo the item scores are usually not taken into account. The total scores that are used in the main analysis are left incomplete when one or more item scores are missing. The available item information is then ignored, while from previous studies it is known that it is best to include all available item information in the missing data handling method (Eekhout, et al., 2014; Gottschall, West, & Enders, 2012). For that reason, the item information can be included in the auxiliary part of the model (Eekhout et al., in press). This means that

the item information is included as auxiliary variables. As previously mentioned, in the context of multiple imputation the auxiliary variables are simply included in the imputation model during the imputation phase. In the analysis phase the auxiliary variables are not of influence in the interpretation of the final estimates of the main analysis. In a model that uses FIML estimation, the auxiliary variables should be included in the main analysis, because that is where the missing data are handled. An auxiliary variable can be included as an additional predictor in the main analysis; however, this method would change the interpretation of the parameter estimates. As an alternative, the auxiliary variables should be included so that the interpretation of the parameter estimates is the same as it would have been had the data been complete. One way to do this is by using a structural equation model to analyze the data and include auxiliary variables as described by Graham (2003). Accordingly, the rules for including auxiliary variables in a structural equation model are to correlate the auxiliary variables with (1) measured predictor and covariate variables, (2) other auxiliary variables, and (3) with the residual terms of the measured outcome variables. The resulting parameter estimates have the same interpretation as the complete data analysis results, but the power has increased due to the item information that is included (Eekhout, et al., in press). For examples of applications of structural equation models for longitudinal data that include auxiliary item information to deal with missing data see Eekhout et al. (under review).

References

Bellamy, N. (2000). WOMAC osteoarthritis index: user guide IV. Brisbane, Australia.

Bernaards, C. A., & Sijtsma, K. (2000). Influence of Imputation and EM Methods on Factor Analysis when Item Nonresponse in Questionnaire Data is Nonignorable. Multivariate Behavioral Research, 35(3), 321-364.

van Buuren, S. (2012). Flexible Imputation of Missing data. New York: Chapman & Hall/ CRC.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological Methods, 6(4), 330-351.

Curran, D., Bacchi, M., Schmitz, S. F., Molenberghs, G., & Sylvester, R. J. (1998). Identifying the types of missingness in quality of life data from clinical trials. Stat.Med., 17(5-7), 739-756.

Eekhout, I., de Boer, R. M., Twisk, J. W., de Vet, H. C., & Heymans, M. W. (2012). Missing data: a systematic review of how they are reported and handled. Epidemiology, 23(5), 729-732.

Eekhout, I., De Vet, H. C. W., Twisk, J. W. R., Brand, J. P. L., De Boer, M. R., & Heymans, M. W. (2014). Missing data in a multi-item instrument were best handled by multiple imputation at the item score level. Journal of Clinical Epidemiology, 67(3), 335-342.

Eekhout, I., Enders, C. K., Twisk, J. W., De Boer, M. R., De Vet, H. C., & Heymans, M. W. (under review). Longitudinal data analysis with auxiliary item informaiton to handle missing questionnaire data. Journal of Clinical Epidemiology.

Eekhout, I., Enders, C. K., Twisk, J. W. R., De Boer, M. R., de Vet, H. C. W., & Heymans, M. W. (in press). Analyzing Incomplete Item Scores in Longitudinal Data by Including Item Score Information as Auxiliary Variables. Structural Equation Modeling: A Multidisciplinary Journal.

Enders, C. K. (2010). Applied Missing Data Analysis. New York, NY: The Guilford Press.

Fayers, P. M., Curran, D., & Machin, D. (1998). Incomplete quality of life data in randomized trials: missing items. Stat.Med., 17(5-7), 679-696.

van Ginkel, J. R., Sijtsma, K., van der Ark, L. A., & Vermunt, J. K. (2010). Incidence of missing item scores in personality measurement, and simple item-score imputation. Methodology, 6(1), 17-30.

Gold, M. S., & Bentler, P. M. (2000). Treatments of Missing Data: A Monte Carlo Comparison of RBHDI, Iterative Stochastic Regression Imputation, and Expectation-Maximization. Structural Equation Modeling: A Multidisciplinary Journal, 7(3), 319-355.

Gottschall, A. C., West, S. G., & Enders, C. K. (2012). A Comparison of Item-Level and Scale-Level Multiple Imputation for Questionnaire Batteries. Multivariate Behavioral Research, 47(1), 1-25.

Graham, J. W. (2003). Adding Missing-Data-Relevant Variables to FIML-Based Structural Equation Models. Structural Equation Modeling: A Multidisciplinary Journal, 10(1), 80-100.

Group, T. E. (1990). EuroQoL-a new facility ffor the measurement of helth-related quality of life. Health Policy, 16(3), 199-208.

Hardt, J., Gerbershagen, H. U., & Franke, P. (2000). The symptom check-list, SCL-90-R: its use and characteristics in chronic pain patients. European Journal of Pain, 4(2), 137-148.

Hawthorne, G., & Elliott, P. (2005). Imputing cross-sectional missing data: comparison of common techniques. Aust.N.Z.J.Psychiatry, 39(7), 583-590.

SPSS Inc. (2008). SPSS Statistics for Windows (Version Version 17.0). Chicago: SPSS Inc.

Kraaimaat, F. W., & Evers, A. W. (2003). Pain-coping strategies in chronic pain patients: psychometric characteristics of the pain-coping inventory (PCI). Int J Behav Med, 10(4), 343-363.

Kwok, O. M., Underhill, A. T., Berry, J. W., Luo, W., Elliott, T. R., & Yoon, M. (2008). Analyzing Longitudinal Data with Multilevel Models: An Example with Individuals Living with Lower Extremity Intra-articular Fractures. Rehabil Psychol, 53(3), 370-386.

Little, R. J. (1988). Missing-data adjustments in large surveys. Journal of Business and Economic Statistics, 6, 287-296.

Little, R. J., & Rubin, D. B. (1989). The Analysis of Social Science Data with Missing Values. Sociological Methods & Research, 18(2-3), 292-326.

Little, R. J., & Rubin, D. B. (2002). Statistical Analysis with Missing Data (Second Edition ed.). Hoboken, NJ: John Wiley & Sons.

Meng, X.-L. (1994). Multiple-Imputation Inferences with Uncongenial Sources of Input. 538-558.

Ridout, M. S. (1991). Testing for random dropouts in repeated measurement data. Biometrics, 47(4), 1617-1619; discussion 1619-1621.

Roth, P. L., Switzer, F. S., & Switzer, D. M. (1999). Missing Data in Multiple Item Scales: A Monte Carlo Analysis of Missing Data Techniques. Organizational Research Methods, 2(3), 211-232.

Rothman, K. J. (2012). Epidemiology: An Introduction: OUP USA.

Rubin, D. B. (1976). Inference and missing data. Biometrika, 63(3), 581-592.

Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: J. Wiley & Sons.

Schafer, J. L. (1997). Analysis of Incomplete Multivariate Data. London, UK: Chapman & Hall.

Schafer, J. L. (1999). Multiple imputation: a primer. Stat.Methods Med.Res., 8(1), 3-15.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. Psychol.

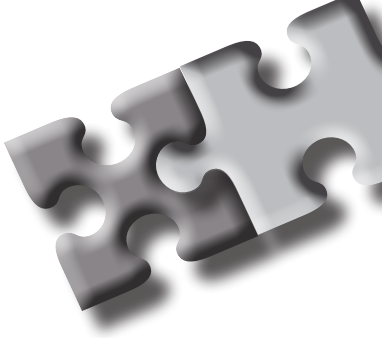
Methods., 7(2), 147-177.

Twisk, J. W. R. (2013). Applied Longitudinal Data Analysis for Epidemiology: A Practical Guide, Second Edition. New York: Cambridge University Press.

de Vet, H. C. W., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011). Measurement in Medicine. Cambridge: Cambridge University Press.

Ware, J. E., Jr., Kosinski, M., & Keller, S. D. (1994). SF-36 Physical and Mental Health Summary Scales: A User's Manual. Boston, MA: Health Assessment Lab.

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. Stat Med, 30(4), 377-399.



Chapter 2

Missing data: a systematic review of how they are reported and handled

Published: Eekhout, I., de Boer, M.R., Twisk, J.W.R., de Vet, H.C.W., & Heymans, M.W. (2012). Missing data: a systematic review of how they are reported and handled. Epidemiology, 23(5), 729-732.

Abstract

The objectives of this systematic review are to examine how researchers report missing data in questionnaires and to provide an overview of current methods for dealing with missing data. We included 262 studies published in 2010 in three leading epidemiological journals. Information was extracted on how missing data were reported, types of missing, and methods for dealing with missing data. 78% of studies lacked clear information about the measurement instruments. Missing data in multi-item instruments were not handled differently from other missing data. Complete-case analysis was most frequently reported (81% of the studies), and the selectivity of missing data was seldom examined. Although there are specific methods for handling missing data in item scores and in total scores of multi-item instruments, these are seldom applied. Researchers mainly use complete-case analysis for both types of missing data, which may seriously bias the study results.

Keywords: bias (epidemiology), data interpretation, statistical, epidemiological research design, missing data, questionnaires, regression analysis, research report

Introduction

Missing data are a problem in most epidemiological studies, especially with questionnaires containing multi-item instruments. Multi-item instruments measure one construct with multiple items (de Vet, Terwee, Mokkink, & Knol, 2011); for example, the CES-D uses 20 items to assess depressive symptoms (Radloff, 1977). On multi-item instruments, several or all item scores can be missing. Single-item instruments assess constructs by one question, for example pain by a visual analog scale. Missing cases, when eligible subjects do not fill out or return the questionnaire, can also occur (McKnight, McKnight, Sidani, & Figueredo, 2007).

Missing total scores on multi-item instruments are equivalent to missing scores on a single-item instrument. Commonly used methods to deal with such missingness are complete-case analysis, mean imputation, or single regression imputation. More advanced techniques that account for missing data uncertainty are multiple imputation or maximum likelihood estimation (Baraldi & Enders, 2010; DeSouza, Legedza, & Sankoh, 2009; Donders, van der Heijden, Stijnen, & Moons, 2006; Enders, 2010; Greenland & Finkle, 1995; McKnight et al., 2007). Specific methods have also been developed for missing-item scores in multi-item instruments, for example, person mean imputation, two-way imputation, response-function imputation, and multivariate normal imputation (Bernaards et al., 2003; Sijtsma & van der Ark, 2003; van Ginkel, Van der Ark, & Sijtsma, 2007).

Several reviews of missing data methods in medical and epidemiological studies have observed that complete-case analyses and single-imputation techniques are the most frequently used (Burton & Altman, 2004; Wood, White, & Thompson, 2004). These reviews have not distinguished between missing data in single-item and multi-item instruments and, for the latter, between missing several or all items. Previous reviews were published at least five years ago, and it might be expected that missing-data methods have improved. We review how recent epidemiological reports have handed missing data, and whether distinctions are made between types of missing data. We also provide an overview of current methods to handle various types of missing data.

Methods

The 2010 volumes of the American Journal of Epidemiology, Epidemiology, and International Journal of Epidemiology (Impact Factor > 5.0 (Reuters, 2009)) were searched by one researcher (IE) (846 articles). We selected articles in which studies used questionnaires to assess the predictors, covariates, or outcomes. 285 studies fulfilled the inclusion criteria (online Appendix, http://links.lww.com). In 4 studies the authors explicitly reported that no data were missing and 19 studies contained no

information on missing data, leaving 262 studies for analysis (Figure 2.1).

Information was extracted using an inventory list containing 28 items (Appendix 2.1). The list was based on the guidelines provided by Sterne et al. (2009) and the STROBE statement (Von Elm et al., 2008). The list assessed information on number and type of missing data, and the methods used to handle missing data. Items could be answered by "yes", "no", "unclear", "not applicable" or "no information."

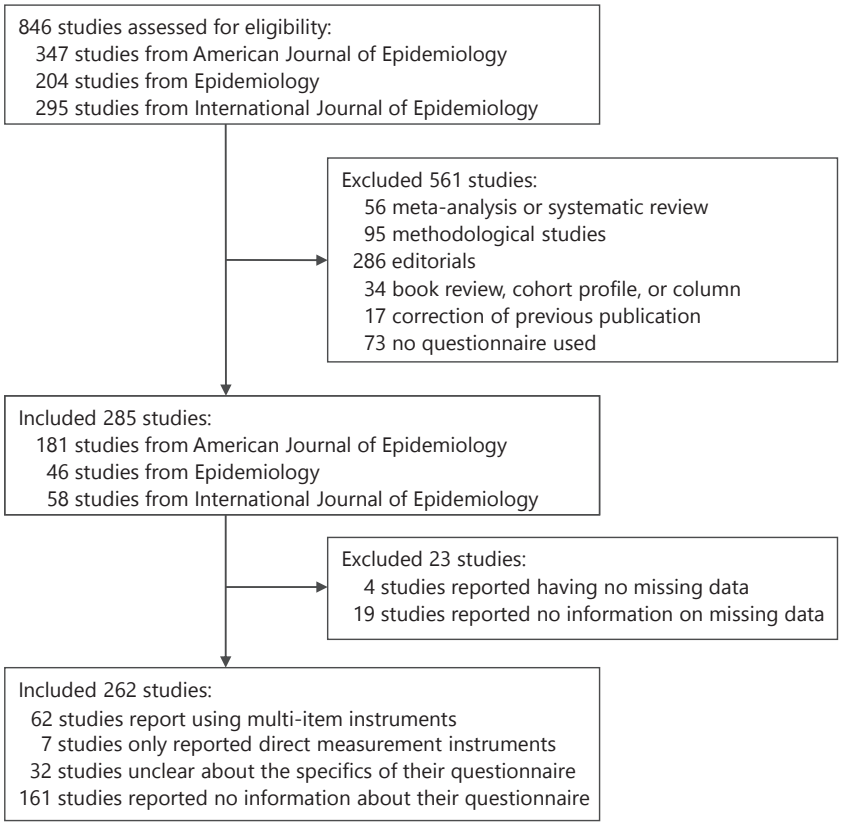


Figure 2.1. Study selection process of studies using questionnaires in all publications in 2010 in the American Journal of Epidemiology, Epidemiology, and the International Journal of Epidemiology

One reviewer (IE) evaluated all studies. 50 studies were randomly selected for independent assessment by another rater (MWH). Agreement was scored for the selection of studies and the items on the inventory list. Overall agreement was 75%, with discrepancies settled by consensus.

Results

In 262 studies having missing data, the type of missing data could not be clearly defined in 46% (Table 2.1). Missing data were most frequently reported for total scores (76%). Among the types of missing data, the average percent missing data was highest for missing cases, although with a wide range for all types. Results were similar for the subgroup of 62 studies that used a multi-item instrument, except that the percent that reported missing item scores was greater (19%) by definition; only multi-item instruments can have missing item scores.

Table 2.1.
Number and percentage of various types of missing data

Type of missing data	All studies (n=262)[a] %	Average percent missing (range)
Item scores	5	11(1 – 44)
Total scores	76	15 (<1 – 68)
Cases	32	26 (<1 – 82)
Unclear	46	13 (<1 – 62)

Note: [a] 105 studies reported multiple types of missing data.

22% of studies considered the possible selectivity of missing data (Table 2.2), presumably by examining differences in characteristics of responders and non-responders. Only 14% of studies that considered selection were specific about the assumed mechanism of missingness (e.g., missing at random).

Table 2.2.
Exploration of selectivity of missing data and reported mechanism in 262 studies that reported missing data

Methods Applied or Reported	% of Studies
Selectivity of missing data examined (n=262)	22
Method used to compare responders to non-responders (n=58)	
Differences in characteristics described in text	54
Descriptive statistics presented	22
Statistical test performed[a]	14
Unclear	10
Reported mechanism (n=58)	
MCAR	0
MAR	14
MNAR	0
Unclear	2
No Information[b]	85

Note: [a] T-test, Chi Square test or Logistic regression; [b] No information means that the assumed mechanism was not explicitly reported.

The methods used to handle missing data are presented in Table 2.3. Most studies (81%) performed a complete-case analysis. 14% used a single imputation technique (e.g., mean imputation, single regression imputation, last observation carried forward, etc.). Multiple imputation, full information maximum likelihood estimation and inverse probability weighting, which assume that data are missing at random, were reported in 8%, 2%, and 3% of the studies, respectively. Results were similar in studies with a multi-item instrument. Just 11% of all studies performed sensitivity analysis to investigate the influence of the handling of missing data on the study results.

Table 2.3.
Reported methods to handle missing data

	All studies (n=262)[a]	Multi-item instrument studies (n=62)[b]
	%	%
Complete-case analysis	81	79
Single imputation techniques	14	19
Including missing indicator as extra answer category	6	7
Inverse probability weighting	3	2
Multiple imputation	8	5
Full information maximum likelihood estimation	2	2
Unclear	13	13
No Information	2	3

Note: [a] 67 studies reported multiple methods to handle their missing data. [b] 16 studies reported multiple methods to handle their missing data.

Discussion

In a review of recent papers published in the three leading epidemiological journals, the routine approach to missing data was to include analysis of complete cases only. For many studies it was unclear whether missing data were from a single-item or multi-item instrument. In studies that use multi-item instruments, researchers generally did not pay attention to the different types of missing data and their corresponding handling methods. Methods designed to handle missing item scores in multi-item instruments, have the advantage that the total score of the construct can be estimated from other items within the same scale. When scores of single-item instruments or total scores are missing, information from other scales and variables is needed — which are usually less efficient. A review by van Ginkel, et al. (2010) found similar lack of distinction between methods to handle missing item scores in multi-item instruments and missing total scores. A broader appreciation of this difference might lead to the application of more valid missing data methods for

multi-item instruments.

More generally, complete-case analysis and single imputation techniques can bias study results, depending on the underlying mechanism of missingness (Huisman, 1999; Little & Rubin, 2002). Knowledge about the selectivity of missing data and the corresponding mechanism forms an important starting point to effectively handle missing data. A proper approach to missing data depends on whether the data are missing at random (MAR) or missing not at random (MNAR). Where data are missing at random (or can be assumed to be MAR), observed data can be used to estimate the missing values. When data are MNAR, such estimation is not possible. A third mechanism assumes that subjects with missing data are a random subset of the whole study sample and therefore even less prone to bias (MCAR) (Rubin, 1976). In the reviewed studies the average mean proportion of missing data was larger than 10%, which might lead to potential biased results. Even when the MCAR assumption holds, loss of power may cause unreliable estimates (Enders, 2010). Furthermore, for both MAR and MNAR data, complete-case analysis and single imputation methods can result in incorrect parameter estimates (Donders et al., 2006; Enders, 2010; Enders & Bandalos, 2001; Graham, 2009; Huisman, 2000; Roth, 1994). Recommended methods such as multiple imputation, full information maximum likelihood estimation, and inverse probability weighting were used in only 13% of studies. These techniques have been shown to work well when MCAR and MAR assumptions hold (Baraldi & Enders, 2010; van Buuren, 2010). In general, studies should report on how many cases their inferences are based. Where a substantial proportion of data are missing, (selective) missings can lead to biased results.

We used publications of three leading epidemiological journals to represent current epidemiological research. We expect that the practices described in these papers are at least as good as actual practice for the field as a whole. The majority of the studies were rated by only one rater, which is presumably less valid than if several raters had sought consensus. Also, our results are based only on the reported information in the studies. This might underestimate the actual extent of missing data. The lack of clarity in many studies over whether a multi-item instrument had been used makes it impossible to assess whether optimal methods for dealing with missing data in multi-item questions were applied.

The reporting of missing data in epidemiological studies is highly variable and mostly poor. Most epidemiologists do not distinguish between missing item scores and missing total scores in multi-item instruments, either in reporting their missing data or in the application of missing data methods. Many researchers may not be aware of the impact of the different types of missing data (i.e., item or total scores) on their study results.

References

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. Journal of School Psycholology, 48(1), 5-37.

Bernaards, C. A., Farmer, M. M., Qi, K., Dulai, G. S., Ganz, P. A., & Kahn, K. L. (2003). Comparison of Two Multiple Imputation Procedures in a Cancer Screening Survey. Journal of Data Science, 1(3), 19.

Burton, A., & Altman, D. G. (2004). Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. British Journal of Cancer, 91(1), 4-8.

van Buuren, S. (2010). Item Imputation Without Specifying Scale Structure. Methodology, 6(1), 31-36.

DeSouza, C. M., Legedza, A. T., & Sankoh, A. J. (2009). An overview of practical approaches for handling missing data in clinical trials. Journal of Biopharmaceutical Statistics, 19(6), 1055-1073.

Donders, A. R., van der Heijden, G. J., Stijnen, T., & Moons, K. G. (2006). Review: a gentle introduction to imputation of missing values. Journal of Clinical Epidemiology, 59(10), 1087-1091.

Enders, C. K. (2010). Applied Missing Data Analysis. New York, NY: The Guilford Press.

Enders, C. K., & Bandalos, D. L. (2001). The Relative Performance of Full Information Maximum Likelihood Estimation for Missing Data in Structural Equation Models. Structural Equation Modeling: A Multidisciplinary Journal, 8(3), 430-457.

van Ginkel, J. R., Sijtsma, K., van der Ark, L. A., & Vermunt, J. K. (2010). Incidence of missing item scores in personality measurement, and simple item-score imputation. Methodology, 6(1), 17-30.

van Ginkel, J. R., Van der Ark, L. A., & Sijtsma, K. (2007). Multiple Imputation of Item Scores in Test and Questionnaire Data, and Influence on Psychometric Results. Multivariate Behavioral Research, 42(2), 387-414.

Graham, J. W. (2009). Missing data analysis: making it work in the real world. Annual Review of Psychology, 60, 549-576.

Greenland, S., & Finkle, W. D. (1995). A critical look at methods for handling missing covariates in epidemiological regression analyses. American Journal of Epidemiology, 142(12), 1255-1264.

Huisman, M. (1999). Item Nonresponse: Occurence, Causes, and Imputation of Missing Answers to Test Items. Leiden: DSWO Press.

Huisman, M. (2000). Imputation of Missing Item Responses: Some Simple Techniques. Quality & Quantity, 34(4), 331-351.

Little, R. J., & Rubin, D. B. (2002). Statistical Analysis with Missing Data (Second Edition ed.). Hoboken, NJ: John Wiley & Sons.

McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). Missing Data A Gentle Introduction. New York: The Guilford Press.

Radloff, L. S. (1977). The CES-D Scale. Applied Psychological Measurement, 1(3), 385-401.

Reuters, T. (2009). Journal Citation Reports (Science Edition ed.). New York: ISI Web of Knowledge.

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. Personnel Psychology, 47(3), 537-560.

Rubin, D. B. (1976). Inference and missing data. Biometrika, 63(3), 581-592.

Sijtsma, K., & van der Ark, L. A. (2003). Investigation and treatment of missing scores in test and questionnaire data. Multivariate Behavioral Research, 38(4), 505-528.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., . . . Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ, 338, b2393.

de Vet, H. C. W., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011). Measurement in Medicine. Cambridge: Cambridge University Press.

Von Elm, E., Altman, D. G., Egger, M., Pocock, S., Gotzsche, P. C., Vandenbroucke, J. P., & Initiative, S. (2008). The strengthening the reporting of observational studies in epidemiology (STROBE) statement. Journal of Clinical Epidemiology, 61, 344-349.

Wood, A. M., White, I. R., & Thompson, S. G. (2004). Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. Clinical Trials, 1(4), 368-376.

Appendix 2.1|Inventory list

First Author:.....
Pages: .....

Study design
1. Type of design:
☐ Treatment studies
☐ RCT
☐ Non-randomized trial (quasi-experiment)
☐ Observational studies
☐ Cohort study
☐ Prospective cohort
☐ Retrospective cohort
☐ Case-control study
☐ Cross-sectional study

2. Researched population
☐ Patients
☐ Healthy individuals

3. If patients: type of Patients included:.....

4. Number of participants:.....

5. Principal analysis used in the study:
☐ Linear/logistic/Cox/Poisson regression
☐ Mixed models
☐ Generalized Estimating Equations (GEE)
☐ T-tests
☐ ANOVA / MANOVA / Repeated measures ANOVA
☐ Item response theory (IRT)
☐ Chi-square test
☐ Nonparametric test
☐ Other:.....

6a. A Questionnaire was used for the assessment of:
☐ Covariates/Predictor
☐ Outcome
☐ Both
☐ None

6b. Did the questionnaire consist of different items resulting in a total score (or total scores per dimension/subscale)?
☐ Yes
☐ No
☐ Unclear
☐ No information
☐ Not applicable

6c. Is more information available on the questionnaire (e.g., response/answer categories, score calculation, etc.)?
☐ Yes
☐ No

Missing data information:
7a. Is the percentage/number of missing data described?
☐ Yes
☐ No
☐ Unclear

7b. Is the location of missing data described?
☐ Yes
☐ No
☐ Unclear
☐ Not applicable

7c. What is the location of missing data presented?
☐ Missing in variables
☐ Missing total scores (also attrition)
☐ Missing items
☐ Missing cases (unit nonresponse: did not show up/return questionnaire)
☐ Planned missingness (e.g.. missing by design)
☐ Other:.....
☐ Unclear
☐ Not applicable

7d. What type of missing data is reported?
☐ Nonresponse (item/variable/unit)
☐ Dropout
☐ Attrition
☐ Lost to follow up
☐ Intermittent
☐ Other:.....
☐ Unclear
☐ Not applicable
☐ No Information

7e. What is the percentage of missings in the data reported?
Total score:.....
Item: .....
Cases:.....

8. Is the fraction of missing information presented?
☐ Yes
☐ No
☐ Unclear

9. Are the potential reasons for missing data discussed (e.g.. exhaustion, deceased, lack of motivation, lost in mail, etc.)?
☐ Yes
☐ No
☐ Unclear

10a. Was the missing data mechanism evaluated?
☐ Yes
☐ No
☐ Unclear

10b. Which method was used to test the mechanism?
☐ Differences in characteristics between missing and non-missing group described
☐ Analysis between cases with complete and missing data:
☐ Descriptive statistics (e.g., comparing means / percentages)
☐ Chi-square tests
☐ T-tests
☐ Univariate t-Test comparisons
☐ Little's MCAR test
☐ Logistic regression analysis with missing data as outcome
☐ Other:.....
☐ Unclear
☐ Not applicable

10c. What category of missing data mechanism is reported?
☐ MCAR
☐ MAR
☐ MNAR
☐ Other:.....
☐ Unclear
☐ No information
☐ Not applicable

Methods used to handle the missing data:
11. Handling method
☐ Missing total score/unit score methods
☐ Complete-case analysis
☐ Pairwise deletion
☐ Mean substitution/arithmetic mean imputation/unconditional mean imputation/median imputation
☐ Single regression imputation (e.g.. Stochastic)
☐ Hot-deck imputation – matching nonrespondents to resembling respondent
☐ Last value carried forward
☐ Multiple imputation
☐ Full Information Maximum Likelihood Estimation
☐ Missing item score methods
☐ Unconditional random imputation
☐ Item mean substitution/person mean substitution
☐ Corrected item mean substitution
☐ Two-way imputation
☐ Response-function imputation
☐ Multivariate normal imputation
☐ Fully conditional specification
☐ Similar Response pattern imputation
☐ Item correlation substitution
☐ Multiple response-function imputation
☐ Including a missing category
☐ Unclear
☐ No information

☐ Other.....
☐ Not Applicable

If multiple imputation is used (only 12a, 12b and 12c):
12a. Are the number of variables used in the imputation model clearly described?
☐ No
☐ No but normality discussed
☐ Yes
☐ Not applicable

12b. Is the number of multiple imputations presented?
☐ Yes
☐ No
☐ Unclear
☐ Not applicable

12c. Was the imputation process evaluated (i.e., convergence studied, imputed values compared with observed values, etc)?
☐ Yes
☐ No
☐ Unclear
☐ Not applicable

13. Is it described how non-normal / categorical variables were dealt with in the missing data method?
☐ Yes
☐ No
☐ Unclear
☐ Not applicable

14a.Was a sensitivity analysis performed to investigate the influence of how the missing data were handled on the study results?
☐ Yes
☐ No
☐ Unclear
☐ Not applicable

14b. What kind of sensitivity analysis was performed?
☐ Complete case analysis versus imputation method
☐ Different imputation techniques were compared
☐ Other:.....
☐ Not applicable

14c. Are the results of the sensitivity analysis clearly described?
☐ Yes
☐ No
☐ Unclear
☐ Not applicable

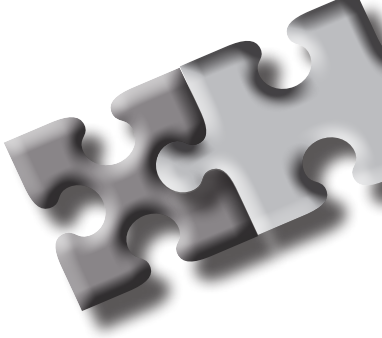
15. What Software package was used for the missing data method?
☐ SPSS
☐ AMOS – Structural equation modeling tool in SPSS
☐ EQS – Structural equation modeling software
☐ HLM – Hierarchical data modeling
☐ LISREL – Structural equation modeling software

- ☐ Mplus – Statistical modeling program
- ☐ SAS
- ☐ SOLAS for missing data analysis
- ☐ Stata
- ☐ EMCOV
- ☐ S-Plus
- ☐ R
- ☐ S-Plus en R packages
- ☐ Amelia - Amelia II: A Program for Missing Data
- ☐ NORM - analysis of multivariate normal datasets with missing values
- ☐ CAT - Analysis of categorical-variable datasets with missing values
- ☐ MICE - Multivariate imputation by chained equations
- ☐ MI - Missing Data Imputation and Model Checking
- ☐ MIX - Multiple imputation for Mixed Categorical and Continuous Data
- ☐ missMDA - Handling missing values with/in multivariate data analysis (principal component methods)
- ☐ mitools - Tools for multiple imputation of missing data
- ☐ mlmmm - ML estimation under multivariate linear mixed models with missing values
- ☐ mvnmle - ML estimation for multivariate normal data with missing values
- ☐ PAN - Multiple imputation for multivariate panel or clustered data
- ☐ MIXED
- ☐ No information
- ☐ Unclear
- ☐ Not applicable

16. What software was used for the primary/general analyses?

SPSS Version .....
SAS Version .....
Stata Version .....
Statistica Version .....
R Version.....
Mplus Version.....
S-plus Version.....
EQS Version .....
LISREL Version .....
SUDAAN Version .....
Other:.....

- ☐ No Information



Chapter 3

Missing data in a multi-item questionnaire are best handled by multiple imputation at the item score level

Published: Eekhout, I., de Vet, H.C.W., Twisk, J.W.R., Brand, J.P.L., de Boer, M.R., & Heymans, M.W. (2014). Missing data in a multi-item instrument were best handled by multiple imputation at the item score level. Journal of Clinical Epidemiology, 67(3), 335-342.

Abstract

Regardless of the proportion of missing values, complete-case analysis is most frequently applied, although advanced techniques such as multiple imputation are available. The objective of this study is to explore the performance of simple and more advanced methods for handling missing data in case some, many, or all item scores are missing in a multi-item instrument. Real-life missing data situations were simulated in a multi-item variable used as a covariate in a linear regression model. Various missing data mechanisms were simulated with an increasing percentage of missing data. Subsequently, several techniques to handle missing data were applied to decide on the most optimal technique for each scenario. Fitted regression coefficients were compared using the bias and coverage as performance parameters. Mean imputation caused biased estimates in every missing data scenario when data are missing for more than 10% of the subjects. Furthermore, when a large percentage of subjects had missing item scores (>25%), multiple imputation methods applied to the items outperformed methods applied to the total score. We recommend applying multiple imputation to the item scores in order to get the most accurate regression model estimates. Moreover, we advise not to use any form of mean imputation to handle missing data.

Keywords: missing data, multiple imputation, multi-item questionnaire, item imputation, methods, bias, simulation

Introduction

Missing data on multi-item instruments is a frequently seen problem in epidemiological and medical studies. Multi-item instruments can be used to measure for example quality of life, coping ability or other psychological states. A multi-item instrument generally consists of several items that measure one construct (de Vet, Terwee, Mokkink, & Knol, 2011), for example the Pain Coping Inventory assesses active coping skills for people with pain complaints by 12 items (Kraaimaat & Evers, 2003). Missing data on these kinds of instruments can occur as missing item scores, when several items are not completed, or as missing data in total scores when the entire instrument is not filled out. Furthermore, missing item scores impair the calculation of the total score which can lead to missing total scores as well. For missing data in item and total scores, different missing data handling methods are available, with complete-case analysis (CCA) as the most frequently used method (Eekhout, de Boer, Twisk, de Vet, & Heymans, 2012). In general, CCA tends to perform well under the strict assumption that missing data are a completely random subsample of the data, in other words missing completely at random (MCAR) (Rubin, 1976). However, CCA reduces power caused by a decreased sample size. Single imputation methods, such as mean imputation of the total score and item mean imputation, may be used to preserve the sample size by replacing the missing values by the mean score, but these methods reduce the variability in the data. Single stochastic regression imputation (SRI) uses observed data to predict the missing value and adds residual error to the imputed data to restore the variability in the data, but this method does not take the uncertainty of the imputed values into account.

Mostly, the probability of missing data depends on other observed variables, indicated as missing at random (MAR) (Rubin, 1976). In contrast to traditional methods as CCA and mean imputation, more advanced methods, such as multiple imputation, produce reliable and unbiased results under the MAR mechanism and take missing data uncertainty into account (Janssen et al., 2010; Little & Rubin, 2002). Both traditional and advanced methods can be applied either to the missing item scores, or directly to the missing total scores.

The comparison between missing data methods for item level and total score level missingness in questionnaire data is seldom made in one study (Eekhout, et al., 2012). Other simulation studies have researched the performance of missing data methods applied to non-questionnaire data (Collins, Schafer, & Kam, 2001; Marshall, Altman, Royston, & Holder, 2010) or only studied methods applied to the item scores of a multi-item instrument (Burns et al., 2011; Hawthorne & Elliott, 2005; Roth, Switzer, & Switzer, 1999; van Buuren, 2010; van Ginkel, Sijtsma, van der Ark, & Vermunt, 2010). For example Burns et al. (2011) studied the performance of multiple imputation of missing item scores, but did not compare this to imputing at the total

score level of their questionnaire. So far it is still unclear if it is better to apply a missing data handling method to the missing item scores or to the total scores when some or many items in a multi-item instrument are missing. Moreover, the impact on the study results of different missing data methods when multi-item data are missing on the covariate has not been researched extensively yet. The current study aims to explore the performance of different missing data handling methods designed for missing item scores and missing total scores in a multivariable regression model. This objective is considered in the following two aspects: (1) which missing data methods should be used to handle missing data; and (2) should this missing data method be applied to the item scores or to the total scores.

Methods

Simulation set up

In order to investigate the differences between several imputation methods, we used a simulation procedure comparable to the study performed by Marshall et al. (2010). We based our simulation on an empirical dataset, which was previously used in a prospective cohort study investigating the prognosis of low back pain (Heymans et al., 2006). In this study we used a cross-sectional part of these data that contained the multi-item variable active coping of the Pain Coping Inventory (PCI-active) (Kraaimaat & Evers, 2003). The PCI consists of 12 items with four ordered response categories that result in a total sum score, which we consider as a continuous scale. Additionally, five other covariates were selected to be included in this dataset: gender, health status, job demands, number of working years, absenteeism, and the outcome variable was low back pain intensity. Using the means and covariance matrix of these empirical data, 500 simulated data samples of 500 subjects were generated using the mvrnorm function in package MASS in R statistical software (Venables & Ripley, 2002). Subsequently, in each simulated sample, missing data were created in only the multi-item covariate PCI-active under several missing data mechanisms. After this step, several techniques were applied to handle the incomplete datasets. The implications of these different techniques were compared by fitting a multivariable regression model to the data. This model regressed PCI-active total score, gender, health status and job demands on pain intensity. In order to have an imputation model that differs from the regression model, number of working years and absenteeism were only included in the imputation models, but not in the final regression model. Model coefficients fitted to the 'handled data' were compared to the 'true' model coefficients fitted to the same samples without missings.

Step 1 Generating missing data

The generation of missing data in the multi-item covariate PCI-active was performed using a program that was translated from SAS software (Brand, van Buuren, Groothuis-Oudshoorn, & Gelsema, 2003) into R statistical software (R Core Development Team, 2012) by the first (IE) and last author (MWH). This program was used to generate multivariate missing data, according to the MCAR, MAR and MNAR mechanisms. In the MCAR situations the selection of item scores that were made missing in the PCI-active covariate was completely random. In the MAR situations item scores of the PCI-active covariate were made missing depending on the values of the observed items and the other covariates gender, health status, job demands number of working years, absenteeism and the outcome low back pain intensity. For the MNAR situations, the scores that were made missing also depended on the values of the PCI-active item scores themselves.

We generated missing item data in the PCI-active covariate according to the following four patterns: (1) a pattern where 25% of the item scores were made missing within subjects, (2) a pattern where 50% of the item scores were made missing, (3) a pattern where 75% of the item scores were made missing, (4) or a pattern where 100% of the items were considered missing. These missing item patterns were applied to 10%, 25%, 50%, or 75% of the subjects. This resulted for example in a situation where in 10% of the subjects, 25% of the item scores were missing.

Total scores of the PCI-active covariate can only be calculated if all item scores are available. Consequently, in the first three patterns mentioned above, where some of the item scores in the PCI-active covariate were made missing within a subject, the PCI-active total score for that subject was also missing. This situation was reflected by the fourth pattern. This made it possible to study separately, in each simulated sample, the influence of missing data methods when they were applied to missing values in item scores or total scores. By generating the incomplete data according to the above described scenarios, 48 different situations were investigated.

Step 2 Methods to handle missing data

The generated incomplete datasets were handled using different methods, summarized in Table 3.1. As previously mentioned, these methods can either be applied to the missing item scores, after which the total score can be calculated or to the missing total score directly. Both of these possibilities were explored in this study. After applying the method, a multivariable regression model was fitted and regression coefficients were estimated.

1. Complete-case analysis.

In a CCA, only the subjects with complete observations for the PCI-active covariate were included in the analysis. Accordingly, all subjects with missing item scores were removed from the data and the model was fitted to the remaining sample. Consequently this method would yield the same results when applied to the missing item data as when applied to the missing total scores data directly.

Table 3.1.
Summary of the application of different missing data methods to item scores and/or total scores.

Label	Method name	Applied to item scores	Applied to total scores
1 CCA	Complete-case analysis	X[a]	X
2a Mean	Mean imputation		X
2b Item mean	Item mean imputation	X	
2c Person mean	Person mean imputation	X	
2d Two-way	Two-way imputation	X	
3 SRI	Single stochastic regression imputation	X	X
4a MI-SR	Multiple Imputation by stochastic regression	X	X
4b MI-PMM	Multiple Imputation by predictive mean matching	X	X
4c MI-PO	Multiple Imputation by proportional odds model	X	

Note: [a] Equal to application to total scores.

2. Mean imputation.

In mean imputation the missing scores were imputed with the mean score of the non-missing data. The missing data on the PCI-active covariate total scores were imputed with the mean total score of all observed subjects in mean imputation applied to the total scores (2a). In item mean imputation (2b) a missing item score was imputed with the mean score for all complete data on that item. In person mean imputation (2c), the mean score of the items per subject was calculated, and for each subject missing item scores were imputed with this 'personal mean score'. Two-way imputation (2d) combines the person mean and the item mean to account for both the personal effect and the item effect. The person mean was added to the item mean, and then the overall mean was subtracted. Furthermore, a random error term was added to account for variability in the data (Bernaards & Sijtsma, 2000).

3. Stochastic regression imputation.

In SRI the missing values were imputed with the regression estimates from the observed variables when applied to the total score. For methods applied to the items, the regression estimate of the observed variables and the observed items was used. Regression assumes that the imputed values fall directly on the regression line, so it implies a correlation of 1 between the predictors and the incomplete outcome variable (PCI-active). Stochastic regression imputation aims to reduce the bias by an

extra step of augmenting each predicted score with a normally distributed random error with a variance equal to the variance of the regression model.

4. Multiple imputation.

The multiple imputation (MI) method has three phases, the imputation phase, the analysis phase and the pooling phase. First, the incomplete data was completed by imputing a value for the missing scores in the imputation phase. When applied to the item scores, the missing item scores were imputed, but when applied to the missing total scores the PCI-active total score was imputed. The imputed values were estimated from the observed variables in the dataset by an imputation algorithm and a random residual term was added to each resulting estimate. The imputation algorithm is a regression equation specified in the imputation model using the observed variables to estimate the missing value. In case of the item score application, the observed item scores of the PCI-active covariate were also included in the imputation model. Multiple datasets were generated, each with different imputed values for the missing items or total scores. In this simulation study we generated 15 imputed datasets, which is higher than the minimum recommended number of 5 (van Buuren, 2012), and still computationally and practically possible in this simulation study. During the analysis phase, the analysis was carried out on each dataset using the same procedures that would have been used had the data been complete. So, when applied to the item scores, the item scores were added to form a total score and the regression analysis was performed. In case of the missing total scores, the imputed total scores were used in the analysis. Finally, in the pooling phase the multiple sets of results, or parameter estimates, were combined into one single set of results according to Rubin's rules (Rubin, 2004). In this study three different imputation algorithms were applied and compared to impute the missing data in the imputation phase. These were stochastic regression (SR; 4a), predictive mean matching (PMM; 4b), and a proportional odds model (PO; 4c) (van Buuren, 2007, 2010). The latter model is recommended for missing ordinal categorical data (van Buuren, 2012). The multiple imputations were done using the mice function in package MICE (van Buuren & Groothuis-Oudshoorn, 2011).

Step 3 Comparing missing data handling methods

The model coefficients of the 'handled datasets' were compared to the coefficients based on the 'true data'. The true data were generated by running the simulation process without missing data. Accordingly, the regression model was fitted to 500 simulated complete samples and the average regression coefficients formed the true values against which the missing data simulations were compared. Regression coefficients were considered 'biased' when the estimate was outside a limit of 0.5

standard error from the true coefficient (Schafer & Graham, 2002). We also looked at the estimates for the standard error (SE) of the regression estimates, which were required to be estimated somewhat larger than the true SE in order to incorporate the appropriate missing data uncertainty (Enders, 2010). Furthermore, the coverage of the true value of the regression coefficients within the confidence limits of the estimated coefficients was computed. This was calculated as the percentage of times that the true regression coefficient was within the confidence interval of the estimates from the datasets after the missing data handling methods were applied. Coverage of 95% is optimal whereas higher coverage indicates that the method might be too conservative and lower coverage value suggests a higher than expected type I error (Burton, Altman, Royston, & Holder, 2006). Decreased coverage can be caused by too narrow confidence intervals as a result from underestimated standard errors. Standard errors should incorporate uncertainty of missing data to overcome this problem (Siddique, Harel, & Crespi, 2012). All analyses and simulations were performed in R statistical software (R Core Development Team, 2014), scripts are available by the first author upon request.

Results

In Table 3.2, the regression coefficient and standard error estimates for the PCI-active total score under the three missing data mechanisms are presented. Not surprisingly, for the MCAR data the coefficient estimate was the same as the true coefficient value, but the standard error increased with higher data missing rates. A similar trend was seen in the MAR and MNAR missing data situations, however accompanied by much larger deviations in SE's.

Table 3.2.
Coefficient and standard error (SE) of the PCI-active total score covariate according to the three missing data mechanisms MCAR, MAR and MNAR after a complete-case analysis.

	'True' Coefficient (SE)	10%[a] Coefficient (SE)	25%[a] Coefficient (SE)	50%[a] Coefficient (SE)	75%[a] Coefficient (SE)
MCAR	0.1411 (0.0135)	0.1413 (0.0142)	0.1413 (0.0156)	0.1410 (0.0191)	0.1424 (0.0274)
MAR	0.1411 (0.0135)	0.1414 (0.0142)	0.1413 (0.0157)	0.1418 (0.0198)	0.1428 (0.0259)
MNAR	0.1411 (0.0135)	0.1412 (0.0143)	0.1416 (0.0159)	0.1498 (0.0207)	0.1729 (0.0285)

Note: [a] Percentage of cases that had a missing PCI-active total score.

Figures 3.1 and 3.2 present the effect of the missing data handling methods on the estimates of regression coefficients for an increasing amount of missing total scores and an increasing amount of missing item scores when 25% of the subjects

have missing data, respectively. A full tabulation of the results is presented in the online Appendices (www.jclinepi.com).

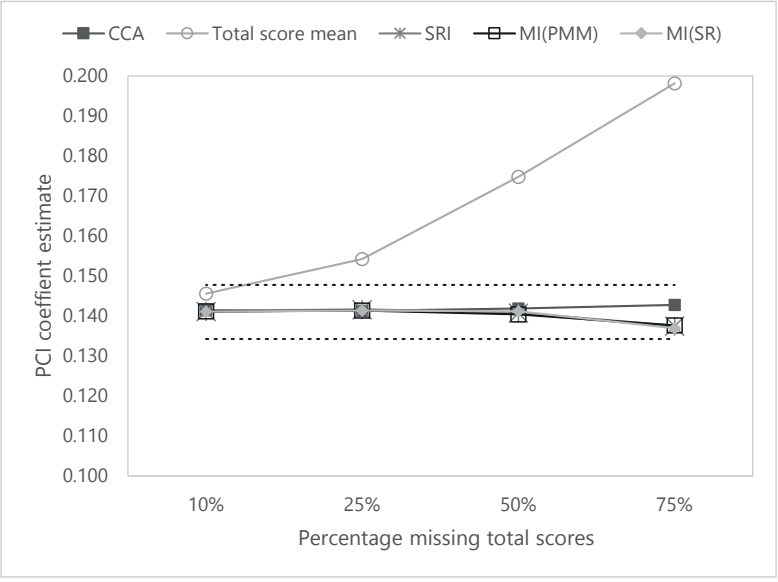


Figure 3.1. Regression coefficient estimates of the PCI-active covariate for different missing data methods applied to the total score for when an increasing percentage of subjects had missing at random data in total score. The black dashed lines depict the thresholds for bias at 0.5 SE from the true coefficient.

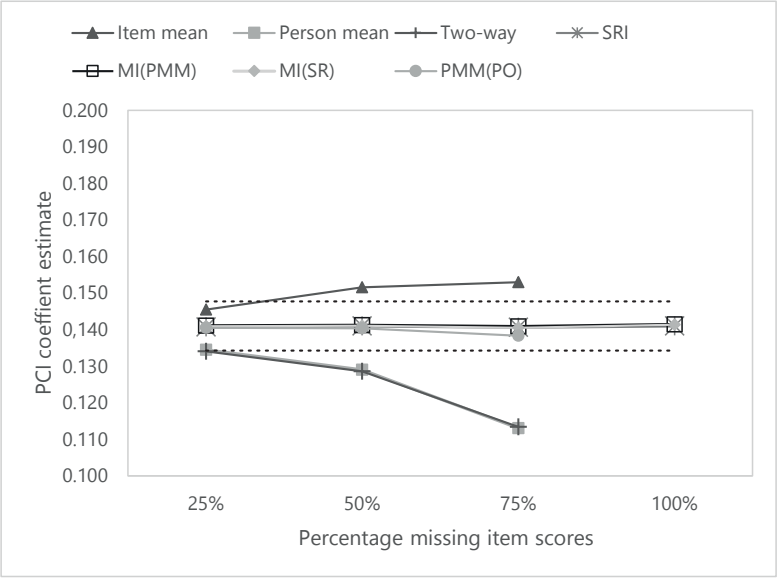


Figure 3.2. Regression coefficient estimates of the PCI-active covariate for different missing data methods applied to the items when 25% of the subjects had missing at random data in an increasing percentage of missing item scores. The black dashed lines depict the thresholds for bias at 0.5 SE from the true coefficient.

Regression coefficients

In the MAR data situation multiple imputation as well as SRI and CCA gave unbiased regression coefficient estimates of the PCI-active variable, regardless of the number of missing total scores and items (Figure 3.1 and 3.2). The worst method applied to the total scores was mean imputation and the worst methods applied to item scores were item and person mean imputation and two-way imputation which yielded biased coefficients. For example, item mean imputation had values in the range of 0.150 and 0.172 compared to the true value of 0.141 for the PCI-active variable, when 50% of the subjects had missing item scores (online Appendix, www.jclinepi.com). Mean imputation of the total score yielded biased regression estimates when more than 10% of the cases had a missing total score. Results were the same in MCAR and MNAR data; mean imputation methods were the worst solutions as well. For example item mean imputation had estimates ranging from 0.152 to 0.185 compared to the true value of 0.141 when 50% of the subjects had MNAR data. With 50% or more MNAR data in item or total scores, also advanced MI methods gave biased coefficient estimates.

Standard errors and coverage

In MAR data the methods that yielded the largest bias in the regression coefficient estimates as reported above, also showed biased SE's (online Appendix, www.jclinepi.com). This was most evident for the mean imputation methods in all situations of missing item and total scores. Figure 3.3 and 3.4 display the SE estimates in MAR data for the methods that gave the best results with respect to the regression coefficients.

Despite the unbiased coefficient estimates for PCI-active in a CCA, the standard error was highly overestimated for this method in MAR data (Figure 3.3). Overall multiple imputation showed slightly smaller bias and better SE estimates when applied to the items than when this method was applied to the total scores when less than 50% of the item scores were missing. Similar SE results for the different methods were seen in MCAR and MNAR data (online Appendix, www.jclinepi.com). Standard error estimates in MI methods correctly incorporated missing data uncertainty both when applied to item scores and to total scores. Coverage rates, which measure the combined performance of the coefficient estimate and SE by evaluating the confidence interval, confirm this. Coverage rates were worst for single imputation methods and best for multiple imputation of the item scores (Table 3.3).

Other covariates

Other covariates in the model were also affected by the missing data in the PCI-active variable (online Appendix, www.jclinepi.com). In general, mean imputation methods resulted in biased estimates on the other covariates when an increasing

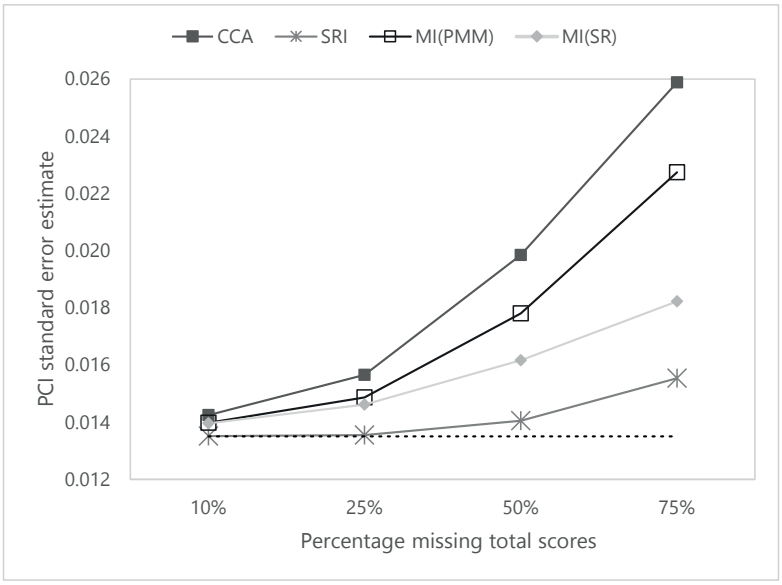


Figure 3.3. Standard error estimates for the missing data methods that have unbiased PCI-active coefficient estimates when an increasing percentage of subjects had missing at random data in total scores. The black dashed line depicts the SE estimate of the true coefficient (0.0135).

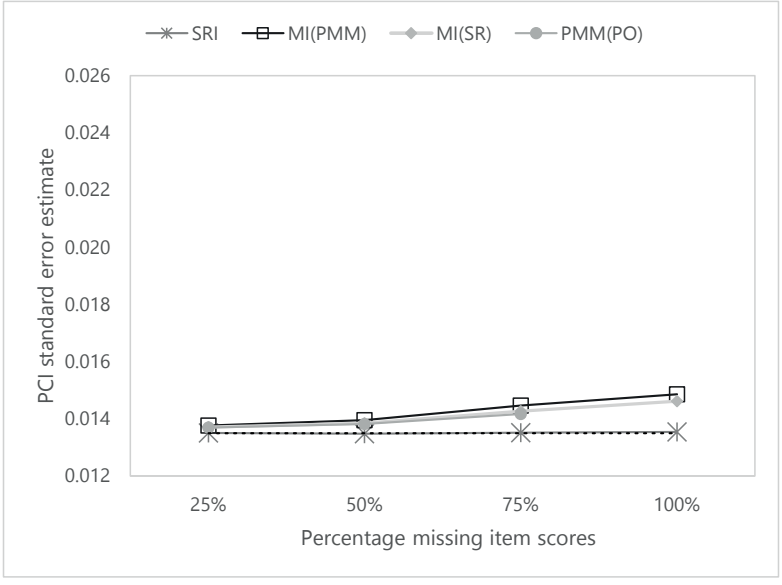


Figure 3.4. Standard error estimates for the missing data methods that have unbiased PCI-active coefficient estimates when 25% of the subjects had missing at random data in an increasing percentage of missing item scores. The black dashed line depicts the SE estimate of the true coefficient (0.0135).

amount of data were missing on the PCI-active covariate. MI methods applied to the item scores mostly resulted in unbiased coefficient estimates for all other covariates in all missing data situations. The SE estimates in the other covariates only showed large deviations for the CCA. The coverage was mostly acceptable, round 95%, except for the mean imputation methods, especially when more items were missing in a large amount of data (>50%).

Table 3.3.
Coverage rates for all missing data methods when 25% of the subjects had missing at random data in an increasing percentage of missing item scores.

	Percentage missing item scores			
Method	25%	50%	75%	100%
Item mean	94%	87%	85%	-
Person mean	94%	83%	41%	-
Two-way	92%	83%	40%	-
Total score mean	-	-	-	85%
SRI	94%	95%	93%	91%
MI(SR)	96%	96%	95%	95%
MI(PMM)	95%	96%	95%	95%
MI(PO)	96%	95%	93%	-

Note: All empty cells are not-applicable conditions.

Discussion

The results of our study are that missing item data are best handled by applying MI based on PMM or SR to the item scores regardless of how many subject scores and item scores are missing. Furthermore, SRI also seems to yield acceptable results, and mean imputation of the total scores performs worst. Additionally, we showed that the underlying mechanism influences the performance of the missing data handling method, especially when large amounts of data are missing. This is not only of concern when working with total scores but also when working with missing item scores in multi-item instruments.

Moreover, our results showed that complete-case analysis performed satisfactory with respect to the regression coefficient estimates when data are MAR on the covariate, which was also found in the simulation study by Marshall et al. (2010). However, standard errors are largely overestimated and hence power is reduced. For that reason it is not recommended to perform a complete-case analysis, especially when more than 10% of the subjects have missing data. Nevertheless, in about 80% of epidemiological studies a complete-case analysis is still used (Eekhout, et al., 2012).

Item mean imputation, which is advised in user-manuals of widely used multi-item questionnaires as the SF-36 (Ware, Kosinski, & Keller, 1994) and the PCI (Kraaimaat & Evers, 2003), results in highly biased estimates in all of the missing item data patterns when more than 10% of the subjects have missing data. Therefore

we would not recommend item mean imputation in any situation of missing data. Furthermore, person mean and two-way imputation result in underestimated coefficient and SE estimates. Of the single imputation methods, SRI performs best by far. Other studies with missing covariate data also found this method to exceed the performance of other single imputation methods (Enders, 2010; Pastor, 2003). Even though SRI produced valid regression coefficient estimates at first sight, the imputed values are treated as real data and imputation uncertainty is ignored. This leads to narrow confidence intervals resulting in decreased coverage (Enders, 2010). In studies that investigated more complicated missing data situations, SRI proved to underestimate the standard error (Gold & Bentler, 2000; Newman, 2003). Repeating the imputation process multiple times to account for the missing data uncertainty, as is done in multiple imputation, is therefore recommended. Multiple imputation outperforms the ad-hoc methods and produces minimal bias in model estimates. Other studies have found these same results when comparing multiple imputation to ad-hoc methods applied to continuous variables (Donders, van der Heijden, Stijnen, & Moons, 2006; Kneipp & McIntosh, 2001; Marshall, et al., 2010; Schafer & Graham, 2002). We found that this works for missings in total scores as well as missing values in item scores.

When missing data occurred in 50% or more of the cases, all methods applied to the total score yield largely overestimated standard errors. However, when comparable missing data methods are applied to the item scores of this same data, resulting coefficient estimates and standard errors are much more accurate, with multiple imputation methods showing the best results. Therefore, multiple imputation applied to the item scores is preferred over imputation methods applied to the total scores in these situations.

In addition, we found that the estimation of regression coefficients of other covariates was disturbed by the missing data in the PCI-active covariate. Hence missing data on one variable in the model has an effect on the estimates of all other covariates in the model even though these covariates do not contain any missing data. This effect is larger in MAR and MNAR data than in MCAR data. It would be expected that this influence is typically seen in highly correlated data. However, in our simulated data correlations between the variables were low to moderate ranging from 0.10 to 0.45.

In this simulation study, an example dataset was used as a template to simulate missing data. This is a beneficial point because the simulated scenarios reflect real-life research situations and therefore give a realistic view of the magnitude of the effects on the results. Moreover, the simulated MAR data depended on all covariates and on the outcome which reflects a probable missing data situation. However, the PCI-active items in the example data had a Cronbach's alpha of 0.74, which reflects

an acceptable but not excellent internal consistency (George & Mallery, 2009). This might be an explanation for the disappointing results for person mean, two-way imputation and MI with a proportional odds model, because these methods are based on the internal consistency of a multi-item instrument (Bernaards & Sijtsma, 2000). Also, previous simulation studies that reported optimistic results for two-way imputation only investigated small amounts of missing data (≤20%), in which situations we found unbiased estimates as well, or included repeating two-way imputation multiple times (van Ginkel, et al., 2010; van Ginkel, van der Ark, & Sijtsma, 2007a, 2007b). Furthermore, missing data were only generated on one covariate. Results might be different when missing data occurs in the outcome or both in the covariate(s) and the outcome, which can be explored in a future study.

The comparison between missing data methods applied to item scores and methods applied to total scores as investigated in our study has never been made before. Previous studies investigated solely single item imputation methods applied to missing item scores (Hawthorne & Elliott, 2005; Roth, et al., 1999). These studies did not compare these methods to multiple imputation methods, or to methods applied to total scores. Our results are important for all researchers working with multi-item instruments. Manuals of most multi-item questionnaires have been developed prior to the development and exploration of most advanced methods applied in this study. Therefore it is important to follow-up on current literature to make a fair assessment of the missing data solutions available aside from the methods described in the questionnaire guidelines. Moreover, many major statistical packages such as SAS (SAS Institute Inc., 2011), R statistical software (R Core Development Team, 2012), Stata (StataCorp., 2011), and also SPSS (SPSS Inc., 2008), presently offer applications to multiply impute missing data. Nevertheless, a previously conducted review showed that only 8% of epidemiological studies currently use multiple imputation to handle missing data (Eekhout, et al., 2012).

To sum up, as expected the more advanced methods, such as multiple imputation, perform better than the traditional methods. Furthermore, multiple imputation applied to the item scores performs better than methods applied to the total scores and is therefore advised. However, when only a small amount of item scores are missing (<25%) in only a small amount of data (<10%), single imputation such as SRI or CCA might be preferred with MAR data in the covariate over multiple imputation purely for practical reasons (van Ginkel, et al., 2010; van Ginkel, et al., 2007a). We advise not to use mean imputation applied to missings in items or total scores.

References

Bernaards, C. A., & Sijtsma, K. (2000). Influence of Imputation and EM Methods on Factor Analysis when Item Nonresponse in Questionnaire Data is Nonignorable. Multivariate Behavioral Research, 35(3), 321-364.

Brand, J. P. L., van Buuren, S., Groothuis-Oudshoorn, K., & Gelsema, E. S. (2003). A toolkit in SAS for the evaluation of multiple imputation methods. Statistica Neerlandica, 57(1), 36-45.

Burns, R. A., Butterworth, P., Kiely, K. M., Bielak, A. A., Luszcz, M. A., Mitchell, P., . . . Anstey, K. J. (2011). Multiple imputation was an efficient method for harmonizing the Mini-Mental State Examination with missing item-level data. J Clin Epidemiol.

Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. [Research Support, Non-U.S. Gov't]. Stat Med., 25(24), 4279-4292.

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. Stat.Methods Med.Res., 16(3), 219-242.

van Buuren, S. (2010). Item Imputation Without Specifying Scale Structure. Methodology, 6(1), 31-36.

van Buuren, S. (2012). Flexible Imputation of Missing data. New York: Chapman & Hall/CRC.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological Methods, 6(4), 330-351.

Donders, A. R., van der Heijden, G. J., Stijnen, T., & Moons, K. G. (2006). Review: a gentle introduction to imputation of missing values. Journal of Clinical Epidemiology, 59(10), 1087-1091.

Eekhout, I., de Boer, R. M., Twisk, J. W., de Vet, H. C., & Heymans, M. W. (2012). Missing data: a systematic review of how they are reported and handled. Epidemiology, 23(5), 729-732.

Enders, C. K. (2010). Applied Missing Data Analysis. New York, NY: The Guilford Press.

George, D., & Mallery, P. (2009). SPSS for Windows Step by Step: A Simple Guide and Reference, 17.0 Update: Allyn & Bacon.

van Ginkel, J. R., Sijtsma, K., van der Ark, L. A., & Vermunt, J. K. (2010). Incidence of missing item scores in personality measurement, and simple item-score imputation. Methodology, 6(1), 17-30.

van Ginkel, J. R., van der Ark, L. A., & Sijtsma, K. (2007a). Multiple imputation for item scores when test data are factorially complex. Br.J.Math.Stat.Psychol., 60(Pt 2), 315-337.

van Ginkel, J. R., Van der Ark, L. A., & Sijtsma, K. (2007b). Multiple Imputation of Item

Scores in Test and Questionnaire Data, and Influence on Psychometric Results. Multivariate Behavioral Research, 42(2), 387-414.

Gold, M. S., & Bentler, P. M. (2000). Treatments of Missing Data: A Monte Carlo Comparison of RBHDI, Iterative Stochastic Regression Imputation, and Expectation-Maximization. Structural Equation Modeling: A Multidisciplinary Journal, 7(3), 319-355.

Hawthorne, G., & Elliott, P. (2005). Imputing cross-sectional missing data: comparison of common techniques. Aust.N.Z.J.Psychiatry, 39(7), 583-590.

Heymans, M. W., de Vet, H. C., Bongers, P. M., Knol, D. L., Koes, B. W., & van Mechelen, W. (2006). The effectiveness of high-intensity versus low-intensity back schools in an occupational setting: a pragmatic randomized controlled trial. Spine (Phila Pa 1976), 31(10), 1075-1082.

SPSS Inc. (2008). SPSS Statistics for Windows (Version Version 17.0). Chicago: SPSS Inc.

SAS Institute Inc. (2011). SAS/IML software, Version 9.2 of the SAS System for Windows. Cary, NC, USA.

Janssen, K. J., Donders, A. R., Harrell, F. E., Jr., Vergouwe, Y., Chen, Q., Grobbee, D. E., & Moons, K. G. (2010). Missing covariate data in medical research: to impute is better than to ignore. J Clin Epidemiol, 63(7), 721-727.

Kneipp, S. M., & McIntosh, M. (2001). Handling missing data in nursing research with multiple imputation. [Review]. Nurs Res., 50(6), 384-389.

Kraaimaat, F. W., & Evers, A. W. (2003). Pain-coping strategies in chronic pain patients: psychometric characteristics of the pain-coping inventory (PCI). Int J Behav Med, 10(4), 343-363.

Little, R. J., & Rubin, D. B. (2002). Statistical Analysis with Missing Data (Second Edition ed.). Hoboken, NJ: John Wiley & Sons.

Marshall, A., Altman, D. G., Royston, P., & Holder, R. L. (2010). Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. [Comparative Study Research Support, Non-U.S. Gov't]. BMC Med Res Methodol., 10, 7.

Newman, D. A. (2003). Longitudinal Modeling with Randomly and Systematically Missing Data: A Simulation of Ad Hoc, Maximum Likelihood, and Multiple Imputation Techniques. Organizational Research Methods, 6(3), 328-362.

Pastor, J. B. N. (2003). Methods for the Analysis of Explanatory Linear Regression Models with Missing Data Not at Random. Quality & Quantity: International Journal of Methodology, 37(4), 363-376.

Roth, P. L., Switzer, F. S., & Switzer, D. M. (1999). Missing Data in Multiple Item Scales: A Monte Carlo Analysis of Missing Data Techniques. Organizational Research Methods, 2(3), 211-232.

Rubin, D. B. (1976). Inference and missing data. Biometrika, 63(3), 581-592.

Rubin, D. B. (2004). Multiple Imputation for Nonresponse in Surveys. New York: John

Wiley and Sons.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. Psychol. Methods., 7(2), 147-177.

Siddique, J., Harel, O., & Crespi, C. M. (2012). Addressing Missing Data Mechanism Uncertainty using Multiple-Model Multiple Imputation: Application to a Longitudinal Clinical Trial. Ann Appl Stat, 6(4), 1814-1837.

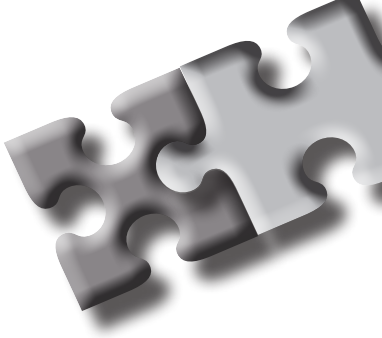
StataCorp. (2011). Stata Statistical Software (Version Release 12). College Station, TX: StataCorp LP.

R Core Development Team (2012). R: A language and environment for statistical computing. Vienna, Austria.

Venables, W. N., & Ripley, B. D. (2002). Modern Applied Statistics with S. Fourth Edition. New York: Springer.

de Vet, H. C. W., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011). Measurement in Medicine. Cambridge: Cambridge University Press.

Ware, J. E., Jr., Kosinski, M., & Keller, S. D. (1994). SF-36 Physical and Mental Health Summary Scales: A User's Manual. Boston, MA: Health Assessment Lab.



Chapter 4

Multiple imputation at the item level when the number of items is very large

Under review: Eekhout, I., de Vet, H.C.W., de Boer, M.R., Twisk, J.W.R., Heymans, M.W. Passive imputation of missing values in studies with many multi-item questionnaire outcomes. Quality of Life Research.

Abstract

Previous studies showed that missing data in multi-item questionnaires should be handled by multiple imputation. However, when many questionnaires are used the number of item variables with missing values will become too large to reliably estimate imputations. Passive imputation methods have been developed to combine variables in the imputation model to reduce information, which has never been studied before in the situation of missing item scores. In a simulation study we compared five methods as part of the multiple imputation procedure in RCTs with complete-case analysis when item scores were made missing in five different multi-item questionnaires. Method 1 and 2 used passive imputation, which updated the questionnaire total scores from imputed item variables between imputations, method 3 used parcel summary scores of the items, method 4 used all items at once and method 5 directly imputed the total scores. Descriptive statistics of questionnaire total scores and treatment coefficient estimates from linear regression were compared to 'true' parameters on bias, mean squared error and coverage. Passive imputation and using parcel summaries showed a standardized bias of less than 10%, while imputing the total score directly a standardized bias of over 60% for the questionnaire total scores. The sample size for imputing total scores needs to be at least 23% larger to attain the same mean squared error in regression coefficients compared to passive imputation. Item imputations are most valid when passive imputation or parcel summary scores are used. These methods are therefore recommended for missing data in multi-item questionnaires.

Keywords: multi-item questionnaires, missing data, multiple imputation, passive imputation, large survey studies

Introduction

Many epidemiological and medical studies use multiple questionnaires to measure patient characteristics or disease outcome. These questionnaires consist of several items which are usually measured at several time-points resulting in a dataset with many variables representing the items. Subsequently, the item scores for each questionnaire are summed up and the total scores can be used in the analyses as predictors, covariates or outcomes. However, often these questionnaires contain missing data on the item scores, which impairs the calculation of the total scores.

An advanced method to handle missing data is multiple imputation (MI) (Rubin, 1987; Schafer, 1997). In MI the missing values are replaced by multiple plausible values resulting in multiple copies of the dataset, each with different imputed values for the missing entries. The plausible values are estimated in an imputation model utilizing regression models to predict and replace missing values based on the observed data. The data analysis is then performed on each imputed dataset, resulting in multiple sets of results. In the end, these sets of results are pooled into one final result. Multiple imputation can either be applied to the questionnaire item scores before the total score is calculated, or to the total scores directly, in which case the total scores are incomplete when one or more items are missing. From previous studies we know that it is most advantageous to handle missing data in multi-item questionnaires at the item level (Eekhout et al., 2014; Gottschall, West, & Enders, 2012).

Since multiple imputation involves using regression models to estimate the imputed values, the rules and assumptions for regression analyses also apply to multiple imputation. One limitation of regression analysis is that the number of independent variables cannot be too large, which can be a problem when all items are included at once in the multiple imputation model (Green, 1991; Hardt, Herke, & Leonhart, 2012). By running the imputation process for each questionnaire or outcome separately, information is lost because the questionnaires might be related to each other. It is recommended to include all possible information in the imputation model and therefore, it is most informative to incorporate all questionnaires at once to deal with missing data (Collins, Schafer, & Kam, 2001).

A possible solution to avoid the problem of having imputation models that are too large is to use total scores of questionnaires in the imputation model as predictors for the missing item scores instead of using only item scores. This seems like a straightforward solution, however, these total scores often contain missing values, which are also caused by missing item scores. A solution is to adapt the imputation process in such a way that the total score will be calculated after each imputation run (i.e., iteration) using the imputed item scores. This is possible by an application called passive imputation. Passive imputation can be used to make sure

that a derived variable (e.g., a questionnaire total score calculated by the sum of the item scores) always depends on the most recently generated item imputations in the original data (van Buuren, 2012). Accordingly, between imputation iterations, the total score is updated from the most recently imputed item scores, which is the passive part of the imputation. Furthermore, during the iteration process the total scores of the questionnaire cannot be used as a predictor for items of that specific questionnaire, but only as a predictor for the items of other questionnaires. Passive imputation seems perfectly designed to handle missing values in different items using several different questionnaires, with the benefit of maintaining the imputation model without the problem of a large number of variables in the imputation model.

Passive imputation in the context of interaction variables (i.e., ratios of variables) has been studied previously and was found to result in biased regression estimates (Morris, White, Royston, Seaman, & Wood, 2014; Von Hippel, 2009). The application of passive imputation for questionnaires was briefly proposed by Van Buuren (2010), however the validity of this method under different data situations has not been studied before. This study therefore evaluated two procedures of the passive imputation method for the imputation of item scores in simulated data. Furthermore, these passive imputation methods were compared to a practical method that imputes the items by using a parcel summary score of the other questionnaires as predictors, which can be applied in any software package. Finally, these methods were also compared to imputing the item scores, imputing total scores directly and to a complete-case analysis (CCA). The latter two methods are mostly used in practice (Eekhout, de Boer, Twisk, de Vet, & Heymans, 2012).

Methods

Simulation study design

We simulated data for five questionnaires (Q1 to Q5). The first simulated questionnaire (Q1) contained 5 items, the second (Q2) and third (Q3) questionnaires contained 10 items and the fourth (Q4) and fifth (Q5) 15 items. All items were measured on a five-point Likert scale. We simulated a randomized controlled trial situation with a pre and a post measurement for the questionnaires, because this is a frequently applied study design in epidemiological studies. Additionally, two baseline continuous covariates were simulated and a random dichotomous treatment variable. This resulted in a total of 55 items measured at two time-points, two time-invariant continuous covariates and one dichotomous time-invariant covariate (i.e., treatment). For the simulation we used a predefined treatment effect of 0.50 for the total scores. Furthermore, we varied between two sample size conditions: 150 subjects and 250 subjects per simulated dataset separated in two equal treatment groups. We

generated 1000 samples in each sample size condition. The complete data samples were used as a reference to compare the performance of the missing data methods to. The complete samples were created in Mplus (Muthén & Muthén, 1998-2012). Subsequently, missing data in the items were generated in all questionnaires by the missing at random mechanism (Rubin, 1976). For 10-25% of the subjects only some items were incidentally made missing within subjects (i.e., < 75% of the items) and for 0-12% of the subjects a whole questionnaire was made missing (i.e., > 75% of the items were missing). These percentages varied per questionnaire. The probability of missing an entire questionnaire was larger after treatment than at baseline (i.e., 0-6% at baseline and 6-12% post-treatment). That way a realistic data situation was simulated where some people skipped some questionnaire items and other people didn't fill out an entire questionnaire. The overall percentage of subjects with missing data was simulated to be 30%. The missing data in the samples were generated in R statistical software (R Core Development Team, 2014).

Compared multiple imputation methods

In the simulated datasets, the missing data were handled with multiple imputation by multivariate imputation by chained equations (MICE) (van Buuren & Groothuis-Oudshoorn, 2011). The ordinal Likert items of the questionnaires in our simulated data were imputed with the predictive mean matching method, which assumes normality but was shown to work for ordinal items as well (Eekhout, et al., 2014). There are several options for specification of the imputation model when questionnaire items and total scores at multiple time-points (e.g., baseline and post-treatment) are involved. We compared four different imputation models that were targeted at imputing missing item scores and one model that imputed the total scores of the questionnaires directly and a CCA. Each imputation model included the treatment group variable and the two other covariates. The following methods were compared:

Method 1: Passive imputation A (M1-Passive)

For this passive imputation procedure the imputation model consisted of the following variables: the item variables of a questionnaire assessed at a certain time-point, the item variables from this questionnaire assessed at other time-points and the total scores of the other questionnaires at both time-points. With this method the total scores are updated after each imputation iteration by the imputed items from the previous iteration. Then the updated total scores become the predictors for the item variables that contain missing values during the next iterations. For the smaller questionnaire (Q1) the imputation model contained 21 variables (i.e., 5 items at baseline, 5 items post-treatment for Q1, 8 total scores for Q2-Q5 at baseline and post-treatment, the treatment group variable and the two covariates), for Q2 & Q3

31 variables and for the larger questionnaires (Q4 & Q5) 41 variables.

Method 2: Passive imputation B (M2-Passive)

The second imputation model also included passive imputation, where the items for a questionnaire at a certain time-point, the total score of that questionnaire at the other time-point and the total scores of all other questionnaires at both time-points were included in the imputation model. Accordingly, the method includes fewer variables in the imputation model at once compared with M1-Passive. The model contained 17 variables for the smaller questionnaire (Q1) (i.e., the 5 items of Q1 at baseline, the total score for Q1 post-treatment, 8 total scores for Q2-Q5 at baseline and post-treatment, the treatment group variable and the two covariates), 22 variables for Q2 & Q3 and 27 variables for the larger questionnaires (Q4 & Q5).

Method 3: Parcel summary model (M3-Parcel)

The third imputation model included the average of the available items of the other questionnaires as predictors for the missing item scores of a questionnaire. The average over the available items is a parcel summary score of the item information, which was calculated once for each questionnaire prior to the start of the imputation process. This method does not use passive imputation, but the same parcel summary scores were used as predictors in the imputation model for each questionnaire. In M3-Parcel the imputation for each separate questionnaire was done independently and we merged the resulting imputed datasets after the imputation process was completed. In this method 21 variables were in the imputation model for the smaller questionnaire (Q1) (i.e., 5 items at baseline, 5 items post-treatment for Q1, 8 parcel summary scores for Q2-Q5 at baseline and post-treatment, the treatment group variable and the two covariates), for Q2 & Q3 31 variables and 41 variables for the larger questionnaires (Q4 & Q5).

Method 4: All item scores (M4-Items)

In the fourth imputation model all items in the dataset were included at once. Consequently, 110 items, the treatment group variable and the two covariates were all entered in the imputation model at once. This model was expected to encounter convergence problems, especially in the smaller sample size condition, but it was included as a comparison.

Method 5: Total scores (M5-TS)

The fifth imputation model was targeted at total scores directly. The total scores were computed prior to imputation and were incomplete when one or more items

were missing. The imputation model included only the total scores of all questionnaires Q1 to Q5 at both time-points, the treatment group variable and the two covariates (i.e., 13 variables). The item scores were ignored in this procedure.

Method 6: Complete-case analysis (M6-CCA).

We also performed a CCA on the data where the total scores were also left incomplete when one or more item scores were missing. In the CCA only the subjects with completely observed data were included in the data analysis. This resulted in a reduction of the sample size of 30%.

As described above, our simulated data contained subjects who had incidental missing item scores as well as subjects that missed data on an entire questionnaire. To accommodate this, we applied the method 1 to 4 as follows. For the subjects that had the entire questionnaire missing (i.e., >75% of the item scores missing within a subject), the total scores were imputed directly. Subsequently, for the subjects that had less than 75% of the item scores missing (i.e., incidental missing item scores within subjects), the item scores were imputed according to the method 1 to 4. Then after the imputation, but before analysis, we selected the total score from the imputed items for the people who missed only some items of a questionnaire and for the people that missed the entire questionnaire, we selected the directly imputed total score.

Analyses

The imputed and thus complete data were analyzed by linear regression analysis using the questionnaire total score after treatment as the outcome and the treatment variable as covariate, adjusted for the baseline measurement of the questionnaire. We analyzed each of the five questionnaires (the outcome variables) separately and we were interested in the treatment coefficients. Furthermore, we computed the means of the questionnaire total scores at baseline and post-treatment. The results for all methods were compared to the complete data results using bias, mean squared error (MSE) and coverage. The bias was evaluated by examining the standardized bias, which reflects the bias relative to the overall uncertainty in the sampling (Collins, et al., 2001). The standardized bias was calculated by

$$Standardized\ bias = \frac{\bar{\hat{\beta}} - \bar{\beta}_c}{sd(\beta_c)}\ 100\%$$

where $\bar{\hat{\beta}}$ *is* the average parameter estimate (e.g., the questionnaire mean or the treatment regression coefficient) obtained from the estimates in the simulated datasets after the missing data method was applied, $\bar{\beta}_c$ *is* the average true parameter of the simulated complete reference data and $sd(\beta_c)$ is the standard deviation of

the estimates of the complete data. The MSE represents the precision and accuracy of the estimates and was calculated by

$$MSE = \left(\bar{\hat{\beta}} - \bar{\beta}_c\right)^2$$

The coverage was calculated by the percentage of times the average complete data estimate $\bar{\beta}_c$ was within the 95% confidence interval of the estimated parameters $\hat{\beta}_i$. For a 95% confidence interval, the coverage rate should be at 95%. Coverage rates higher than 95% indicate that the method might be too conservative, and a lower coverage rate suggests higher than expected type I error (Burton & Altman, 2004). All imputations and analyses were performed in R statistical software (R Core Development Team, 2014). A detailed manual on how to apply the imputation methods is available by the first author.

Results

In Figure 4.1 the standardized bias of the regression coefficients for the treatment coefficient for each questionnaire outcome are presented for the missing data methods at both sample size conditions.

For the sample size of 150, the method that included all items in the imputation model (M4-Items), could not be computed because of an excess of items in the imputation model. Overall we can observe that all methods including the CCA resulted in small standardized bias for the regression coefficients. In both sample size conditions the standardized bias in the regression coefficients (i.e., in treatment effect) was smaller than 10%. In the small sample size condition, the methods applied to the item scores (M1-Passive, M2-Passive & M3-Parcel) performed slightly better than the method applied to total scores (M5-TS); however, differences are very small (Figure 4.1a). In the larger sample size condition, multiple imputation applied to the total scores (M5-TS) had a standardized bias comparable to the methods applied to the item scores (Figure 4.1b).

The MSE of the regression coefficient estimates increased when the number of items per questionnaire increased (Figure 4.2). Furthermore, we can observe that the methods that imputed the item scores (M1 to M4) had a smaller MSE (i.e., more precision and accuracy) than the method that was applied to the total scores (M5-TS) and this difference increased when the number of items per questionnaire was increasing. For the smaller questionnaire (Q1) at n=150 the MSE was 1.22 for both passive imputation methods (M1-Passive & M2-Passive) and for imputing total scores (M5-TS) 1.35. The ratio of these MSEs (1.35/1.22=1.11) indicates the sample size increase required for imputing total scores to attain the same precision as the passive imputation methods, which is 11%. For the larger questionnaires (e.g., Q5) at n=150 the MSE of passive imputation was 9.43 and for imputing total scores 12.36,

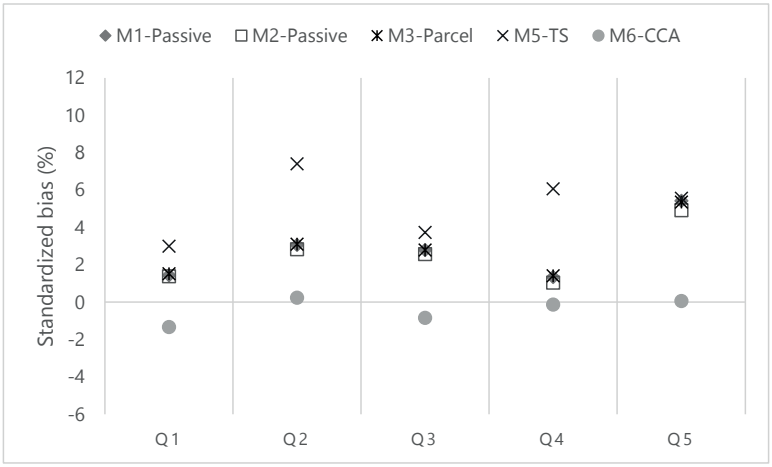


Figure 4.1a. Standardized bias of the regression coefficients for treatment effect for each of the five questionnaires for the missing data methods (n=150). M4-Items could not be performed for n=150.

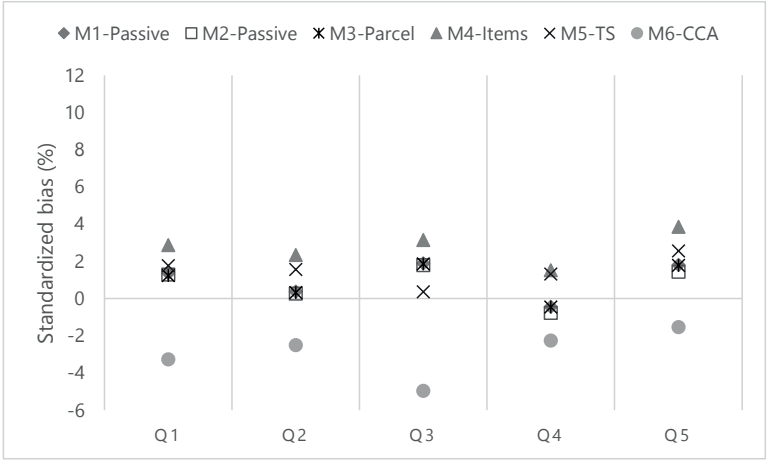


Figure 4.1b. Standardized bias of the regression coefficients for treatment effect for each of the five questionnaires for the missing data methods (n=250).

accordingly the ratio was (12.36/9.43) 1.31, which indicates a required sample size increase of 31% for imputing total scores to reach the same precision as passive imputation. On average over all questionnaires (Q1 to Q5) the ratio of the MSE of passive imputation to the MSE of imputing total scores was 1.23, which means that the sample size should increase by at least 23% for imputing total scores to attain the same precision as passive imputation. The MSE of CCA was even worse (i.e., larger); the average ratio of the MSE of passive imputation to the MSE of CCA was 1.33, which indicates required a sample size increase by 33% for CCA to achieve the same precision as passive imputation. The coverage rates of the regression coefficient estimates were satisfactory for all methods for both sample size conditions (data not shown).

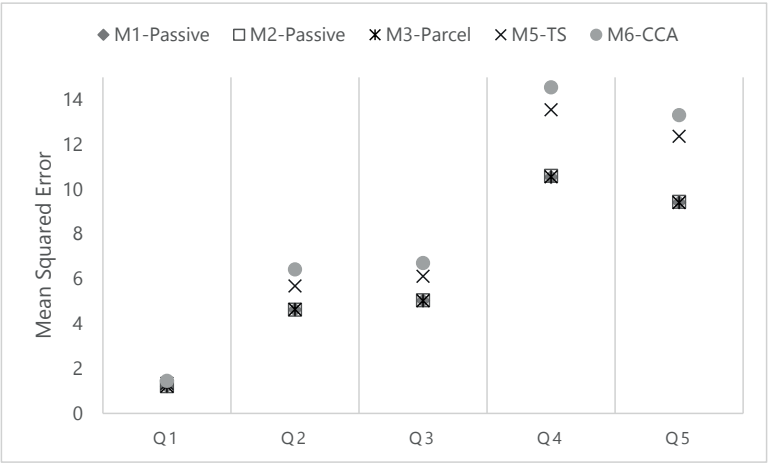


Figure 4.2a. Mean Squared Error (MSE) for the regression coefficients for treatment effect of the analysis of the five questionnaires for the missing data methods (n=150). M4-Items could not be performed for n=150.

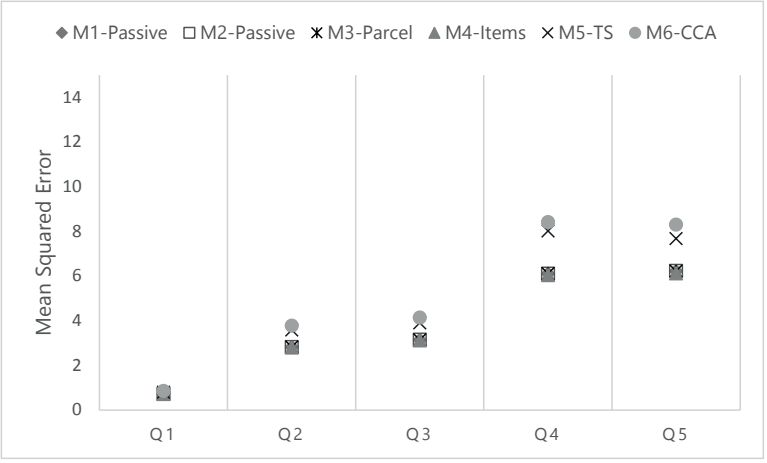


Figure 4.2b. Mean Squared Error (MSE) for the regression coefficients for treatment effect of the analysis on the five questionnaires for the missing data methods (n=250)

The coverage rates of the regression coefficient estimates were satisfactory for all methods for both sample size conditions (data not shown).

In Figure 4.3 the standardized bias of the means of the questionnaire total scores are presented for each method. In both sample size conditions we can observe that the passive imputation methods (M1-Passive & M2-Passive) performed best and were least biased. The parcel summary method (M3-Parcel) performed similarly to the passive imputation methods. In the larger sample size conditions (n=250), the method that included all the items in the imputation model at once (M4-Items), also performed similarly. Imputing the total score directly (M5-TS) performed less

favorable, i.e., larger standardized bias. CCA performed worst and showed bias larger than 1 standard error (i.e., 100%) in both sample size conditions.

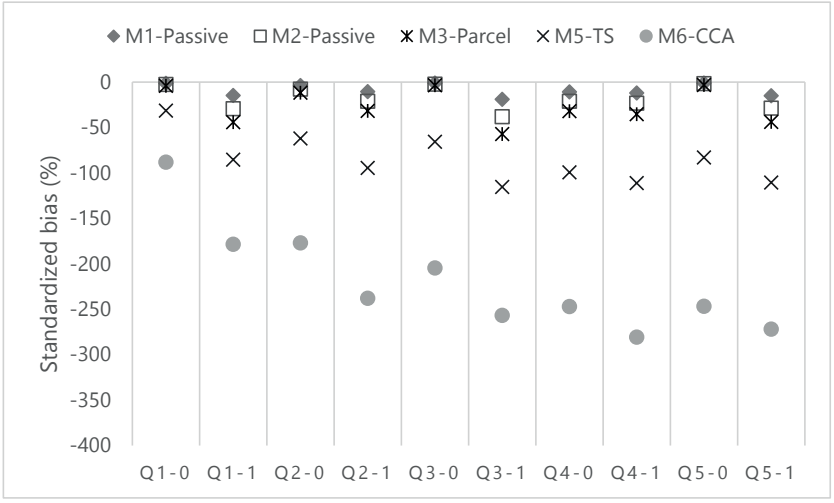


Figure 4.3a. Standardized bias of questionnaire total score mean at baseline (0) and after treatment (1) for each questionnaire for the compared missing data, n=150. M4-Items could not be performed for n=150.

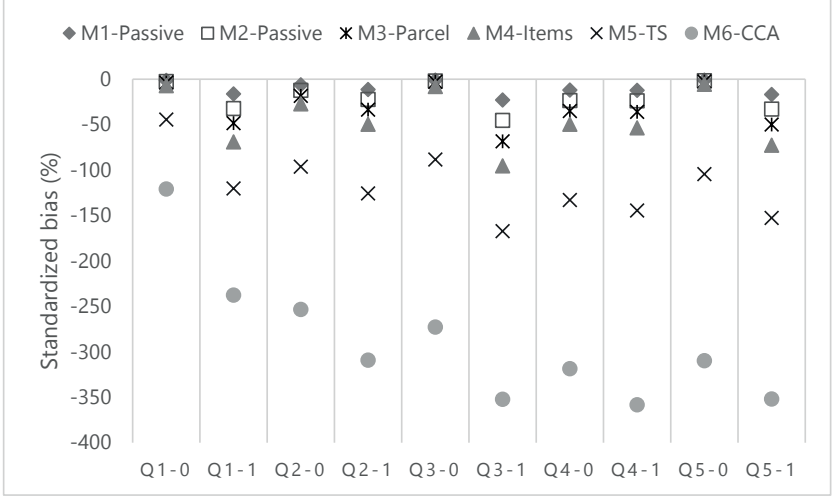


Figure 4.3b. Standardized bias of questionnaire total score mean at baseline (0) and after treatment (1) for each questionnaire for the compared missing data, n=250.

In Figure 4.4 the average MSE of the questionnaire total scores for each imputation method are presented. We can observe the trend that the methods that impute the item scores (M1-M4) had the lowest MSE and performed almost similarly. The imputation of the total scores (M5-TS) resulted in about one-and-a-half times larger MSE and the CCA is even less precise.

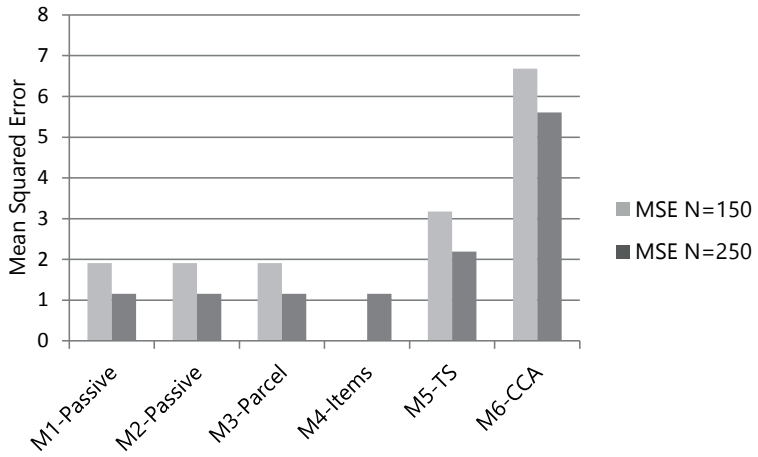


Figure 4.4. MSE of questionnaire total score means averaged for all questionnaires at baseline and after treatment for the missing data methods, at the two sample size conditions (n=150 & n=250). M4-Items could not be performed for n=150.

Discussion

In this paper we compared possible solutions for the problem of missing data when data from many questionnaires are used in one study. To our knowledge, the proposed new methods that use passive imputation (M1-Passive & M2-Passive) had not been validated yet. The results of our simulation show that using passive imputation to impute the item scores of multiple questionnaires is a valid solution for missing item score data in questionnaires. In fact, all methods in which the item scores are imputed by using the total scores as predictors in the imputation model, including the method where a parcel summary score of the items from the other questionnaires was used, performed better than handling the missing data directly at the total score level. This illustrates the importance of including the item scores in the imputation model. Even when the most optimal solution (i.e., including all available item information) is not viable, it is important to apply an advanced missing data method that incorporates the item scores. In this study we found differences in performance for the methods when they were applied to estimate the group means or

the treatment effects. In the analysis of treatment effects, CCA was unbiased, though the precision reflected in the MSE was much better for the imputation methods. For the estimation of the group means the imputation methods performed superiorly compared to a CCA on all evaluation measures (i.e., bias and MSE).

From previous simulation studies we know that it is most beneficial to handle missing data in multi-item questionnaires at the item level (Eekhout, et al., 2014). However, the most appropriate method, i.e., the imputation model where solely the item score information was incorporated (M4-Items), is only viable with a sufficiently large sample size. For that reason, in many situations a solution that includes a smaller imputation model is necessary. The methods that incorporate the item score information most optimally were the passive imputation methods that impute the item scores by using the total scores from the other questionnaires as predictors in the imputation model (M1-Passive & M2-Passive). In these methods, the total scores were updated by the imputed items within the imputation process. That way the most optimal available information was used at every stage of the imputation process. The performance of the passive imputation methods was best on all accounts (i.e., descriptive statistics and regression estimates of the questionnaire data). For that reason this strategy is advised. We compared two different ways to apply these methods. The first strategy (M1-Passive) was to include the item scores for both time-points of one questionnaire together in one batch of imputations. The second strategy was to impute the item scores for the baseline of one questionnaire separately from the post-treatment item scores of that questionnaire, so for each time-point separately. Both strategies performed equally well, so the only preference is related to the number of variables that is included in the imputation model, which might be preferred to be kept lower (i.e., M2-Passive).

Previous studies evaluated the performance of passive imputation when the imputation model contained ratios of variables (Morris, et al., 2014), interactions between variables or squared variables (Von Hippel, 2009). For these kinds of composed variables, the variables are separately imputed in their raw form, and then their composed values are calculated between each iteration of imputations as an update. In that case, the raw variables are imputed in the imputation model, while the analysis model contains the transformed variables (i.e., ratios, interactions or squares). These studies concluded that the use of passive imputation for these kinds of variables can result in biased parameter estimates, because the covariance between the predictor variables and the outcome variable in the imputation model is different from the covariance between the predictors and the outcome in the analysis model. For that reason, these studies advised to transform the variables prior to the imputation. In the present study we did not use any transformations and therefore the covariance matrix in the imputation model and the analysis model are compatible.

Passive imputation is available in the MI procedure in STATA, the MICE package in R and S-PLUS and in IVEware in SAS. The application of this method requires an advanced level of programming and might therefore not be feasible in daily practice for many researchers. For that reason we included an imputation strategy that can be applied in most basic statistical programs that include a multiple imputation option (e.g., SPSS). This method (M3-Parcel), which uses the average of the available items from the other questionnaires to impute the items of one questionnaire, performed satisfactory with regard to most parameters as well. Only the average total scores were biased for this method, but the coefficient estimates for treatment effect and the precision of these estimates were adequate and better than imputing the total scores directly.

Strengths and weaknesses

The strength of this simulation study is that we simulated realistic situations with both items and total scores missing. Furthermore, the design of including several questionnaires in one study is very common in epidemiology. Also different lengths of questionnaires were included with two different sample size conditions. That way the sensitivity of the missing data methods for several data aspects was checked. The tested methods have not been validated before in previous studies and the results of this study underpin the good performance of the newly proposed methods.

A possible weakness of this study might be that missing data methods that are advised in most user-manuals of multi-item questionnaires were not included in the comparison. These advised methods are mostly single imputation methods, for example replacing the missing value with the mean score (Lambert, Lunnen, Umphress, Hansen, & Burlingame, 1994; Ware, Kosinski, & Keller, 1994). However, from a previous simulation study we know that these methods do not perform well and are not recommended to be applied in any missing data situation (Eekhout, et al., 2014). Furthermore, a limited amount of data conditions was compared in this study. For example, we did not vary the total number of items in the entire dataset. However, by varying the sample size we aimed to simulate a situation where the ratio of number of items and sample size varied, which is the most important reason for one of these multiple imputation methods to fail. The amount of missing data was not varied either. However we aimed to simulate a realistic amount of missing data and we expect our methods to perform similarly with less missing data. Moreover, in our previous simulation we found that multiple imputation of item scores remains to perform well up to conditions with 75% missing item scores within 75% of the subjects (Eekhout, et al., 2014).

All compared multiple imputation methods (M1-M5) are advanced methods to handle missing data. For that reason it was expected that the performance of all

methods would be satisfactory to some extent, for example on coverage values. However, this simulation showed that the new methods outperform the current solution of imputing questionnaire data when many questionnaire items are involved (i.e., M5-TS), especially on precision but also on accuracy. In general we advise to include as much item score information as possible in handling the missing data at the item score level of a multi-item questionnaire. It is best to impute the item scores by using the passive imputation procedure.

References

Burton, A., & Altman, D. G. (2004). Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. British Journal of Cancer, 91(1), 4-8.

van Buuren, S. (2010). Item Imputation Without Specifying Scale Structure. Methodology, 6(1), 31-36.

van Buuren, S. (2012). Flexible Imputation of Missing data. New York: Chapman & Hall/ CRC.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological Methods, 6(4), 330-351.

Eekhout, I., de Boer, R. M., Twisk, J. W., de Vet, H. C., & Heymans, M. W. (2012). Missing data: a systematic review of how they are reported and handled. Epidemiology, 23(5), 729-732.

Eekhout, I., De Vet, H. C. W., Twisk, J. W. R., Brand, J. P. L., De Boer, M. R., & Heymans, M. W. (2014). Missing data in a multi-item instrument were best handled by multiple imputation at the item score level. Journal of Clinical Epidemiology, 67(3), 335-342.

Gottschall, A. C., West, S. G., & Enders, C. K. (2012). A Comparison of Item-Level and Scale-Level Multiple Imputation for Questionnaire Batteries. Multivariate Behavioral Research, 47(1), 1-25.

Green, S. B. (1991). How Many Subjects Does It Take To Do A Regression Analysis. Multivariate Behavioral Research, 26(3), 499-510.

Hardt, J., Herke, M., & Leonhart, R. (2012). Auxiliary variables in multiple imputation in regression with missing X: a warning against including too many in small sample research. BMC Med Res Methodol, 12, 184.

Lambert, M. J., Lunnen, K., Umphress, V., Hansen, N., & Burlingame, G. M. (1994). Administration and scoring manual for the Outcome Questionnaire (OQ-45.1). Salt Lake City: IHC Center for Behavioral Healthcare Efficacy.

Morris, T. P., White, I. R., Royston, P., Seaman, S. R., & Wood, A. M. (2014). Multiple imputation for an incomplete covariate that is a ratio. Statistics in Medicine, 33(1), 88-104.

Muthén, L. K., & Muthén, B. O. (1998-2012). Mplus User's Guide. Seventh Edition. Los Angeles, CA: Muthén & Muthén.

Rubin, D. B. (1976). Inference and missing data. Biometrika, 63(3), 581-592.

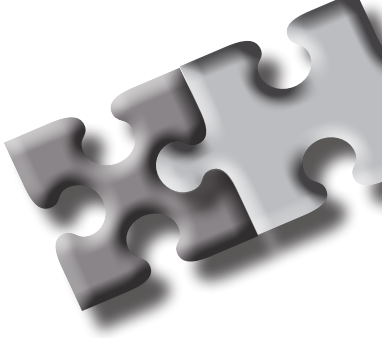
Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: J. Wiley & Sons.

Schafer, J. L. (1997). Analysis of Incomplete Multivariate Data. London, UK: Chapman & Hall.

R Core Development Team (2014). R: A language and environment for statistical computing. Vienna, Austria.

Von Hippel, P. T. (2009). How to impute interactions, squares, and other transformed variables. Sociological Methodology, 39(1), 265-291.

Ware, J. E., Jr., Kosinski, M., & Keller, S. D. (1994). SF-36 Physical and Mental Health Summary Scales: A User's Manual. Boston, MA: Health Assessment Lab.



Chapter 5

Analyzing incomplete item scores in longitudinal data by including item score information as auxiliary variables

Accepted for publication April 2014: Eekhout, I., Enders, C.K., Twisk, J.W.R., de Boer, M.R., de Vet, H.C.W., & Heymans, M.W. (in press). Analyzing incomplete item scores in longitudinal data by including item score information as auxiliary variables. Structural Equation Modeling: A Multidisciplinary Journal.

Abstract

The aim of this study is to investigate a novel method for dealing with incomplete scale scores in longitudinal data that result from missing item responses. This method includes item information as auxiliary variables, which is advantageous because it incorporates the observed item-level data while maintaining the scale scores as the focus of the analysis. These auxiliary variables do not change the analysis model, but improve missing data handling. The investigated novel method uses the item scores or some summary of a parcel of item scores as auxiliary variables, while treating the scale scores missing in a latent growth model. The performance of these methods was examined in several simulated longitudinal data conditions and analyzed through bias, mean squared error, and coverage. Results show that including the item information as auxiliary variables results in rather dramatic power gains compared with analyses without auxiliary variables under varying conditions.

Keywords: structural equation modeling, missing data, questionnaires, longitudinal data, full information maximum likelihood, auxiliary variables

Introduction

Many studies use multi-item questionnaires to collect information about a certain construct of interest. This construct is often measured as a scale score calculated by the sum of the item scores. When the item scores contain missing data, the calculation of the scale score becomes difficult. As a solution, many questionnaire manuals advise to compute scale scores by averaging the available items. This method is also known as person mean imputation, because it is equivalent to imputing the missing values with a person's average item response (Bernaards & Sijtsma, 2000; Fayers, Curran, & Machin, 1998; Hawthorne & Elliott, 2005). However, this method has no theoretical basis, decreases variance and can introduce biased estimates especially when the internal consistency of the scale is not very high and data are missing at random (i.e., other variables fully explain the propensity for missing item responses (Rubin, 1976)). Ergo this method is often sub-par (Eekhout et al., 2014; Schafer & Graham, 2002). A second option for computing scale scores is to limit the analysis to only those cases with complete data on all items, otherwise known as complete-case analysis. This method requires missing completely at random as an assumption (i.e., the missing part of the data is a completely random subsample of the data), otherwise it can result in biased estimates, even when only a small number of items are missing per case. Additionally, complete-case analysis can lead to a considerable loss of power because the sample size is reduced to only fully observed cases. Although complete-case analysis and other ad-hoc methods, such as mean imputation, can bias analysis results, these methods are still popular in many fields of research (Eekhout, de Boer, Twisk, de Vet, & Heymans, 2012; Karahalios, Baglietto, Carlin, English, & Simpson, 2012).

Currently recommended advanced missing data methods are multiple imputation and full information maximum likelihood estimation (FIML). These methods work well when missing questionnaire variables are missing at random (MAR; Little & Rubin, 2002). In multiple imputation and FIML, all available information in the data is used for estimations. Furthermore, with these techniques model estimations are generally unbiased and don not lose power, even when data are missing completely at random. In multiple imputation missing values are imputed prior to the analysis. Multiple imputation is performed in three phases. In the first phase, the imputation phase, incomplete values are imputed according to an imputation model, which is a regression model that estimates the predicted scores for incomplete data. In order to account for uncertainty around the imputed values, random error is added to each predicted score from the regression model, and the sum of the predicted score and the error term is the imputed value. This imputation process is repeated multiple times resulting in multiple imputed datasets. In the analysis phase, the data analysis is performed on each of these imputed datasets and results are pooled afterwards

in the pooling phase, to obtain the final analysis result (Rubin, 1987; Schafer, 1997; van Buuren, 2012). In FIML missing values are not replaced or imputed, instead the observed data are used to estimate the population parameters with the highest likelihood of producing the sample data.

Since multi-item questionnaires are mostly used to measure a scale score by summing items, incomplete item score data on these instruments can be handled at either the item score level or at the scale score level. Thus we can apply a missing data method to the incomplete item scores, then sum these item scores to a scale score and use these total scores for our analysis. Alternatively, we can treat the scale score as missing for cases with one or more missing item responses, then apply a missing handling technique to the scale score (e.g., impute the scale scores, or submit the incomplete scale scores to a FIML analysis). Previous studies have shown that in the context of multiple imputation, incomplete item data are best handled at the item-level (Eekhout, et al., 2014; Gottschall, West, & Enders, 2012). Handling missing values at the item-level has a substantial benefit on power. Gottschall et al. (2012) found in their simulation study that, in certain situations, imputing the incomplete scale scores required a 75% increase in the sample size in order to yield the same power as an analysis that imputed the incomplete item responses. Furthermore, the power benefit for handling missings at the item-level increases as the number of questionnaire items increases. However, when the number of items is larger, including all available information from the items might be computationally difficult. Calculations might even become impossible as the number of items (i.e., variables) to be estimated in the model grows closer to the sample size. Since multiple imputation and FIML are asymptotically equivalent (Collins, Schafer, & Kam, 2001), also in FIML handling missing item scores at the item-level should improve power and accuracy. However, no previous studies have discussed FIML methods for dealing with item-level missing data in a scale score analysis.

Structural equation modeling programs are an ideal method for implementing FIML. One way to handle missing item scores in a structural equation model is to treat the items as indicators of a latent factor. However, when the complete-data analysis is based on the scale scores, treating the individual items as indicators requires altering the analysis model to accommodate the missing data. Since the items are modeled as indicators for a latent factor, the scale scores are not represented as a raw sum of the item scores, but as these latent factors. This means that the interpretation of the model coefficients is different from the interpretation of the coefficients of the model that would have been fitted had the data been complete, which is a practical disadvantage. The aim of this study is to propose two novel methods for dealing with incomplete scale scores that result from missing item responses by including the item information in the model as auxiliary variables. These methods are advantageous

because they incorporate the observed item-level data while maintaining the sum scores as the outcome of the analysis. Our specific interest is in applying this method to growth curve models that use scale scores as indicators.

The organization of the manuscript is as follows. First, we give a brief description of a motivating example we use throughout the paper. Next, an overview of auxiliary variables and their application to the motivating example is provided. In the section that follows the first simulation study that investigates two methods to include auxiliary variables when item scores are missing is described. After that the second simulation study which investigates one of these methods more broadly over many longitudinal conditions is presented followed by the discussion.

Motivating example

In order to illustrate our methods we use an example about physical functioning. The example data was adapted from a randomized controlled trial by Hoving et al. (2002) about neck pain, where the physical functioning scale of the SF-36 was used as a secondary outcome measure (Ware, Kosinski, & Keller, 1994). The effectiveness of three treatments was compared, which were manual therapy (specific mobilization techniques), physical therapy (exercise therapy) and continued care by a general practitioner (analgesics, counseling and education). The short term effects of these treatments on physical functioning were measured at 3 and 7 weeks by 10 items. The original data contained 170 participants with complete data at all waves, in which missing item scores were artificially created for this illustration. The missing item scores were created in the wave at 3 weeks and the wave at 7 weeks and were related to the covariate age and to treatment group in order to satisfy the missing at random assumption. At the 3 week wave item 1, 2, 3, 5, 8 and 9 contained missing values varying from 10% for item 5, 8 and 9 to 15% for item 1, 2 and 3. At the 7 week wave item 1, 2, 4, 7, and 8 had missing values varying from 10% for item 2, 7 and 8 to 15% and 25% for item 1 and 4 respectively.

Throughout the manuscript we used a latent growth model to demonstrate the methods we investigated. For the example about physical functioning we were interested in the change over time of the physical functioning scale score related to the treatment. The physical functioning scale score was calculated by the sum of the 10 items. In the example study participants were separated in three treatment groups and age was used as a covariate. The latent growth model assessed the average change in physical functioning scale score over time in the mean slope and mean intercept per treatment group corrected for age. Furthermore, the model also incorporated the variation between the individuals for both intercept and slope parameters; because person A might have a different initial physical functioning score and a different rate of change in physical functioning score than person B

even though they were both in the same treatment group. This variation between individuals was measured in the variance of both intercept (ζ_i) and slope (ζ_s). Figure 5.1 depicts a path diagram of the growth model for the physical functioning scale scores. The treatment groups were indicated by two dummy variables as predictors. Note that this model could also be parameterized as a multiple group model, with the dummy variables defining group membership. The model that we use is somewhat more parsimonious because it assumes a common covariance structure for the three groups. The loadings for the intercept factors are fixed at 1 and the loadings for the slope factor are set at the time scores (i.e., 0, 3, 7).

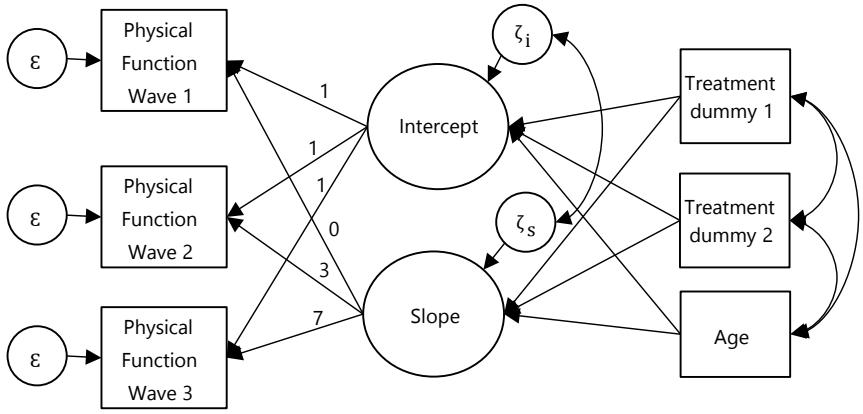


Figure 5.1. Path diagram of a latent growth model with outcomes measured at three time points. Treatment dummy 1 denotes physical therapy versus continued care by a general practitioner; Treatment dummy 2 denotes manual therapy versus continued care by a general practitioner.

Auxiliary variables

Auxiliary variables are variables that are correlated with incomplete variables and/or correlated with missingness (Collins, et al., 2001). Including variables related to missingness in the handling of missing data improves model estimations, because more information is taken into account. Collins et al. (2001) concluded that auxiliary variables are very helpful to reduce estimation bias and restore power lost due to missingness. The benefit from auxiliary variables is the same for multiple imputation and for FIML methods. Including auxiliary variables without adjusting the analysis model is fairly straightforward in multiple imputation; the variables related to missingness can be included in the imputation model (Bell & Fairclough, 2013). After the multiple imputation procedure, the data analysis is performed on the model of substantive interest without regard to the auxiliary variables. Also in FIML it is

beneficial to include the predictors of missingness (i.e., auxiliary variables) in the model (Collins, et al., 2001; Graham, 2003). For example Raykov et al. (2014) present a FIML method to estimate and test measure correlations in incomplete data. In the proposed method auxiliary variables are included in order to enhance the probability of the MAR assumption. However, in a structural equation model, such as a latent growth model, it is somewhat more complicated to incorporate auxiliary variables compared with multiple imputation. Graham (2003) formulated the following rules for including auxiliary variables in a structural equation model with latent variables: (a) auxiliary variables should be correlated to completely exogenous manifest variables, these are independent variables; (b) auxiliary variables should be correlated to the residuals of all manifest (i.e., measured) predicted and outcome variables in the model that are predicted by or indicators for a latent variable; and (c) auxiliary variables should be correlated with one another.

In a latent growth model where scale scores are the outcome, it is feasible to use auxiliary variables to incorporate the item-level information. In our example, the physical functioning scale scores are incomplete due to item-level missings. In this case, the item-level information can be included as auxiliary information to bolster the estimation of the incomplete scale scores. There are multiple ways to include this item-level information as auxiliary variables. One method is to take in the information by including the observed item scores as auxiliary variables, while again treating the scale scores missing. Another method to include item information is to use some function or a parcel summary of the items as auxiliary variables, while treating the scale scores missing. An example of such a parcel summary is the mean of a subset of available item scores. In our study we investigated these two novel methods to include the item information to improve the estimation of scale scores.

Method 1: items as auxiliary variables

In the first procedure to handle the incomplete physical functioning scores caused by the missing item scores, the items from the physical functioning scale are included as auxiliary variables. In this method the scale score is treated missing whenever one or more items are missing. Then when the model estimates the population parameters with the highest likelihood of producing the sample data, as a latent growth model does, the observed item scores are included in this estimation process to recapture power loss. Figure 5.2 depicts the path diagram of the physical functioning data with the item scores of the incomplete scales included as the auxiliary variables. As shown in Figure 5.2 the auxiliary variables are correlated with the independent variables and with the residuals of all measured outcome variables, but not to the latent intercept and slope variables. The auxiliary variables are also correlated with each other. To avoid visual clutter, the diagram shows the model set up with three

auxiliary item scores per wave, but a larger set can be included. Additionally, at least one item must be omitted from the model to circumvent linear dependencies and consequent lack of convergence. In practice, excluding the item with the highest missing data might be desirable because it would likely contain the least amount of auxiliary information. We apply this strategy later in the example data, where we also excluded the item scores from the baseline measurement (wave 1), because all items at this wave are complete. Consequently, we selected all but the first item for wave 2 and all but the fourth item for wave 3 to act as auxiliary variables in the model. Including the item scores into the model this way, should improve the precision of the parameter estimates, because correlations allow the item-level information to be transmitted to the incomplete scale scores.

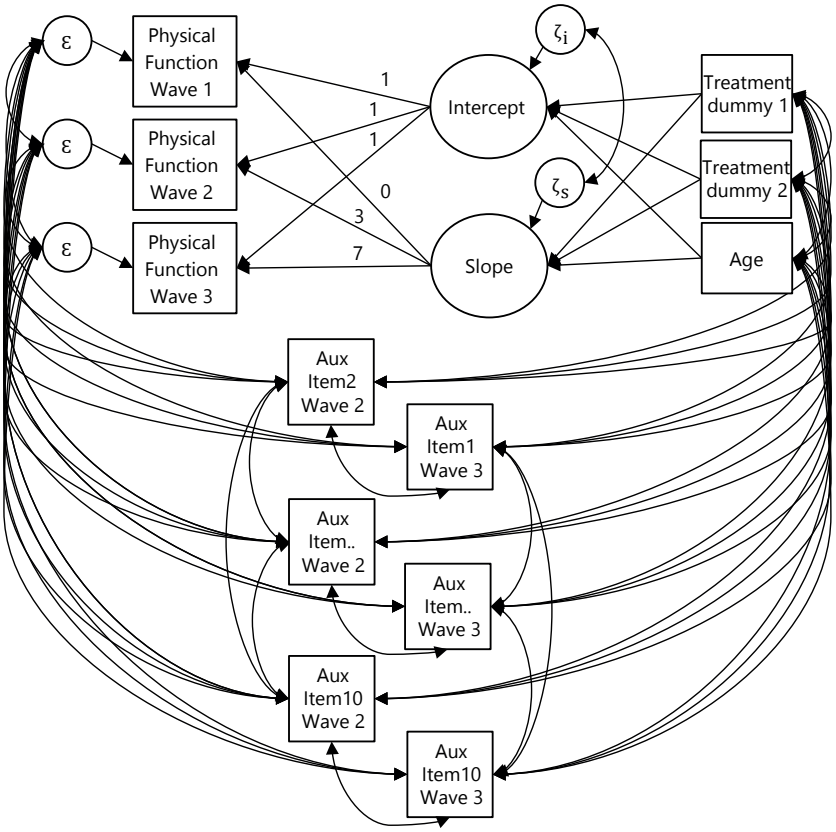


Figure 5.2. Path diagram of a latent growth model with the item scores as auxiliary variables. Treatment dummy 1 denotes physical therapy versus continued care by a general practitioner; Treatment dummy 2 denotes manual therapy versus continued care by a general practitioner.

Method 2: parcel of items as auxiliary variables

Including the items themselves is theoretically most ideal, because all available information is included in the model. However, when scales have many questionnaire items and more measurement waves are studied, including all item scores might be computationally difficult because the number of estimated parameters becomes very large. Perhaps to such an extent, that the sample size might not support the estimation of all the additional correlations in the model. In the context of multiple imputation with a large number of questionnaire items, Enders (2010) suggested to include a summary of the items into the imputation model. The items can be summarized by the average or sum of a subset of the items for each scale which would form a 'duplicate scale score'. We apply a similar logic to FIML growth modeling by using parcels of items that serve as auxiliary variables as a way to reduce model complexity. Because some of the items that contribute to the parcel might be missing, we take the average of the available items. Although averaging the available items is not a good standalone missing data method (Eekhout, et al., 2014; Schafer & Graham, 2002), our later simulations suggest that it works well as an item

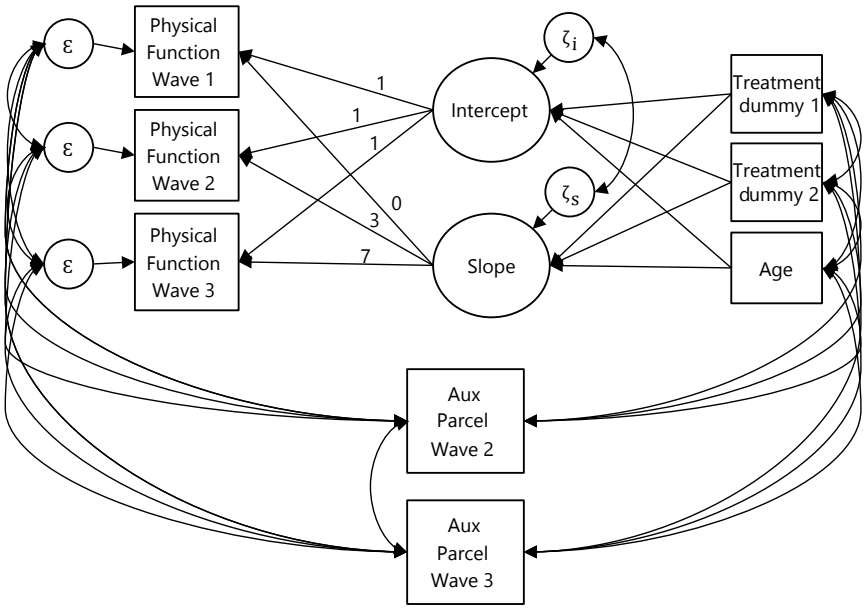


Figure 5.3. Path diagram of a latent growth model with parcel summaries of the items as auxiliary variables. Treatment dummy 1 denotes physical therapy versus continued care by a general practitioner; Treatment dummy 2 denotes manual therapy versus continued care by a general practitioner.

parcel summary. Our rationale for using it is that the average of available items can capture most of the available information in the observed items while dramatically reducing the number of parameters (Enders, 2008). Furthermore, the specification of the parcel summary would ultimately be the average of all but one item, though in some cases this specification might cause convergence problems, because of linear dependencies. In those cases a subset of fewer items can be used. For example, one possibility is to use two-thirds of the items with least missing scores; the goal is to include as much information as possible. In this study we explored the inclusion of a parcel of two-thirds of the items. This way the model incorporated a summary of the items to eliminate mathematical difficulties associated with estimating a large number of parameters. In Figure 5.3 the path diagram is depicted where this procedure is applied to the physical functioning study.

Illustrative analysis

As an illustration we used the proposed models to analyze the example dataset. The models were estimated by Mplus (Muthén & Muthén, 2007) and the syntaxes are provided in the Appendix 5.1. Firstly the data was analyzed by the growth model without auxiliary variables (Figure 5.1). The results from this model were used as a reference. The model was estimated with FIML with the scale scores missing when one or more items were missing. Secondly, the FIML model that includes the items as auxiliary variables was estimated as presented in Figure 5.2. In this example we included all but one item for wave two and wave three. Lastly we estimated the FIML model using a parcel summary of two-thirds of the item scores with least missing data as auxiliary variables (Figure 5.3). The parcel summary of wave two was the average of item 4 to item 10 and the parcel summary of wave three was the average of item 3 and item 5 to item 10.

The resulting parameter estimates of the example data are presented in Table 5.1 for each of the models. If we compare the results of the FIML model without the auxiliary variables to the FIML models that include the auxiliary information, the standard error estimates are smaller for the models that include auxiliary item information, which reflects a precision gain. The model that includes the item scores themselves had slightly smaller standard errors compared to the model that uses the parcel summaries, which is to be expected because the information about the items is more detailed as opposed to including a parcel summary and is therefore expected to result in more precision. However, the difference between the two auxiliary item methods is not very large and generally both methods improve the precision of the estimates as compared to not including auxiliary variables. Furthermore, the number of free parameters that have to be estimated in the models differs extensively. In the model with the items included as auxiliary variables the number of free parameters is

320, while the use of a parcel summary of the items as auxiliary variables decreased this number to 40. The finding that including item information has a precision gain, is consistent with previous studies of multiple imputation by Gottschall et al. (2012) and Eekhout et al. (2014).

Table 5.1.
Model estimates for the fitted models on the example data with the estimated standard errors between brackets.

Parameter	FIML without auxiliary variables (reference)	FIML with parcels as auxiliary variables	FIML with items as auxiliary variables
Number of free parameters	14	40	320
Intercept on age	0.021(0.022)	0.023(0.022)	0.020(0.022)
Intercept physical therapy[a]	-0.029(0.600)	-0.113(0.604)	-0.183(0.598)
Intercept manual therapy[b]	-0.503(0.600)	-0.536(0.604)	-0.598(0.598)
Slope on age	-0.003(0.004)	-0.004(0.003)	-0.003(0.003)
Slope on physical therapy[a]	0.076(0.116)	-0.013(0.087)	-0.013(0.084)
Slope on manual therapy[b]	0.191(0.117)	0.129(0.087)	0.137(0.084)
Intercept latent mean	24.684(1.115)	24.621(1.121)	24.846(1.110)
Slope latent mean	0.160(0.215)	0.304(0.162)	0.272(0.156)
Residual variances[c]	2.462(0.325)	2.433(0.301)	2.289(0.267)
Intercept residual variance	8.110(1.145)	8.337(1.147)	8.308(1.122)
Slope residual variance	0.082(0.028)	0.094(0.025)	0.102(0.024)
Intercept-slope covariance	-0.208(0.143)	-0.345(0.130)	-0.343(0.125)

Note: [a]Dummy that reflects physical therapy versus continued care by a general practitioner; [b]Dummy that reflects manual therapy versus continued care by a general practitioner; [c] model residual variances were constrained to be equal for each wave to eliminate fitting problems due to nonlinear growth.

These results suggest that including item-level auxiliary information (either the items themselves or a parcel summary of the items) can provide substantial precision gains in the form of reduced standard errors. The performance of the described methods was fully explored in a simulation study. In this simulation these methods were investigated in several longitudinal data situations with incomplete outcomes, caused by missing item scores. The proposed methods of including auxiliary variables as described above were compared to a FIML model without auxiliary variables and to a complete-case analysis.

Simulation study 1

The first simulation study was performed to test whether including the parcel summary as auxiliary variables worked equally well as including the item scores themselves under varying conditions. We compared these methods on a small number of conditions, because including the item scores themselves would likely be limited to conditions where a smaller number of free parameters has to be estimated. If the two methods would prove equivalent, then we would favor the approach

where parcel scores are used as auxiliary information, since this simpler approach requires fewer estimated parameters. Subsequently, we would expand exploring the performance of this method in a broader perspective of longitudinal data conditions.

Design

The items as auxiliary variables and the parcels as auxiliary variables methods were studied on simulated longitudinal data conditions, where six items were measured per scale for three measurement waves. The growth factors were predicted by a dichotomous treatment variable along with two normally distributed covariates. The covariates explained about 5% of the variance. The model was simulated to have an effect size around 0.25 at wave three. The effect size was defined as the difference between the model-implied means divided by the standard deviation at baseline. The path diagram of the model is similar to the diagram presented in Figure 5.1, with one extra covariate included. The loadings for the intercept factors were fixed at 1 and the loadings for the slope factor were set at linear time scores (i.e., 0, 1, 2). We generated population datasets containing 250,000 cases with some varying factors. Firstly, the inter-item correlations were varied between 0.6 and 0.8. Secondly, the number of incomplete items varied between 33% or 66% of the items; the first two items or the first four items were made incomplete respectively. Items measured at the first wave were left complete. The generation of missing data was related to the covariates in order to achieve a missing at random situation and incomplete items were missing for 15% of the subjects. Third, the distribution of the items was modified as normal or skewed, with a skewness level of -2.00. Items were generated as normally distributed variables and then threshold cutoffs were used to create discrete items with five categories. For the normal condition we used threshold cutoffs that would result in normally distributed items and for the skewed condition we used threshold cutoffs that generated a skewness of -2.00 in the items. This resulted in a total of eight population datasets generated in Mplus (Muthén & Muthén, 2007). From these population datasets we drew 1000 samples for each of three different sample size conditions (n=100, 500 or 1000). Sampling was performed in R (R Core Development Team, 2014). This resulted in a total of 24 longitudinal data conditions to compare the three methods on.

In the method where the item scores were used as auxiliary variables, all items except the first were included per measurement wave with incomplete items. The items from the complete wave were excluded to limit the number of free parameters to be estimated. The parcel scores were calculated by the average of two-thirds of the items for each wave. In order to include as much complete items as possible we excluded the first third of the items for the calculation of the parcel, because these were incomplete in all simulation conditions. The parcel of two-thirds of the items was

arbitrarily chosen to have a parcel that contains enough auxiliary item information without causing convergence problems due to linear dependencies. Accordingly, item 3 to item 6 were averaged for each measurement wave, irrespective of the percentage of incomplete items, to be the auxiliary parcel scores.

The performance of the two methods that included item information as auxiliary variables was compared to the growth model without auxiliary variables to evaluate the gain of including auxiliary variables when scale scores are incomplete simultaneously. This growth model was estimated using FIML and the scale scores were left incomplete if one of the items were missing. All models were estimated in Mplus (Muthén & Muthén, 2007).

In the presentation of the results we show the effects on all model parameters, however in the evaluation of the methods we focused on the model parameters that are most meaningful for researchers studying effects within a trial by using a latent growth model. These are the slope on treatment parameter which depicts the change in outcome for the treatment group and the latent mean of the slope factor which depicts the change in outcome for the control group. Since the difference between these two parameters, which depicts the benefit of treatment over the control condition, is of most interest, both are investigated. We also evaluated the estimation of the effect size parameter, which was calculated by dividing the mean difference between the treatment and control group at the end of the study to the standard deviation at baseline (Cohen, 1988). The effect size reflects the relative difference in change of outcome between the treatment and control group. (3)

The performance of the compared methods was evaluated through bias, mean squared error (MSE) and coverage of the confidence interval. These evaluation measures were calculated by comparing the estimates from the simulation samples to the parameter estimates from the reference populations without missing data. Bias was defined as the difference between average sample estimates within each condition and the population parameter from that condition. We report the standardized bias, which is calculated by

$$Standardized\ bias = \frac{\bar{\hat{\beta}} - \bar{\beta}_c}{sd(\beta_c)}\ 100\% \quad (1)$$

where $\bar{\hat{\beta}}$ is the average sample estimate within a condition for the applied method, β the population parameter from that condition, and $sd(\hat{\beta})$ the standard deviation of the sample estimates for the FIML method without auxiliary variables (Collins, et al., 2001). We use the standard deviation of one method to hold the sampling variance constant and that way to be able to really compare the bias across methods. The MSE is a measure of precision and incorporates both the bias and the variability of the estimates. MSE was calculated by squaring the difference between the average

sample estimate $(\hat{\beta})$ and the population parameter (β) in a condition:

$$MSE = \left(\bar{\hat{\beta}} - \bar{\beta}_c\right)^2 \quad (2)$$

Precision of parameter estimates is related to the sample size (Cohen, 1988). For easier interpretation we also report the MSE ratio for the MSE of the FIML model without auxiliary variables and the FIML model with auxiliary information, as follows

$$MSE_{ratio} = \frac{MSE_{model\ without\ auxiliary\ variables}}{MSE_{model\ with\ auxiliary\ item\ information}}(100\%) \quad (3)$$

Because the MSE is inversely related to sample size, the MSE_{ratio} represents the proportional increase in the sample size that is required for the model without auxiliary variables to achieve the same precision as the models with auxiliary item information. For example an MSE_{ratio} of 125 indicates that the sample size should be increased by 25% to achieve the same level of precision for the model without auxiliary variables as the model with the auxiliary item information. Coverage was evaluated by the proportion of replications in each cell, that the confidence interval of the sample estimate included the population parameter. The coverage of the confidence intervals should be approximately equal to the nominal confidence interval rate, in our study 95%. Coverages above the 95% rate indicate that the method might be too conservative i.e., yield standard errors that are too large, and a lower coverage rate suggests higher than expected type I error i.e., yield standard errors that are too small (Burton & Altman, 2004).

For each condition in the simulation study we checked whether there were important differences on each of the performance measures. This was done by doing a factorial analysis of variance to explore to what extent the performance measures differed for the conditions. We looked at the effect sizes of interactions between the methods and other conditions. If the effect size of the condition was larger than the threshold for a small effect size which is 0.1, then we would examine to what extent the condition affected the performance measure (Cohen, 1988). Only the conditions that actually affected the performance of the methods are reported in the results section.

Results

In this simulation study, we did not find significant differences between including the individual items as auxiliary variables and including the parcel scores in any of the three performance measures for any of the data conditions. Bias was generally small for all three methods; in all conditions standardized bias was smaller than 5%, so we have omitted these results. All of the methods showed good coverage in the conditions of simulation study 1 (data not shown). However, both methods that

included auxiliary variables performed better than the FIML model without auxiliary variables with respect to the MSE. Furthermore, the parameters estimated by the FIML model without auxiliary variables had a larger MSE when more items were incomplete than in the condition where less items were incomplete. In Table 5.2 the MSE_{ratios} are presented for the FIML method without auxiliary variables relative to (1) the model that includes parcel scores and (2) the model that includes the item scores. The MSE_{ratio} results in Table 5.2 are split for the two conditions of the number of incomplete items, which are 33% and 66%, but all other conditions are joined within these conditions. For example, the MSE_{ratios} of the slope on treatment parameter when 66% of the items are incomplete in the FIML model without auxiliary variables relative to the model including parcel scores is 141.36. This ratio suggests that the sample size needs to be increased by 41% for the model without auxiliary information to achieve the same power at the model with the auxiliary information included, regardless of the correlation between the items, the skewness of the items, or the original sample size. The increase in the MSE_{ratio} when 33% versus 66% of the items are incomplete suggests that there is a large effect of the percentage of incomplete items on results from the method where no auxiliary variables are used. The difference between the two auxiliary variable approaches did not vary as a function of the percentage incomplete items.

Table 5.2.
MSE_{ratio} of FIML model without auxiliary variables versus FIML models with auxiliary item information by parameter estimate separated for amount of incomplete items in simulation study 1.

	No auxiliary versus parcels as auxiliary variables		No auxiliary versus items as auxiliary variables	
	Amount of incomplete items		Amount of incomplete items	
	33%	66%	33%	66%
Intercept on covariate1	103.36	103.68	103.64	104.43
Intercept on covariate 2	103.45	102.91	103.54	103.31
Intercept on treatment	102.59	103.59	102.64	103.83
Slope on covariate 1	134.18	170.99	135.89	177.34
Slope on covariate 2	136.12	166.85	137.23	174.37
Slope on treatment	121.60	141.36	122.21	144.74
Intercept latent mean	103.30	103.73	103.44	104.27
Slope latent mean	125.95	148.29	126.84	153.29
Residual variance scale wave 1	136.53	163.47	138.25	172.09
Residual variance scale wave 2	138.22	143.77	141.82	182.47
Residual variance scale wave 3	153.80	189.95	157.18	213.77
Intercept residual variance	139.53	168.47	142.13	178.55
Slope residual variance	147.09	180.85	149.77	193.89
Intercept-slope covariance	146.47	175.69	149.19	187.79
Effect size	126.03	148.65	127.12	153.63

Conclusions

The results showed that both the model where item scores are included as auxiliary variables and the method where parcel summaries of the item scores were used performed equally well. Both methods showed more precision than the FIML model without auxiliary variables; however the differences in bias are small. This indicates that there is primarily a power advantage for using a model with auxiliary item information. Furthermore we can conclude that a summary of the items (i.e., the parcel summary score) contains enough information to improve the precision of estimates effectively. As previously mentioned the method where the item scores are included as auxiliary variables requires a significant amount of computational effort and might sometimes fail to converge. The parcel summary method requires far fewer calculations and is therefore preferred.

Simulation study 2

The first simulation showed that the parcel summary method improves estimates compared to not including any auxiliary information. However this first simulation was only performed on a limited number of longitudinal data conditions; the simulated data had three measurement waves with six items per scale. The second simulation study was conducted to investigate the performance of the parcel scores as auxiliary variables with varying numbers of waves, number of items per scale, inter-item correlations, and deviation from a normal distribution. In these conditions the number of repeated measures can increase to seven and the number of items per scale to 18.

Design

Longitudinal data situations were simulated by varying six data features. The first four varying aspects were (1) the number of items per scale (6, 12, or 18 items), (2) the distribution of the items (normal or skewed at a level of -2), (3) the inter-item correlation (r=0.6 or 0.8), and (4) the amount of incomplete items (33% or 66% of the items per scale). The items from the population data were again generated as normally distributed continuous variables. In order to modify the distribution of the items as normal or skewed, threshold cutoffs were used to create discrete items with five categories. For the normal condition we used threshold cutoffs that would result in normally distributed items and for the skewed condition we used threshold cutoffs that generated a skewness of -2.00 in the items. According to the four varying aspects settings we generated 24 different population datasets of 250,000 cases.

The population datasets were generated for seven measurement waves and included two time-invariant continuous normally distributed covariates and a binary

treatment variable. The path diagram of the population model for simulation study 2 is similar to the diagram presented in Figure 5.1, but with one extra covariate and 4 more measurement waves. The loadings for the intercept factors were fixed at 1 and the loadings for the slope factor were set at the time scores (e.g., from 0 to 6 for the 7 wave condition). The effect size of the model was simulated to be 0.50 at wave 7 and the covariates explained about 5% of the variance in the model. Items were made missing at the item-level of the repeated measurements, so the baseline scale was complete. Items were made missing independently from the other items in the scale but related to the covariates, so that the missing data was modeled to be missing at random. Each incomplete item was made missing for 15% of the subjects. The population data was generated in Mplus (Muthén & Muthén, 2007). From each of these populations, samples were drawn with a varying sample size (n=100, 500 or 1000) and two conditions of repeated measures (3 or 7 measurement waves). The population data was created for seven waves; for the three-wave condition we only selected the first three waves. This resulted in a total of 144 simulated conditions; for each condition 1000 samples were drawn. The sampling was performed in R version 2.15.3 (R Core Development Team, 2014).

The parcel scores were calculated by averaging over two-thirds of the items for each measurement wave. For the condition with 6 items per scale, item 3 to item 6 were averaged; for the 12 item scale, item 5 to item 12 were averaged; and for the 18 item scale the average over item 7 to item 18 was taken. The method that uses parcel scores as auxiliary variables was compared with the model without auxiliary variables estimated with FIML and with a complete-case analysis (CCA). In the CCA only complete cases were included into the analysis. This could result in very small remaining samples, because any missing on an item score would result in the case to be excluded from the analysis. The CCA method was included in the simulation in order show the gain of using a FIML method in our studied data situations. All models were estimated in Mplus (Muthén & Muthén, 2007).

The performance of the compared methods was evaluated in the same way as simulation study 1. The methods were compared through bias, MSE, MSE_{ratio} (Equation 1-3) and coverage. We used the population data results without missing data as the true parameters. The reference for the MSE_{ratio} was the method with the parcel summary as auxiliary variable. Again, we focused on the model parameters that are most meaningful for researchers studying effects within a clinical trial by using a latent growth model i.e., the slope on treatment parameter, the slope latent mean and the estimation of the effect size parameter, but all model parameters are presented in Table 5.3. To explore the simulation results, we performed a factorial analysis of variance to identify differences in performance for the studied methods in the simulated conditions. In the results section we reported the conditions that

indicated differences in method performance. The main results are described in text; tabular presentations are omitted in order to save space. A full tabulation of the simulation results is available by the first author upon request.

Results

The bias, presented as the standardized bias in Table 5.3, was generally small for all methods. However for the CCA there are some estimates of the important parameters that have larger bias values. For example 6.6% for the intercept latent mean and even 18.1% for the effect size. The standardized bias presented in Table 5.3 is the average bias over all simulated conditions.

Table 5.3.
Standardized bias estimates for the parameter estimates of the methods investigated in simulation study 2 over all conditions.

Parameter	Complete-case analysis	FIML without auxiliary variables	FIML with parcels as auxiliary variables
Intercept on covariate1	-4.0141	-3.7495	-3.0084
Intercept on covariate 2	-2.3881	-4.1857	-3.5789
Intercept on treatment	-1.8530	-7.7731	-1.8055
Slope on covariate 1	-0.9642	-3.5099	-5.1619
Slope on covariate 2	4.2501	2.2215	0.9375
Slope on treatment	-1.6486	-1.6526	-2.7890
Intercept latent mean	-3.6323	-1.4568	-2.2926
Slope latent mean	6.5937	-0.8486	-0.0591
Residual variance scale wave 1	-8.4716	-5.0961	-3.7795
Residual variance scale wave 2	-8.6331	-6.8420	-0.4292
Residual variance scale wave 3	-6.7905	-2.7752	0.3884
Residual variance scale wave 4	-4.4338	-0.8554	-0.2022
Residual variance scale wave 5	-12.1554	9.1675	18.2329
Residual variance scale wave 6	2.8426	0.2885	14.9863
Residual variance scale wave 7	2.0509	-12.0206	-1.4746
Intercept residual variance	-21.6582	-9.2685	-6.4967
Slope residual variance	-5.3552	-3.0196	-2.7799
Intercept-slope covariance	14.0178	2.4691	1.7598
Effect size	18.1048	7.9380	3.7600

In the top-panel of Figure 5.4 the MSE_{ratio} of the FIML model without auxiliary variables relative to the method with auxiliary parcel scores for an increasing number of items per scale on the x-axis is presented. The results in the graph are averaged over all other conditions. The necessary increase in sample size grew larger when scales contained more items. The effect of the increase of items per scale was even larger for the CCA than for the FIML model without auxiliary variables. For example for the slope on treatment parameter, the necessary percentage of sample size increase for the FIML model without auxiliary variables ranges from an additional

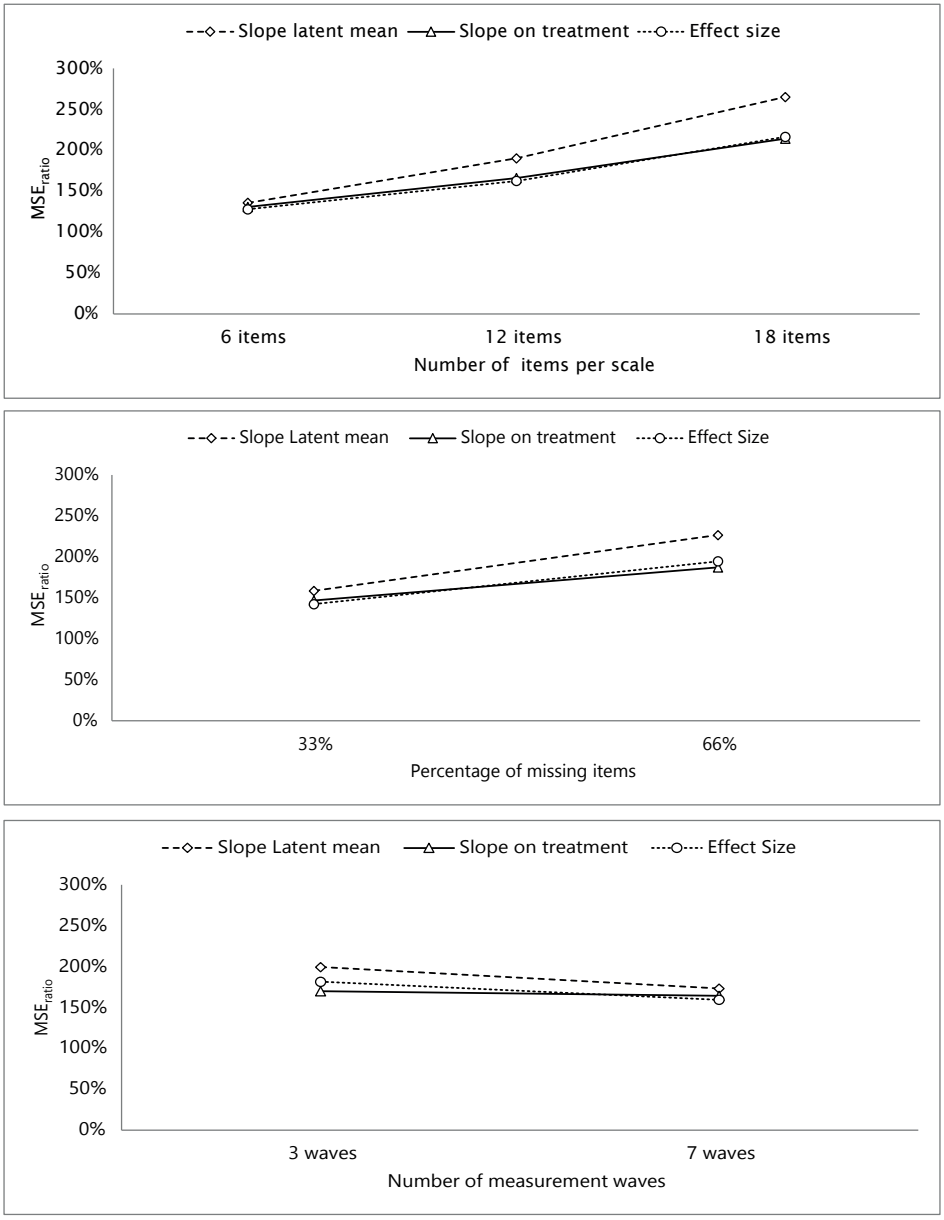


Figure 5.4. MSE_{ratio} for FIML without auxiliary variables relative to the FIML model that includes parcel scores as auxiliary variables for the slope on treatment, slope latent mean and effect size.

31% to 114%, when the number of items per scale was 6 items and 18 items respectively. For the CCA the required percentage of sample size increase was much larger; from the slope on treatment parameter the percentage ranged from 229% to 957%, meaning that the sample size should be multiplied by almost a factor ten to reach the same precision. For both the FIML method without auxiliary variables and the CCA, the increase in MSE decreased when the sample size became larger. For the parcel scores as auxiliary variables we observed the adverse effect on MSE. The MSE decreased slightly when the scale length increased, this effect was smaller when the sample size increased. Moreover, a larger percentage of incomplete items per scale was related to an increased MSE for both CCA and for the FIML model without auxiliary variables. For these two methods this negative effect became larger when the scale length increased. In the middle-panel of Figure 5.4 the MSE_{ratio} is depicted comparing the MSE of the FIML model without auxiliary variables relative to the model with the auxiliary parcel scores. For example for the slope latent mean parameter the necessary increase of sample size was 58% and 127% for the FIML model without auxiliary variables when 33% and 66% of the items were incomplete respectively. For CCA this increase was 651% and 2307% when 33% and 66% of the items contained missings correspondingly. The percentage of items with missings did not affect the MSE for the model that included the parcel scores as auxiliary variables. For the slope latent mean we also found that for a CCA a shorter scale combined with more measurement waves increased the precision, but for larger scales more waves increased the MSE and therefore decreased precision. For the parcel scores as auxiliary variables method more waves generally increased the precision of the slope on treatment parameter and the slope latent mean parameter, irrespective of the scale length. For the FIML model without auxiliary variables more waves increased the precision as well, but overall longer scales decreased precision. For the effect size parameter we found that the number of waves mostly affected the MSE of CCA. The other two methods remained relatively stable with respect to the MSE when more waves were included in the study. The bottom-panel of Figure 5.4 depicts the effect of the number of waves on the MSE_{ratio}, which shows that the MSE_{ratio} of the FIML model without auxiliary variables and the model with parcel scores remained stable for the number of repeated measurements conditions and is even slightly decreasing when the number of waves is larger. The MSE_{ratio} for CCA is much larger and largely increases when more waves are studied.

For all parameters of interest the coverage of the method that uses the parcel scores as auxiliary variables was stable over all conditions and closely around 95%. Table 5.4 presents the average coverage values over all conditions split for the number of items per scale. For the CCA method the coverage was on the lower side and decreased more when the scale length became larger or more waves were

included for the slope on treatment parameter and the slope latent mean parameter. The growth model without auxiliary variables had a good coverage in general.

Table 5.4.
Average coverage rates per parameter estimate separated for the different scale length conditions for each investigated method

	Slope on treatment			Slope latent mean			Effect size		
	CCA	No Aux	Parcel Aux	CCA	No Aux	Parcel Aux	CCA	No Aux	Parcel Aux
6 items	94.1%	94.4%	94.8%	94.1%	94.9%	95.1%	94.5%	94.6%	94.8%
12 items	93.0%	93.7%	94.6%	93.0%	94.8%	94.9%	93.7%	94.1%	94.8%
18 items	91.7%	93.1%	94.8%	91.5%	94.5%	94.9%	92.9%	93.7%	94.8%

Note: CCA is the complete-case analysis; No Aux is the FIML model without auxiliary variables; Parcel aux is the model where parcel scores are used as auxiliary variables.

Conclusions

For the second simulation study we can conclude that including a parcel summary of items as auxiliary variables performed better than the FIML model without auxiliary variables and especially better than CCA. The largest advantage is in precision, reflected in the MSE_{ratio} which indicates the necessary increase in sample size to reach the level of precision in the parcel summary method. Both the FIML model where no auxiliary variables are included and the CCA required a substantial increase in sample size. Essentially, the method that includes the parcel summaries of items seems unaffected by the percentage of incomplete items, while both other two methods decrease in performance when the percentage of incomplete items was increasing. Furthermore, even when 33% of the items were missing, the model that includes the parcel summary scores performs superior to the FIML model without auxiliary variables, and especially much better than a CCA. Moreover, the number of measurement waves and the scale length had an adverse effect on precision compared to CCA. The increase in measurement waves or in the number of items per scale caused CCA to perform increasingly worse, while the FIML model that includes the parcel scores as auxiliary variables only gained in precision.

Furthermore, the MSE_{ratio} of the FIML model without auxiliary variables relative to the model with parcel summaries somewhat decreased when more waves were measured. This can be interpreted as that the precision of the FIML model without auxiliary variables came slightly closer to the precision of the parcel summary method when more waves were measured. However, the MSE_{ratio} was larger than 100% in all conditions of the simulation and is therefore indicating that the precision of the model with parcel summary scores is generally better. Overall, sample sizes should practically be doubled, as indicated by the observed MSE_{ratios} of 200%, for the FIML model without auxiliary variables to achieve the same precision as the model that includes the item information.

Discussion

In this study we proposed new methods for dealing with incomplete scale scores that result from missing item responses. Previous studies have shown that handling missing data at the item-level can provide substantial improvements in power (Eekhout, et al., 2014; Gottschall, et al., 2012). However, these studies were focused on multiple imputation. In longitudinal studies it might be feasible to use a model that measures the development over time and handles the missing data at the same time, as a latent growth model. In these models the item information is usually not included when the scale scores are the outcomes of interest. Thus, the purpose of this study was to propose two novel methods that include the items score information as auxiliary variables, while treating the scale scores missing in a latent growth model. That way the item information is incorporated in the model while leaving the scale scores as the focus of the analysis.

In this study we found that including auxiliary item information into the model when item scores are missing improves results compared to not including this information. The main advantage is in the precision of model coefficient estimates. Previous studies showed that FIML are good methods to handle missing data when missings are in the outcome (Enders, 2011; Enders & Bandalos, 2001). This study showed that in the case of incomplete scale scores that result from missing item scores, precision can be hugely improved by including the item information as auxiliary variables in the model. Theoretically the most improvement was expected when the item scores itself were included as auxiliary variables. This method would incorporate the maximum amount of information in the estimation process and therefore achieve an optimal amount of power. Furthermore, in an asymptotically equivalent method, multiple imputation, it was already demonstrated that using the items scores in the model would estimate the most optimal results for a scale score analysis (Eekhout, et al., 2014; Gottschall, et al., 2012). Both studies showed that the gain of applying multiple imputation to the item scores is in precision, ergo smaller MSE and standard error estimates. Our methods aimed to achieve a similar optimal method for FIML methods. Even though this is not as straight forward in FIML as in multiple imputation, where it would come down to including the auxiliary variables in the imputation model, our method achieved the same success. Including the item score information as auxiliary variables in the estimation process yielded more optimal results than not including this information this way.

In our first simulation study we found that including the item scores as auxiliary variables performed comparable to including parcel summaries of items as auxiliary items. For that reason we conducted the second simulation study with the parcel summary of items method to be able to study more complex longitudinal conditions. In these conditions computation with the inclusion of the item scores itself would

be too demanding and often would fail to converge. Our second simulation study showed that including parcel summaries of items as auxiliary variables improves power and precision in model coefficient estimates compared to not including these variables. Especially in the results of the MSE ratios comparing precision of the model without auxiliary variables with the model including parcel summaries are really convincing. Sample sizes should nearly be doubled to achieve the same level of precision. So even though including a parcel is theoretically not most optimal, our study showed that it is a huge improvement to not including item information into the model.

Benefits and limitations

When the number of items per scale increased, the precision decreased for the FIML model without auxiliary variables and CCA. Including the parcel summary of items as auxiliary variables seems to diminish that effect. The decrease in precision when the scale length was larger for the other two methods can be explained in relation to item missings. Larger scale lengths are related to a larger absolute number of missing cases. For example if one-thirds of the item scores are missing when the scale length is 6 items, two items would contain missings. For each of these items, 15% of the subjects have a missing score. In the condition where the scale length is 18 items, 6 items would have missing scores, each of them for 15% of the subjects dependent on the covariates. Though in more realistic situations, missing item scores are often clustered within subjects, in our way of simulating the missings chances are that in the 18 items condition, more subjects have missing data points than in the 6 items condition and consequently the total score calculation is impaired for more subjects. For that reason, the CCA, which only includes those subjects that have complete data on all time-points, is heavily affected by the scale length. The growth model with FIML estimation uses for incomplete cases on wave 2 the data from wave 1 and 3 to obtain the most likely estimate. For that reason, the included information compared to CCA is larger, though in the FIML model still many scale scores will turn out incomplete due to many subjects with missing item scores. And if a scale has any missing item, the scale score would turn out missing. In the FIML model with auxiliary variables, the information of the available items is used as well. Accordingly, the amount of information used has hugely increased and is even higher with more items in the scale, which relates to better estimates and therefore we observe the adverse effect for the method that uses auxiliary variables.

It is well-known that for missing outcome data in a longitudinal study, FIML estimation outperforms a CCA. Since FIML methods use all available time-points to estimate the most likely parameter estimate, while CCA ignores cases with any missing time-point. Along with the previously described relation between our scale

length condition and the CCA performance it might have been arguable not to include CCA as a comparison method into our simulation. Nevertheless, we wanted to include this method to see how this method would perform in conditions were a small amount of items were missing compared to our proposed method on the one hand and also to show the gain of using a FIML method in our studied data situations. The last argument might seem rather trivial, but a previous review showed that complete-case analysis is still the most widely used method to handle missing data (Eekhout, et al., 2012).

In this simulation study we generated data in order to create a wide variety of data situations. Though we do realize that some of these situations are quite extreme (e.g., four out of six items 66% incomplete), we argue that if our methods perform well in these situations, they will hold in less extreme situations as well. Furthermore, we only fully investigated one way to calculate a parcel summary, which is to average over two-thirds of the most complete items. We did a small simulation to investigate the possibilities of different compositions of the parcel summary by including information from more items (e.g., all but one items) or less items (e.g., half of the items). We found that the most optimal parcel should include maximum information, however contain enough noise not to cause multicollinearity problems. For example, including all but one item would in some occasions result in parcel scores too similar to the scale score and cause estimation problems. Consequently, in order to have a method that would work in many longitudinal situations we chose to study the currently presented parcel summary.

The latent growth model we present uses a dummy predictor to indicate the treatment groups in the model. This model assumes a common covariance structure for the auxiliary variables. Instead of a dummy predictor, the model could also be specified as a multiple group model, when the substantive interest involves a comparison of change. In a multiple group model, it is possible to allow the auxiliary variable correlations to freely vary across groups. The decision to constrain or freely estimate these covariances is related to measurement invariance. In the situation that the scale scores possess measurement invariance, it is expected that the covariance structure would be common to both groups. A lack of between-group invariance prompts to freely estimate at least some of the auxiliary variable correlations. At this point it is unclear whether mis-specifying the auxiliary part of the model would affect the growth model estimates, but future research could investigate this issue.

Concluding remarks

In general we recommend including a parcel summary of the items in the auxiliary part of the latent growth model when incomplete scale scores result from missing item scores. This study shows that the parcel summary of the items improved the

precision of the estimates over not including auxiliary information. Furthermore, the inclusion of a parcel summary is an efficient method that does not over-complicate model estimations.

References

Bell, M. L., & Fairclough, D. L. (2013). Practical and statistical issues in missing data for longitudinal patient reported outcomes. Stat Methods Med Res, 23(5), 440-459.

Bernaards, C. A., & Sijtsma, K. (2000). Influence of Imputation and EM Methods on Factor Analysis when Item Nonresponse in Questionnaire Data is Nonignorable. Multivariate Behavioral Research, 35(3), 321-364.

Burton, A., & Altman, D. G. (2004). Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. British Journal of Cancer, 91(1), 4-8.

van Buuren, S. (2012). Flexible Imputation of Missing data. New York: Chapman & Hall/ CRC.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological Methods, 6(4), 330-351.

Eekhout, I., de Boer, R. M., Twisk, J. W., de Vet, H. C., & Heymans, M. W. (2012). Missing data: a systematic review of how they are reported and handled. Epidemiology, 23(5), 729-732.

Eekhout, I., De Vet, H. C. W., Twisk, J. W. R., Brand, J. P. L., De Boer, M. R., & Heymans, M. W. (2014). Missing data in a multi-item instrument were best handled by multiple imputation at the item score level. Journal of Clinical Epidemiology, 67(3), 335-342.

Enders, C. K. (2008). A Note on the Use of Missing Auxiliary Variables in Full Information Maximum Likelihood-Based Structural Equation Models. Structural Equation Modeling: A Multidisciplinary Journal, 15(3), 434-448.

Enders, C. K. (2010). Applied Missing Data Analysis. New York, NY: The Guilford Press.

Enders, C. K. (2011). Analyzing longitudinal data with missing values. Rehabilitation Psychology, 56(4), 267-288.

Enders, C. K., & Bandalos, D. L. (2001). The Relative Performance of Full Information Maximum Likelihood Estimation for Missing Data in Structural Equation Models. Structural Equation Modeling: A Multidisciplinary Journal, 8(3), 430-457.

Fayers, P. M., Curran, D., & Machin, D. (1998). Incomplete quality of life data in randomized trials: missing items. Stat.Med., 17(5-7), 679-696.

Gottschall, A. C., West, S. G., & Enders, C. K. (2012). A Comparison of Item-Level and Scale-Level Multiple Imputation for Questionnaire Batteries. Multivariate Behavioral Research, 47(1), 1-25.

Graham, J. W. (2003). Adding Missing-Data-Relevant Variables to FIML-Based Structural Equation Models. Structural Equation Modeling: A Multidisciplinary Journal, 10(1), 80-100.

Hawthorne, G., & Elliott, P. (2005). Imputing cross-sectional missing data: comparison of common techniques. Aust.N.Z.J.Psychiatry, 39(7), 583-590.

Karahalios, A., Baglietto, L., Carlin, J., English, D., & Simpson, J. (2012). A review of the reporting and handling of missing data in cohort studies with repeated assessment of exposure measures. BMC Med Res Methodol., 12(96), 1-10.

Little, R. J., & Rubin, D. B. (2002). Statistical Analysis with Missing Data (Second Edition ed.). Hoboken, NJ: John Wiley & Sons.

Muthén, L. K., & Muthén, B. O. (2007). Mplus User's Guide (Seventh Edition). Los Angeles, CA: Muthén & Muthén.

Raykov, T., Schneider, B. C., Marcoulides, G. A., & Lichtenberg, P. A. (2014). Examining Measure Correlations With Incomplete Data Sets. Structural Equation Modeling: A Multidisciplinary Journal, 21(2), 318-324.

Rubin, D. B. (1976). Inference and missing data. Biometrika, 63(3), 581-592.

Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: J. Wiley & Sons.

Schafer, J. L. (1997). Analysis of Incomplete Multivariate Data. London, UK: Chapman & Hall.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. Psychol. Methods., 7(2), 147-177.

R Core Development Team (2014). R: A language and environment for statistical computing. Vienna, Austria.

Ware, J. E., Jr., Kosinski, M., & Keller, S. D. (1994). SF-36 Physical and Mental Health Summary Scales: A User's Manual. Boston, MA: Health Assessment Lab.

Appendix 5.1|Mplus syntaxes

Mplus syntax for the growth model without auxiliary variables

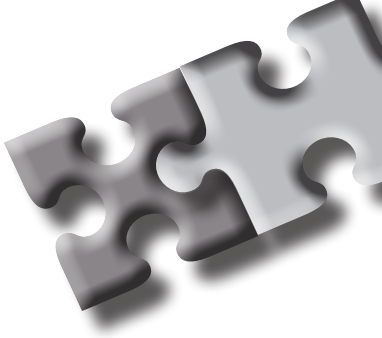
```
Mplus VERSION 7.11
MUTHEN & MUTHEN

INPUT INSTRUCTIONS

data:
file ='C:\filelocation\dataset_missing.dat';
variable:
names =
y1_1 - y1_10
y2_1 - y2_10
y3_1 - y3_10
tx age;
usevariables = age dummy1 dummy2 scale1 scale2 scale3;
missing = all(-9);
define:
!define dummies for the treatment groups;
IF (tx==1) THEN dummy1=0;
IF (tx==2) THEN dummy1=1;
IF (tx==3) THEN dummy1=0;
IF (tx==1) THEN dummy2=1;
IF (tx==2) THEN dummy2=0;
IF (tx==3) THEN dummy2=0;
! calculation of scale scores;
scale1 = sum(y1_1-y1_10);
scale2 = sum(y2_1-y2_10);
scale3 = sum(y3_1-y3_10);
model:
i s | scale1@0 scale2@3 scale3@7;
i on age
dummy1 (iontx1)
dummy2 (iontx2);
s on age
dummy1(sontx1)
dummy2(sontx2);
i with s;
[i] (i);
[s] (s);
i (ivar);
s;
scale1 - scale3 (resvar);
```

Mplus syntax for the growth model with item scores as auxiliary variables

```
Mplus VERSION 7.11
MUTHEN & MUTHEN

INPUT INSTRUCTIONS

data:
file ='C:\filelocation\dataset_missing.dat';
variable:
names =
y1_1 - y1_10
y2_1 - y2_10
y3_1 - y3_10
tx age;
usevariables = age dummy1 dummy2 scale1 - scale3;
missing = all(-9);
! including the items as auxiliary variables;
auxiliary = (m) y2_2-y2_10
y3_1-y3_3 y3_5-y3_10;
define:
!define dummies for the treatment groups;
IF (tx==1) THEN dummy1=0;
IF (tx==2) THEN dummy1=1;
IF (tx==3) THEN dummy1=0;
IF (tx==1) THEN dummy2=1;
IF (tx==2) THEN dummy2=0;
IF (tx==3) THEN dummy2=0;
! calculation of scale scores;
scale1 = sum(y1_1-y1_10);
scale2 = sum(y2_1-y2_10);
scale3 = sum(y3_1-y3_10);
model:
i s | scale1@0 scale2@3 scale3@7;
i on age
dummy1 (iontx1)
dummy2 (iontx2);
s on age
dummy1(sontx1)
dummy2(sontx2);
i with s;
[i] (i);
[s] (s);
i (ivar);
s;
scale1 - scale3(resvar);
```

Mplus syntax for the growth model with the parcel scores as auxiliary variables

```
Mplus VERSION 7.11
MUTHEN & MUTHEN

INPUT INSTRUCTIONS

data:
file ='C:\filelocation\dataset_missing.dat';
variable:
names =
y1_1 - y1_10
y2_1 - y2_10
y3_1 - y3_10
tx age;
usevariables = age dummy1 dummy2 scale1-scale3 parcel2-parcel3;
missing = all(-9);
! including the parcel scores as auxiliary variables;
auxiliary = (m) parcel2 - parcel3;
define:
!define dummies for the treatment groups;
IF (tx==1) THEN dummy1=0;
IF (tx==2) THEN dummy1=1;
IF (tx==3) THEN dummy1=0;
IF (tx==1) THEN dummy2=1;
IF (tx==2) THEN dummy2=0;
IF (tx==3) THEN dummy2=0;
! calculation of scale scores;
scale1 = sum(y1_1-y1_10);
scale2 = sum(y2_1-y2_10);
scale3 = sum(y3_1-y3_10);
! calculation of the parcel summary scores;
parcel2 = mean(y2_4-y2_10);
parcel3 = mean(y3_3 y3_5-y3_10);
model:
i s | scale1@0 scale2@3 scale3@7;
i on age
dummy1 (iontx1)
dummy2 (iontx2);
s on age
dummy1(sontx1)
dummy2(sontx2);
i with s;
[i] (i);
[s] (s);
i (ivar);
s;
scale1 - scale3 (resvar);
```



Chapter 6

Including auxiliary item information to handle missing questionnaire data in two longitudinal data examples

Under review: Eekhout, I., Enders, C.K., Twisk, J.W.R., de Boer, M.R., de Vet, H.C.W., & Heymans, M.W. Including auxiliary item information to handle missing questionnaire data in two longitudinal data examples. Journal of Clinical Epidemiology.

Abstract

Previous studies show that missing values in multi-item questionnaires can best be handled at item score level. The aim of this study is to demonstrate two novel methods for dealing with incomplete item scores in outcome variables in longitudinal studies. The performance of these methods was previously examined in a simulation study. The two methods incorporate item information at the background when simultaneously the study outcomes are estimated. The investigated methods include the item scores or a summary of a parcel of available item scores as auxiliary variables, while using the total score of the multi-item questionnaire as the main focus of the analysis in a latent growth model. That way the items help estimating the incomplete information of the total scores. The methods are demonstrated in two empirical datasets. Including the item information results in more precise outcomes in terms of regression coefficient estimates and standard errors, compared to not including item information in the analysis. The inclusion of a parcel summary is an efficient method that does not over-complicate longitudinal growth estimates. Therefore it is recommended in situations where multi-item questionnaires are used as outcome measure in longitudinal clinical studies with incomplete scores due to missing item scores.

Keywords: missing data, longitudinal data, multi-item questionnaire, auxiliary variables, full information maximum likelihood, methods, latent growth modeling, structural equation modeling

Introduction

Many medical and epidemiological longitudinal studies use patient-reported outcomes such as quality of life as the main focus of their analyses. These patient-reported outcomes are often repeatedly measured by a multi-item questionnaire. The item scores of the questionnaire are summed or averaged to a total score to represent the outcome of interest. In case respondents do not fill out all the questions in a multi-item questionnaire, the calculation of the total scores is impaired. As a solution, manuals of multi-item questionnaires often advise to average over the available items (e.g., (Bracken & Howell, 2004; Lambert, Lunnen, Umphress, Hansen, & Burlingame, 1994)), otherwise known as person mean imputation. Averaging over the available items is algebraically identical to substituting a person's mean item response. This solution can result in biased analysis results, especially when data are not missing completely at random (MCAR) (Eekhout et al., 2014; Gottschall, West, & Enders, 2012). Another option for handling missing data values is to apply a complete-case analysis. In that method only respondents that have all item scores observed are included in the analysis. This method only results in unbiased analyses when data are MCAR. A complete-case analysis always results in a decreased sample size, so power will be suboptimal in all situations. Nevertheless, this method is most often applied in epidemiological studies (Eekhout, de Boer, Twisk, de Vet, & Heymans, 2012).

More advanced methods to handle missing data are multiple imputation or full information maximum likelihood (FIML). Both methods use all observed data in the analyses. In multiple imputation, the missing values are replaced by imputed values. A regression model estimates predicted scores for the incomplete values and random error, drawn from a normal distribution around the estimated value, is added to the predicted score to account for uncertainty around the imputed values. This imputation process is repeated multiple times resulting in multiple imputed datasets. Subsequently, the data analysis is performed on each of these imputed datasets. The multiple results from these datasets are pooled into one final analysis result (Rubin, 1987; Schafer, 1997; van Buuren, 2012). In FIML, missing values are not replaced or imputed; instead all available data are used to estimate the population parameters with the highest likelihood of producing the sample data. Both multiple imputation and FIML perform well when the probability of missing data is related to other variables in the data, which is known as missing at random (MAR) (Rubin, 1976). Furthermore, with these techniques model estimations are generally unbiased and without loss of power.

In a multi-item questionnaire total scores may be missing because of missing item scores. In that case, there are two main approaches to handle the missing data. Missing data can be handled at the item level or at the total score level of the multi-item questionnaire. The missings are handled at the item level when a missing data

method is applied to the incomplete item scores first and then the total scores are calculated (e.g., by summing imputed item responses) and used for the analysis. Handling the missings at the total score level means that the total scores will be incomplete when one or more item scores are missing. The missing data handling method is applied to these total scores directly. Previous studies have shown that it is most beneficial to handle the missing data in a multi-item questionnaire at the item level. Handling missing item scores at the item level improves precision (Eekhout et al., 2014; Gottschall et al., 2012). In the context of multiple imputation it is quite straightforward to handle the missings at the item level. The item scores are imputed in the imputation model, and after the imputation part, the item scores are summed to the total scores in each of the imputed datasets, which are used for the analysis. However, when the number of items is very large, for example in longitudinal studies where item scores from multiple time points are included in the analysis, multiple imputation of the item scores might cause complications. When the number of items in the study gets close to the sample size, there is not enough information in the data to estimate the imputation model parameters. For example in a study where a multi-item questionnaire with 20 items is measured at six time points, the total number of variables in an imputation model would be at least 120. Green (1991) described a rule of thumb where the sample size should be larger than 53+k to do a regression analysis for a medium effect size (i.e., 0.13), where k is the number of predictors. In the example we outline below with 120 variables, the minimum sample size should then be 173. Hence, the number of variables in an imputation model could easily exceed the maximum allowed number in a longitudinal study with many time-points and a multi-item questionnaire as outcome measure. Moreover, when outcomes are measured at multiple time-points in a longitudinal study it might be feasible to analyze the data with a longitudinal analysis method such as a latent growth model. Usually these models are estimated with FIML, which produces unbiased model estimates when missing outcomes are missing at random. Nevertheless, the item scores are generally not included in such an analysis, because mostly only the total scores are modeled. Ergo, growth models estimated by FIML encourage users to deal with missing data at the scale level rather than the item level. Since, as previously mentioned, it is better to handle missings in a multi-item questionnaire at the item level, it would be beneficial to include the item scores in the analysis as well.

In a previous simulation study, we investigated two novel methods for including item-level information in a latent growth model while still focusing on change at the scale score level; the purpose of that study was to outline a FIML analog to item-level imputation (Eekhout et al., in press). We showed that these methods yield valid and precise parameter estimates in a latent growth model when total scores were missing due to missing item scores. In that study, the item information was included in the

model as auxiliary variables using well-established methods outlined by Graham (2003). Auxiliary variables are variables that are used to include extra information about the missingness of the data. They are related to the missingness in the data and/or are correlated with the incomplete variables. Including these variables in a missing data analysis will reduce bias and improve precision lost due to missing data (Collins, Schafer, & Kam, 2001). Auxiliary variable techniques are usually employed to incorporate predictors of missingness, thereby increasing the plausibility of the MAR assumption. We use these techniques to incorporate item-level information into a total score analysis. In this paper we will explain and demonstrate these two methods which include different item score information as auxiliary variables in a longitudinal study by using a latent growth model in two data examples.

Methods

Data examples

Data example 1 is a dataset from a study where the longitudinal effects of a randomized controlled trial were analyzed in which three treatments for neck pain were compared: manual therapy (specific mobilization techniques), physical therapy (exercise therapy), and usual care (analgesics, counseling and education) (Hoving et al., 2002). The main outcomes in the study were global perceived recovery, physical functioning, pain intensity, and neck disability. One of the secondary outcomes in the study was physical functioning. Physical functioning was measured by the physical functioning scale of the SF-36, which contains 10 items measured at a three point likert scale (Ware, Kosinski, & Keller, 1994). The item scores can be summed to obtain a total score for physical functioning. The outcome was measured at baseline and after 3, and 7 weeks. For this example we used the multi-item data from the physical function scale and included the treatment variable as central independent variable and age as a covariate in a latent growth analysis. In this dataset 170 out of 183 participants had completely observed data. We generated missing values in the items of the outcome measures in order to create situations for which we could compare the models that include the item information as auxiliary variables with the model without this auxiliary information. We used the 170 cases with complete data as a reference. In a copy of this dataset, missings were generated on the item-level of the SF-36 subscale for Physical Functioning by using the treatment variable and the age variable as predictors for missingness. That way the missing data on the items was missing at random. The baseline wave was complete. For the measurements at three and seven weeks about half of the items had about 15% missing data.

Data example 2 is from a randomized controlled trial about low back pain. The study population consisted of 299 workers that were listed as sick for a period of

three weeks due to low back pain. Three treatment groups were compared in a randomized controlled trial. The treatments were high-intensity and low-intensity back schools compared to the usual treatment by the occupational physician. The outcomes were measured at baseline, after three and after six months and were days until return to work, days of sick-leave, pain, functional disability, kinesiophobia, and perceived recovery. The results for the treatment effects for the main outcomes were published previously (Heymans et al., 2006). For this example we used the data from the passive coping scale of the Perceived Coping Inventory as the outcome which was also measured at the three time points (Kraaimaat & Evers, 2003). This subscale contains 21 items measured on a four point Likert scale. Data example 2 is a dataset that already contained missing data. The missing data in this dataset was mostly due to participants that missed an entire wave. At baseline 4% of the participants didn't return the questionnaire, at wave 2 26% and at wave 3 30%. We generated additional missing values for the item scores to present a data situation with missing total scores due to item scores as well as missings caused by participants not returning the questionnaire. The resulting overall average percentage of missing item scores was 25%.

In summary, in data example 1 we only generated incidental missing item scores and in data example 2 we present a situation where missing data on the total scores were caused by both the item score missings and by participants missing entire measurement waves. Additionally, the data in data example 2 also contains missing data for the baseline measurement. Both missing data situations are realistic and common in epidemiological studies. Furthermore, the number of items per scale of data example 2 is twice as high as the number of items per scale in data example 1.

Full Information Maximum Likelihood analyses

The data for both examples were analyzed by a latent growth model estimated with FIML. In a latent growth model the change in total scores over time is modeled, where the individual growth of each case in the study can be treated as a random effect. That way the variance between persons is taken into account, because person A might have a different development over time than person B. So the intercept and slope coefficients may vary across individuals, and are therefore referred to as random effects, or latent growth factors (Kwok et al., 2008). In models that use questionnaire total scores as the outcome, the total scores are computed prior to including them in the analysis. The total score is only computed when all items are observed. When some or all items are missing, the total score is missing. So for each wave the observed item scores are ignored for the cases with incomplete item scores.

In order to examine the change in physical functioning over the three time-points for data example 1 we used the model of Figure 6.1. The factor loadings of the latent

intercept were fixed at 1 and the factor loadings of the slope factor were set at the time-scores, which were 0, 3, and 7. The age and treatment covariates were included in the model. The three treatment categories were included as two dummy variables in order to distinguish between the effects of each treatment. The total scores are the sums of the item scores at each measurement wave. The estimates from the reference dataset with the 170 participants with complete observations were compared to the model estimates of the dataset with incomplete total scores due to the generated missing item scores.

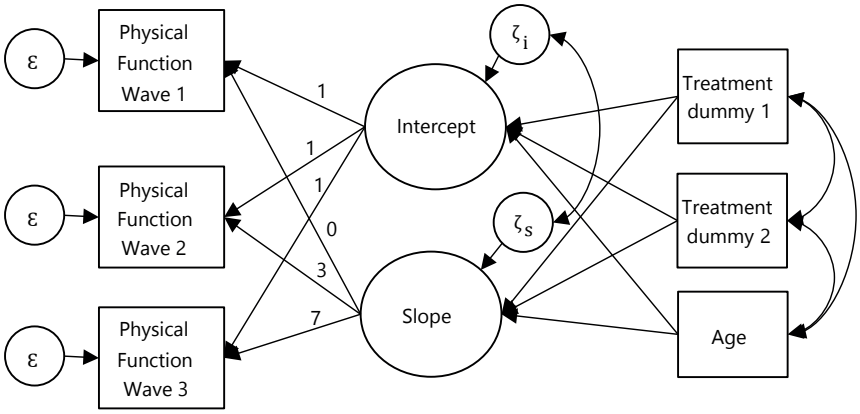


Figure 6.1. Latent growth model diagram for example data 1. Treatment dummy 1 denotes physical therapy versus continued care by a general practitioner; treatment dummy 2 denotes manual therapy versus continued care by a general practitioner.

For data example 2, the latent growth model presented in Figure 6.2 was fitted to measure the change in the passive coping score. The factor loadings for the growth factor were 0 for baseline and 3 and 6 for the follow-up waves. The loadings for the intercept factor were fixed at 1. The treatment variable was included as a dummy variable in the model in order to distinguish between the effects of the separate treatments. The total scores for passive coping are the sum of the item scores for each measurement wave and these were incomplete when one or more items were missing.

For each model we compared estimates for the average baseline score for the control group (intercept latent mean), the average difference at baseline for each treatment group relative to the control group (intercept on treatment), the average growth of the control group (slope latent mean) and the difference of linear growth for each treatment group relative to the control condition (slope on treatment).

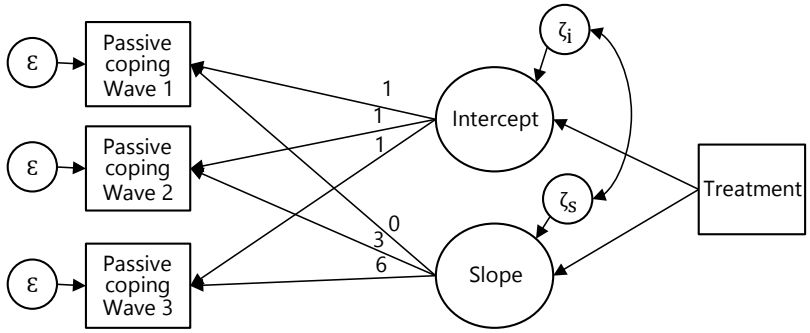


Figure 6.2. Latent growth model diagram for example data 2. Treatment dummy 1 denotes low-intensity treatment versus usual care; treatment dummy 2 denotes high-intensity care versus usual care.

Including the item information as auxiliary variables

Usually, for each wave only the cases with completely observed item data are included in the latent growth analyses (as in the analysis described above), since only for those cases the total score can be computed. This leads to a decreased precision of estimates, because less than an optimal amount of information is included. Furthermore, the scale scores at different waves could have different missingness rates. In order to improve estimates the item information was included as auxiliary variables in the models for data in the datasets with missing values. Graham (2003) described a method to include auxiliary variables that can be applied to structural equation models that use latent variables. The auxiliary variables should be (a) correlated to the manifest independent variables, (b) correlated to the residuals of all manifest endogenous variables (e.g., repeated measured scale scores); and (c) correlated with each other. The item scores would be ideal candidates as auxiliary variables, since the item scores are related to the scale scores and to the missingness on the scale scores as well. Accordingly we can include item scores by (a) correlating them in the model to the independent variables (e.g., treatment and age from data example 1), and (b) to the residuals of the scale scores from each wave and (c) correlating them with each other.

In Figure 6.3 an example of two auxiliary item scores included in the model from data example 1 is displayed. Item information can be included in the model by two methods: using the item scores as auxiliary variables or using a parcel summary score of the items as auxiliary variables (Eekhout et al., in press). For the method where the item scores were used, we included the observed item scores for each time-point in the auxiliary part of the model. That way, additional to the main model information, also the information from the items is used to estimate the most likely model parameters. It would be most ideal to include as many item scores as

possible while still reaching convergence of the model. The process of obtaining the full information maximum likelihood estimates is called convergence. Convergence problems can be related to the fact that the auxiliary part of the model is too similar to the total score outcomes, i.e., collinearity. In that case some extra noise should be added by removing some items, minimally one item per wave with the most missing data. Another reason for a lack of convergence might be that the number of correlations that have to be estimated in the model exceeds the sample size. For example when number of included item scores is large in longer questionnaires measured at many time-points. Therefore a second method that includes a summary of the item scores or a summary of a parcel of the item scores as auxiliary variables can be used. An example of such a summary is the average over the available items. That way the item information is included in the model without over-complicating the model estimation process. Our rationale for using a parcel summary score, is that the average of available items can capture most of the available information in the observed items while dramatically reducing the number of parameters (Enders, 2008). The average over the available items proved to be a valid and efficient method to include the item information (Eekhout et al., in press). Both of these methods accomplish the same end-point, which is smaller standard errors and therefore increase the precision of model estimates.

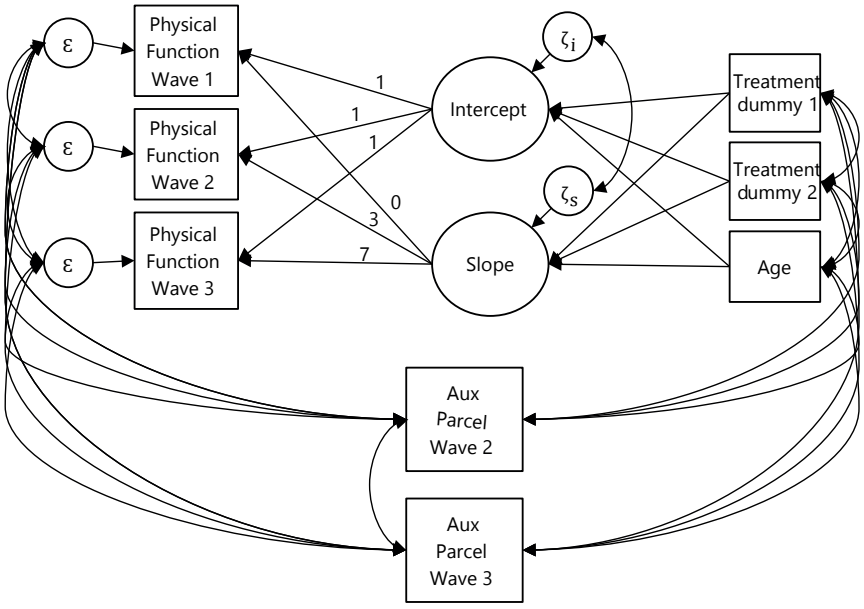


Figure 6.3. Auxiliary variables included in the latent growth model of data example 1. Treatment dummy 1 denotes physical therapy versus continued care by a general practitioner; treatment dummy 2 denotes manual therapy versus continued care by a general practitioner

For the first method that includes the item information it would be desirable to include 50% or more of the items as auxiliary variables. In data example 1, we included all but one items per wave with missing data. For data example 2, there was also some missing data on the baseline wave. So we also included the items from that wave in the auxiliary part of the model. Including all but one items per wave in the model could not be estimated, so we included 17 out of the 21 items per wave. We included the items with the lowest percentages of missing values.

For the second method, using a parcel summary of the items in order to include the item information, it is again desirable to include at least 50% of the items in the parcel summary. The parcel summary score was computed for each wave by taking the average value of the available items in the parcel. For data example 1, the parcel summary score of all but one item for wave 2 and 3 were included in the parcel. For the data example 2, the parcel summary scores for all but two items were used for all waves, because in this example also the baseline contained missing data. We excluded two items per wave, because the parcel summary scores of all but one item were too similar to the total score outcomes in the model and therefore caused computational problems. For each wave, we excluded the two items with the most missing data.

In summary, for each data example we compared two procedures that include the auxiliary item information. The first method is the inclusion of the item scores separately and the second is the summary scores of the items. In data example 1 these procedures were compared to the reference results from the complete data and to the results from a model on the incomplete data without auxiliary variables included. In data example 2 we compared results from the methods with auxiliary variables to the results from the model in the incomplete data without auxiliary variables included. All models were estimated by full information maximum likelihood in Mplus (Muthén, Asparouhov, Hunter, & Leuchter, 2010). A detailed manual on how to apply these methods in Mplus is available from the first author upon request; the Mplus syntaxes for the two example datasets are presented in the Appendix 6.1 and 6.2.

Results

For all models of data example 1 the parameter estimates are presented in Table 6.1. In the first column the results of the complete data analysis are presented as a reference. The estimates of the incomplete data from the model without auxiliary variables show that the standard errors for the slope parameters are increased compared to the results from the complete data. This is what was expected from the results of the simulation study that we previously performed (Eekhout et al., in press). When the item scores were used as auxiliary variables, the increase in standard errors

relative to the complete data model results was minimal. The same can be observed in the estimates from the model where the parcel summary scores were included. By computing the ratio of the squared standard errors for the model without auxiliary variables relative to the model with the auxiliary item information, we can compare the precision different on the sample size metric. Accordingly, for the slope on manual therapy parameter this ratio is: 0.1172/0.0842 = 1.94, which means that the model without auxiliary variables would require a 94% increase in the sample size to achieve the same precision as FIML with auxiliary variables.

Table 6.1.
Coefficient and standard error estimates for the compared methods for data example 1

Parameter	Complete data	No auxiliary	Item scores	Parcel summary
Intercept latent mean	25.678(1.108)*	24.684(1.115)*	24.846(1.110)*	24.897(1.116)*
Intercept physical therapy[a]	-0.131(0.597)	-0.029(0.600)	-0.183(0.598)	-0.225(0.602)
Intercept manual therapy[b]	-0.533(0.598)	-0.503(0.600)	-0.598(0.598)	-0.635(0.602)
Slope latent mean	0.281(0.153)	0.160(0.215)	0.272(0.156)	0.357(0.157)
Slope on physical therapy[a]	-0.007(0.083)	0.076(0.116)	-0.013(0.084)	-0.058(0.084)
Slope on manual therapy[b]	0.136(0.083)	0.191(0.117)	0.137(0.084)	0.084(0.084)

*Note: [a]The dummy that reflects physical therapy versus continued care by a general practitioner; [b]The dummy that reflects manual therapy versus continued care by a general practitioner; *P < 0.01.*

For data example 2 the results for the models on the original data of are presented in Table 6.2. In dataset, 20 subjects had missings on all repeated measurements and these were excluded for the model without auxiliary variables model. For the models with auxiliary variables, 16 cases had some observed items which were included in the auxiliary part of the model.

Table 6.2.
Coefficient and standard error estimates for the compared methods for data example 2

Parameter	No auxiliary	Item scores	Parcel summary
Intercept latent mean	44.214(0.982)*	44.054(0.927)*	44.069(0.922)*
Intercept low-intensity[a]	1.864(1.417)	2.091(1.338)	1.628(1.329)
Intercept high-intensity[b]	0.832(1.407)	0.692(1.341)	0.540(1.335)
Slope latent mean	-1.355(0.213)*	-1.020(0.170)*	-1.116(0.164)*
Slope on low-intensity[a]	-0.491(0.312)	-0.896(0.242)*	-0.674(0.233)*
Slope on high intensity[b]	-0.031(0.306)	-0.064(0.244)	0.047(0.236)

*Note: [a]The dummy that reflects low-intensity back school treatment versus usual care; [b]The dummy that reflects high-intensity back school treatment versus usual care by an occupational therapist; *P<0.01.*

In the results from Table 6.2 we can observe a gain in precision reflected in the smaller standard errors for the models with items scores and parcel summary scores included in as auxiliary variables, this gain was most apparent for the slope parameters. In addition, the regression coefficients show a stronger effect. The model without auxiliary variables did not present any significant treatment effect

on passive coping. However, the models that included the auxiliary item information (i.e., item scores or parcel summary scores) showed a significant slope for low-intensity back school treatment. In this example we can see how improved methods can actually affect study conclusions. As for the previous example, we can calculate the effect of the precision gain also by putting the standard error differences on the sample size metric. For the slope on low-intensity treatment this ratio would be 0.3122/0.2422=1.66 and 0.3122/0.2332=1.79 for including item scores or a parcel summary of the items respectively. These ratios imply a required increase in sample size for a model without auxiliary variables of 66% and 79% to reach the same precision as in the models with auxiliary item information.

Discussion

In this paper we presented two examples of longitudinal data analyses with a growth model when total scores are missing due to missing item scores. The compared models that include auxiliary item information improve the precision of the growth estimates which is important to correctly estimate a treatment effect. The level of precision that was obtained in the models that include the auxiliary item information can only be obtained in a model without auxiliary variables by increasing the sample size substantively. As was shown in the examples the required increase in sample size to reach the same level of precision can be as high as 94% (i.e., doubling the sample size). Furthermore, in data example 2 we showed that smaller standard errors caused by the auxiliary item information resulted in a significant treatment effect for the low-intensity back school. Especially in such clinical research situation it is important to estimate a model with optimal precision.

We presented two different methods to include item information in the auxiliary part of a latent growth model. In the first method the item scores were included separately and the most optimal amount of information is included in the model. However, the amount of correlations that have to be estimated in such a model can become problematically large. For that reason we also presented a method where a parcel summary score of the items is included. Including the item scores separately or including a parcel summary score of the item scores both performed well and improve precision. However, the model with the parcel summary score is easier to estimate and is therefore advised.

A previous simulation study of our group showed the performance with respect to bias (i.e., better coefficient estimates) and precision (i.e., smaller standard errors) in many longitudinal data situations (Eekhout et al., in press). Thought the bias was minimal for all tested FIML models, the effect on precision was substantial; for the models that did not include auxiliary item information, sample sizes should nearly be doubled to achieve the same level of precision as in the models with a parcel

summary score of the items. The results from the data example 1 in the current study were compared with a complete data situation, so in that case we have a true reference situation to show that the inclusion of the item score information does not change the model interpretation, but improves the growth estimates in the model. This example contained incomplete total scores due to missing item data. Data example 2 presents a situation where missing total scores result from incidental missing item scores but also from participants that did not return the questionnaire. Also in that missing data situation the inclusion of item information in the auxiliary part of the model is beneficial. Furthermore, when data are missing at the baseline wave, cases are excluded from analyses as the 20 subjects in data example 2. By including the auxiliary item information for the cases that have observed item scores available, more cases are part of the analysis.

Strengths and limitations

This study shows the performance of including item information in the auxiliary part of a latent growth model to improve precision of parameter estimates in an empirical longitudinal dataset. The applications of our methods to empirical data correspond to the results from a previous simulation study. The presentation of data example 2 showed that the improvement of precision and accuracy of parameter estimates can be crucial in some data situations.

For both example datasets we generated extra missing data at the item level. Many epidemiological studies encounter missing data problems and when multi-questionnaires are used the missing data often occurs at the item level. By generating situations with extra missing data at the item level we can present the robustness of the methods we propose in such realistic missing item data situations.

The parameter estimates presented for the data examples sometimes seem to vary a little across the methods. For example, in data example 1 the slope on manual therapy estimate varied between 0.136 for the complete data to 0.191 in the model without auxiliary variables and 0.137 and 0.084 in the models with auxiliary item information. However, in our simulation study we found that the data with missing total scores analyzed in a latent growth model estimated by FIML does not bias parameter estimates (Eekhout et al., in press). Furthermore, the parameter estimates for intercept and slope vary together. For example, the intercept and slope on manual therapy for complete data in data example 1 were -0.533 and 0.136 and for the incomplete data model without auxiliary variables the intercept and slope were -0.503 and 0.191, so in the complete data the intercept is a bit further from zero but the slope is closer to zero and the opposite is true for the model on the incomplete data.

The datasets we used in the examples were chosen to demonstrate our methods

in different situations. Data example 1 includes a questionnaire of 10 items, while data example 2 includes a questionnaire with 21 items. For the method where the item scores are included separately in the auxiliary part of the model, all but one item is most optimal. However, when questionnaires contain many items, including all but one item score per wave can cause computational difficulties. In data example 2, we included 17 items per wave in the auxiliary part of the model. Nevertheless, this model still improved the precision of the estimates compared to not including auxiliary item information. This showed that even including a smaller part of the items can be very beneficial.

This paper aimed to explain how the inclusion of item information in the auxiliary part of a latent growth model works and to show the feasibility and the effects of the inclusion of item information in empirical longitudinal data. As previously mentioned, it is most feasible to include as much information as possible in the auxiliary part of the model. This is also applicable for the composition of the parcel summary score. The performance of different compositions of parcel scores should be further explored in a simulation study. In a small simulation previously conducted (data not shown) we found that using 50% of the available item scores already improves estimates. The item scores can then be included either separately or as a summary score. However, the most optimal number of items relative to scale length or number of repeated measures was not explored extensively yet, but can be studied in future research.

References

Bracken, B. A., & Howell, K. (2004). Clinical Assessment of Depression: Professional manual. Odessa, FL: Psychological Assessment Resources.

van Buuren, S. (2012). Flexible Imputation of Missing data. New York: Chapman & Hall/ CRC.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological Methods, 6(4), 330-351.

Eekhout, I., de Boer, R. M., Twisk, J. W., de Vet, H. C., & Heymans, M. W. (2012). Missing data: a systematic review of how they are reported and handled. Epidemiology, 23(5), 729-732.

Eekhout, I., De Vet, H. C. W., Twisk, J. W. R., Brand, J. P. L., De Boer, M. R., & Heymans, M. W. (2014). Missing data in a multi-item instrument were best handled by multiple imputation at the item score level. Journal of Clinical Epidemiology, 67(3), 335-342.

Eekhout, I., Enders, C. K., Twisk, J. W. R., De Boer, M. R., de Vet, H. C. W., & Heymans, M. W. (in press). Analyzing Incomplete Item Scores in Longitudinal Data by Including Item Score Information as Auxiliary Variables. Structural Equation Modeling: A Multidisciplinary Journal.

Enders, C. K. (2008). A Note on the Use of Missing Auxiliary Variables in Full Information Maximum Likelihood-Based Structural Equation Models. Structural Equation Modeling: A Multidisciplinary Journal, 15(3), 434-448.

Gottschall, A. C., West, S. G., & Enders, C. K. (2012). A Comparison of Item-Level and Scale-Level Multiple Imputation for Questionnaire Batteries. Multivariate Behavioral Research, 47(1), 1-25.

Graham, J. W. (2003). Adding Missing-Data-Relevant Variables to FIML-Based Structural Equation Models. Structural Equation Modeling: A Multidisciplinary Journal, 10(1), 80-100.

Green, S. B. (1991). How Many Subjects Does It Take To Do A Regression Analysis. Multivariate Behavioral Research, 26(3), 499-510.

Heymans, M. W., de Vet, H. C., Bongers, P. M., Knol, D. L., Koes, B. W., & van Mechelen, W. (2006). The effectiveness of high-intensity versus low-intensity back schools in an occupational setting: a pragmatic randomized controlled trial. Spine (Phila Pa 1976), 31(10), 1075-1082.

Hoving, J. L., Koes, B. W., de Vet, H. C., van der Windt, D. A., Assendelft, W. J., van Mameren, H., . . . Bouter, L. M. (2002). Manual therapy, physical therapy, or continued care by a general practitioner for patients with neck pain. A randomized, controlled trial. Ann Intern Med, 136(10), 713-722.

Kraaimaat, F. W., & Evers, A. W. (2003). Pain-coping strategies in chronic pain patients: psychometric characteristics of the pain-coping inventory (PCI). Int J Behav Med, 10(4), 343-363.

Kwok, O. M., Underhill, A. T., Berry, J. W., Luo, W., Elliott, T. R., & Yoon, M. (2008). Analyzing Longitudinal Data with Multilevel Models: An Example with Individuals Living with Lower Extremity Intra-articular Fractures. Rehabil Psychol, 53(3), 370-386.

Lambert, M. J., Lunnen, K., Umphress, V., Hansen, N., & Burlingame, G. M. (1994). Administration and scoring manual for the Outcome Questionnaire (OQ-45.1). Salt Lake City: IHC Center for Behavioral Healthcare Efficacy.

Muthén, B. O., Asparouhov, T., Hunter, A., & Leuchter, A. (2010). Growth Modeling with Non-Ignorable Dropout: alternative analysis of the STAR*D antidepressant trial. February version wide.

Rubin, D. B. (1976). Inference and missing data. Biometrika, 63(3), 581-592.

Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: J. Wiley & Sons.

Schafer, J. L. (1997). Analysis of Incomplete Multivariate Data. London, UK: Chapman & Hall.

Ware, J. E., Jr., Kosinski, M., & Keller, S. D. (1994). SF-36 Physical and Mental Health Summary Scales: A User's Manual. Boston, MA: Health Assessment Lab.

Appendix 6.1|Mplus syntaxes example data 1

Growth model without auxiliary variables

```
Mplus VERSION 7.11
MUTHEN & MUTHEN

INPUT INSTRUCTIONS
data:
!file ='C:\filelocation\dataset1_complete.dat';
file ='C:\filelocation\dataset1_missing.dat';
variable:
names =
y1_1 - y1_10
y2_1 - y2_10
y3_1 - y3_10
tx age;
usevariables = age dummy1 dummy2 scale1 scale2 scale3;
missing = all(-9);
define:
!define dummies for the treatment groups;
IF (tx==1) THEN dummy1=0;
IF (tx==2) THEN dummy1=1;
IF (tx==3) THEN dummy1=0;
IF (tx==1) THEN dummy2=1;
IF (tx==2) THEN dummy2=0;
IF (tx==3) THEN dummy2=0;
! calculation of scale scores;
scale1 = sum(y1_1-y1_10);
scale2 = sum(y2_1-y2_10);
scale3 = sum(y3_1-y3_10);
model:
i s | scale1@0 scale2@3 scale3@7;
i on age
dummy1
dummy2;
s on age
dummy1
dummy2;
i with s;
[i];
[s];
i;
s;
scale1 - scale3 (resvar);
```

Growth model with item scores as auxiliary variables

```
Mplus VERSION 7.11
MUTHEN & MUTHEN

INPUT INSTRUCTIONS
data:
file ='C:\filelocation\dataset1_missing.dat';
variable:
names =
y1_1 - y1_10
y2_1 - y2_10
y3_1 - y3_10
tx age;
usevariables = age dummy1 dummy2 scale1 - scale3;
missing = all(-9);
! including the items as auxiliary variables;
auxiliary = (m) y2_2-y2_10
y3_1-y3_3 y3_5-y3_10;
define:
!define dummies for the treatment groups;
IF (tx==1) THEN dummy1=0;
IF (tx==2) THEN dummy1=1;
IF (tx==3) THEN dummy1=0;
IF (tx==1) THEN dummy2=1;
IF (tx==2) THEN dummy2=0;
IF (tx==3) THEN dummy2=0;
! calculation of scale scores;
scale1 = sum(y1_1-y1_10);
scale2 = sum(y2_1-y2_10);
scale3 = sum(y3_1-y3_10);
model:
i s | scale1@0 scale2@3 scale3@7;
i on age
dummy1
dummy2;
s on age
dummy1
dummy2;
i with s;
[i];
[s];
i;
s;
scale1 - scale3(resvar);
```

Growth model with the parcel scores as auxiliary variables

```
Mplus VERSION 7.11
MUTHEN & MUTHEN

INPUT INSTRUCTIONS
data:
file ='C:\filelocation\dataset1_missing.dat';
variable:
names =
y1_1 - y1_10
y2_1 - y2_10
y3_1 - y3_10
tx age;
usevariables = age dummy1 dummy2 scale1-scale3 parcel2-parcel3;
missing = all(-9);
! including the parcel scores as auxiliary variables;
auxiliary = (m) parcel2 - parcel3;
define:
!define dummies for the treatment groups;
IF (tx==1) THEN dummy1=0;
IF (tx==2) THEN dummy1=1;
IF (tx==3) THEN dummy1=0;
IF (tx==1) THEN dummy2=1;
IF (tx==2) THEN dummy2=0;
IF (tx==3) THEN dummy2=0;
! calculation of scale scores;
scale1 = sum(y1_1-y1_10);
scale2 = sum(y2_1-y2_10);
scale3 = sum(y3_1-y3_10);
! calculation of the parcel summary scores;
parcel2 = mean(y2_2-y2_10);
parcel3 = mean(y3_1-y3_3 y3_5-y3_10);
model:
i s | scale1@0 scale2@3 scale3@7;
i on age
dummy1
dummy2;
s on age
dummy1
dummy2;
i with s;
[i];
[s];
i;
s;
scale1 – scale3 (resvar);
```

Appendix 6.2|Mplus syntaxes example data 2

Growth model without auxiliary variables

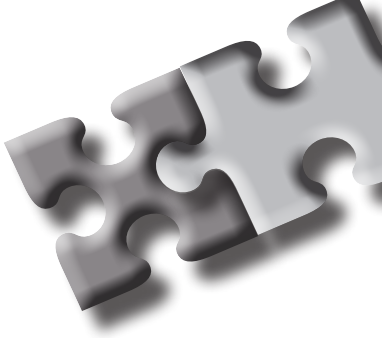
```
Mplus VERSION 7.11
MUTHEN & MUTHEN

INPUT INSTRUCTIONS
data:
file ='C:\filelocation\dataset2_missing.dat';
variable:
names =
y1_1 - y1_21
y2_1 - y2_21
y3_1 - y3_21
tx age;
usevariables = age dummy1 dummy2 scale1 scale2 scale3;
missing = all(-9);
define:
!define dummies for the treatment groups;
IF (tx==0) THEN dummy1=0;
IF (tx==1) THEN dummy1=1;
IF (tx==2) THEN dummy1=0;
IF (tx==0) THEN dummy2=0;
IF (tx==1) THEN dummy2=0;
IF (tx==2) THEN dummy2=1;
! calculation of scale scores;
scale1 = sum(y1_1-y1_21);
scale2 = sum(y2_1-y2_21);
scale3 = sum(y3_1-y3_21);
model:
i s | scale1@0 scale2@3 scale3@6;
i on dummy1
dummy2;
s on dummy1
dummy2;
i with s;
[i];
[s];
i;
s;
scale1 - scale3 (resvar);
```

Growth model with item scores as auxiliary variables

```
Mplus VERSION 7.11
MUTHEN & MUTHEN

INPUT INSTRUCTIONS
data:
file ='C:\filelocation\dataset2_missing.dat';
variable:
names =
y1_1 - y1_21
y2_1 - y2_21
y3_1 - y3_21
tx age;
usevariables = age dummy1 dummy2 scale1 - scale3;
missing = all(-9);
! including the items as auxiliary variables;
auxiliary = (m) y1_1-1_2 y1_6-y1_9 y1_11-y1_21
y2_1-y2_4 y2_6-y2_11 y2_13-y2_15 y2_17 y2_19-y2_21
y3_1-y3_3 y3_5-y3_12 y3_14 y3_16-y3_18 y3_20 y3_21;
define:
!define dummies for the treatment groups;
IF (tx==0) THEN dummy1=0;
IF (tx==1) THEN dummy1=1;
IF (tx==2) THEN dummy1=0;
IF (tx==0) THEN dummy2=0;
IF (tx==1) THEN dummy2=0;
IF (tx==2) THEN dummy2=1;
! calculation of scale scores;
scale1 = sum(y1_1-y1_21);
scale2 = sum(y2_1-y2_21);
scale3 = sum(y3_1-y3_21);
model:
i s | scale1@0 scale2@3 scale3@6;
i on dummy1
dummy2;
s on dummy1
dummy2;
i with s;
[i];
[s];
i;
s;
scale1 - scale3(resvar);
```

Growth model with the parcel scores as auxiliary variables

```
Mplus VERSION 7.11
MUTHEN & MUTHEN

INPUT INSTRUCTIONS
data:
file ='C:\filelocation\dataset2_missing.dat';
variable:
names =
y1_1 - y1_21
y2_1 - y2_21
y3_1 - y3_21
tx age;
usevariables = age dummy1 dummy2 scale1-scale3 parcel1-parcel3;
missing = all(-9);
! including the parcel scores as auxiliary variables;
auxiliary = (m) parcel1 - parcel3;
define:
!define dummies for the treatment groups;
IF (tx==0) THEN dummy1=0;
IF (tx==1) THEN dummy1=1;
IF (tx==2) THEN dummy1=0;
IF (tx==0) THEN dummy2=0;
IF (tx==1) THEN dummy2=0;
IF (tx==2) THEN dummy2=1;
! calculation of scale scores;
scale1 = sum(y1_1-y1_21);
scale2 = sum(y2_1-y2_21);
scale3 = sum(y3_1-y3_21);
! calculation of the parcel summary scores;
Parcel1 = mean(y1_1-1_2 y1_4-y1_9 y1_11-y1_21);
parcel2 = mean(y2_1-y2_11 y2_13-y2_15 y2_17-y2_21);
parcel3 = mean(y3_1-y3_12 y3_14-y3_18 y3_20 y3_21);
model:
i s | scale1@0 scale2@3 scale3@6;
i on dummy1
dummy2;
s on dummy1
dummy2;
i with s;
[i];
[s];
i;
s;
scale1 - scale3 (resvar);
```



Chapter 7

Handling missing data in sub-costs in a cost-effectiveness analysis

Under review: MacNeil-Vroomen, J., Eekhout, I., Dijkgraaf, M.G., Van Hout, H., De Rooij, S.E., Heymans, M.W., Bosmans, J.E. Multiple imputation strategies for zero-inflated cost data in economic evaluations: which method works best? The European Journal of Health Economics.

Abstract

Cost and effect data are prone to missing data because economic evaluations are often "piggy-backed" onto clinical studies where cost data are rarely the primary outcome. Multiple imputation is recommended for handling missing data. The objective of this article was to investigate which multiple imputation strategy is most appropriate to use for missing cost-effectiveness data in a pragmatic randomized controlled trial (RCT). Three incomplete datasets were generated from a complete reference dataset with 17%, 35% and 50% missing data in effects and costs. The strategies evaluated included complete-case analysis (CCA), multiple imputation with predictive mean matching (MI-PMM), MI-PMM on the log-transformed costs (Log MI-PMM), and a two-step MI. Mean cost and effect estimates, standard errors and incremental net benefits were compared with the results of the analyses on the complete reference dataset. The CCA, MI-PMM, and the two-step MI strategy deviated more from the results for the reference dataset when the amount of missing data increased. In contrast, the estimates of the Log MI-PMM strategy remained stable regardless of the amount of missing data. MI provided better estimates than CCA in all scenarios. With low amounts of missing data the MI strategies appeared equivalent but with missing data greater than 35%, we recommend using the Log MI-PMM.

Keywords: missing data, cost data, economic evaluations, multiple imputation, zero-inflated data

Introduction

Missing data may lead to loss of information in epidemiological and clinical research (Sterne et al., 2009). Therefore, researchers should aim for collecting high quality and complete data. However, missing data are unavoidable when performing trials where data is collected through self-report by the participants. Cost data are even more prone to have missing data because economic evaluations are often "piggy-backed" onto clinical studies where cost data are rarely the primary outcome. Moreover costs from several measurements are summed up in a total cost estimate, meaning that one missing measurement results in a missing total cost estimate.

Complete-case analysis (CCA) is the default strategy to deal with missing data in many statistical packages although it is known for deficiencies like biased estimates, wide standard errors and lowered power. Oostenbrink et al. (2003) and Briggs et al. (2003) showed that multiple imputation techniques to deal with missing data performed better than CCA and simple imputation techniques (conditional mean imputation, single imputation with predictive mean matching, hot decking and expectation maximization).

In the last few years, multiple imputation has been recommended as the most appropriate way for handling missing data (Klebanoff & Cole, 2008; Nietert, Wahlquist, & Herbert, 2013; Sterne et al., 2009; van Buuren, 2012; White, Royston, & Wood, 2011). Multiple imputation can be a powerful tool to estimate missing data (van Buuren, 2012), but there are some important points to consider when specifying the multiple imputation model. First, the imputation model should include all variables that explain missing values. Second, it should include all variables included in the analysis model and third the imputation model must account for the distribution of the data. This assumption may not be met when imputing cost data in trials because of the distributional issues posed by cost data including constrained positive values, a large amount of zero values, and right-handed tail skewness.

Multiple imputation with predictive mean matching (PMM) can be a helpful tool to deal with the skewed distribution of cost data because PMM preserves the distribution of the data and, therefore, is robust against violations of the normality assumption (van Buuren, 2012). Another commonly recommended approach to deal with skewed data is to take the log of the skewed variables before imputation and then back transform the variables to their original scale before the target analysis. (Lee & Carlin, 2010; van Buuren, 2012; White et al., 2011). Lee and Carlin (2010) compared multiple imputation with transformation and PMM to deal with non-normality in continuous variables. They recommended transformation of skewed variables to a symmetric distribution to avoid the introduction of biases of study results. Another alternative is to impute missing data in two separate steps. In the first step, the probability of having costs is imputed which takes care of the zero inflation, and in

the second step, an actual cost value is imputed for individuals that are predicted to have costs. In the second step, the skewness of the cost data is taken into account by using the PMM algorithm to impute the cost values for the people that are predicted to have costs using only the observed cost data (Javaras & Van Dyk, 2003).

It is unclear which method to deal with imputation of skewed data is the most appropriate in economic evaluations. Therefore, the objective of this article was to investigate which imputation strategy is most appropriate to impute missing cost and effect data in an economic evaluation alongside a pragmatic randomized controlled study. This study adds to previous studies by looking at costs, effects and cost-effectiveness while comparing different multiple imputation strategies. The strategies compared include CCA, MI with predictive mean matching (MI-PMM), MI with predictive mean matching on log-transformed costs (Log-MI-PMM), and two step multiple imputation with predictive mean matching (two-step-MI).

Methods

Reference dataset

The dataset used in this study was obtained from two open-labeled randomized controlled trials evaluating the cost-utility of medical co-prescription of heroin compared with methadone maintenance treatment alone for chronic, treatment resistant heroin addicts. Full details on this study are presented elsewhere (Dijkgraaf et al., 2005). Outcomes included QALYs based on the EuroQol (EQ-5D) and costs from a societal perspective (Brooks, 1996). Each participant completed the EQ-5D at baseline and at months 6, 10, and 12 during treatment. The health states from the EQ-5D were subsequently converted to utilities using the York tariff (Dolan, 1997). We calculated QALYs by multiplying the utility of each health state by the time in between two measurements and summing the results over the 12 month treatment period. Table 7.1 contains baseline characteristics and the variables used to calculate

Table 7.1.
Baseline characteristics of the reference dataset.

Explanatory variables	Methadone alone (n=237)	Co-prescribed heroin (n=193)
% male (n)	55.1 (190)	44.9 (155)
age (sd)	38.9 (5.7)	39.7 (5.8)
% injected (n)	56.3 (98)	43.7 (76)
% completed (n)	60.2 (204)	39.8 (135)
% abstinent (n)	59.3 (80)	40.7 (55)
% second interview performed (n)[a]	55.7 (59)	44.3 (47)
baseline utility (sd)	0.731 (0.273)	0.739 (0.272)

Note: [a] Those included early in the trials also completed the questionnaire in the second month.

the utilities and total costs respectively. Total costs consisted of program costs, law enforcement costs, costs of damage to victims, health related travel costs and other health care costs. Figures of the frequency distributions of each cost category in the reference dataset is presented in Figure 7.1. Occasionally missing values were imputed using last observation carried forward resulting in a complete dataset for all 430 participants. We used this dataset for the purpose of the present article.

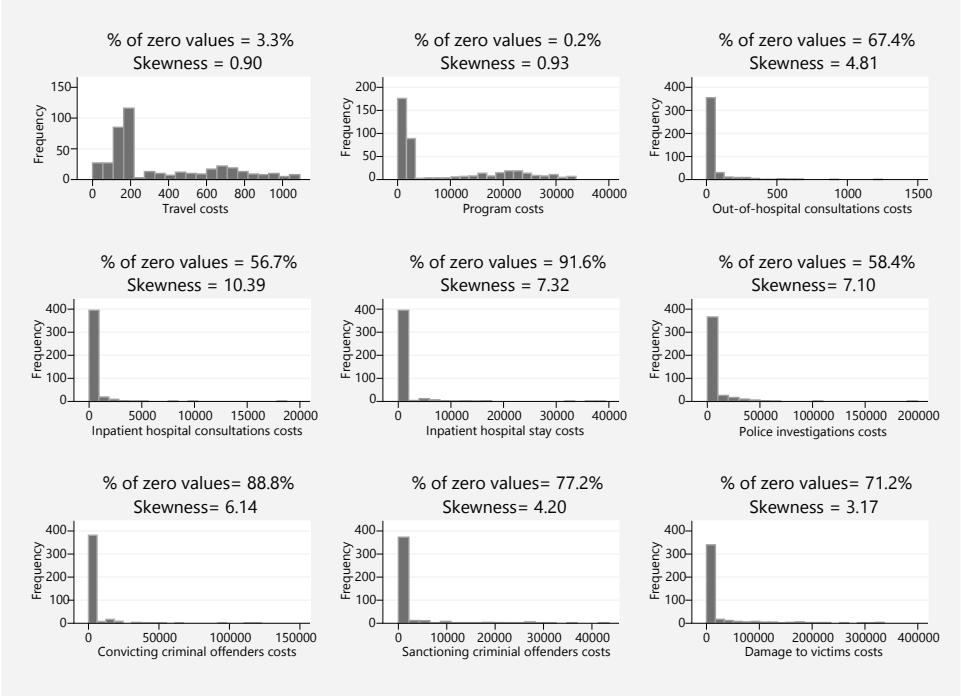


Figure 7.1. Frequency distributions of each cost category in the reference dataset (n=430).

Missing data

Author IE generated missing data in the complete dataset using R statistical software (R Core Development Team, 2014). In order to investigate the effect of the rate of incomplete data on the performance of the imputation methods, three incomplete datasets were created with 17%, 35% and 50% missing data. Missing data points were created in the QALY variable and several cost variables. In order to satisfy a missing at random (MAR) assumption for the missing data, the probability of missing data was related to other variables in the data. For the dataset with 17% missing data, center, location, age, occurrence of a second interview, and abstinence were predictors of missingness in the utility and cost outcome variables. In the

dataset with 35% missing data, predictor variables were treatment group, center, sex, age, and occurrence of a second interview. In the dataset with 50% missing data, the predictor variables were treatment group center, age and occurrence of a second interview. In Appendix 7.1 descriptive tables of all key cost variables with missing data are presented for the different missing data scenarios.

Missing data strategies

CCA

In CCA, analysis was restricted to participants with complete cost and effect data. This resulted in smaller sample sizes than in the reference dataset (see Appendix 7.1).

Multiple imputation procedure

Multiple imputation was done using multivariate imputation by chained equation (MICE). MICE or fully conditional specification is a flexible multivariate model that does not rely on the assumption of multivariate normality (van Buuren, 2012). Regression models are specified for each variable with missing values, conditional on all of the other variables in the imputation model. Imputations are generated by drawing from iterated conditional models (van Buuren, 2012).

The imputed values were estimated using the predictive mean matching (PMM) algorithm. PMM is an algorithm that matches the missing value to the observed value with the closest predicted estimate (White et al., 2011). The predicted mean is estimated in a regression equation where a random residual term is added to the estimate in order to account for missing data uncertainty. In PMM, instead of using the predicted estimate, the imputed value is randomly selected from observed values that are closest to the predicted estimate. For example, if an older single man misses a measurement for blood pressure and the value for this man is estimated to be 102.34 mmHg by regressing blood pressure on age and sex. Five other older single men have observed blood pressures of 103; 103; 102; 101, and 104 mmHg, respectively. The missing value is then imputed with a random draw from these five blood pressures. PMM has several advantages when imputing cost data. It is more robust against non-normal data as it uses the observed distribution of the data. Furthermore, it imputes only plausible values because it randomly draws from observed values. The process of estimating imputed values is repeated in sequential cycles, each time using the updated data with the imputed estimates from the previous cycle. These cycles are called iterations. One of these iterations (e.g., the 100th) is selected and used as an imputed dataset until 'm' datasets were selected in total. We used 200 imputations to minimize internal variation so that the imputation variation would not affect the performance of each imputation method (Enders,

2010; Horton & Lipsitz, 2001; Rubin, 1976; Sterne et al., 2009). We performed MI using the chained command in Stata 12, which uses fully conditional specification to perform the multiple imputations (Statscorp, 2011).

We performed the multiple imputations stratified by treatment group to maintain the possible group effect in the data. PMM uses one to the nearest neighbor as a default to draw from therefore it replaces missing values with an observed value whose linear prediction is the closest to that of the missing value. For all multiple imputation strategies we checked the convergence plots to see if iterations were free from trend and imputations were successful. To solve any occurring convergence problems, we merged highly correlated variables together. For this reason, travel costs were merged together with total program costs (correlation coefficient > 0.9). In-patient hospital consultations and inpatient length of hospital stay were also highly correlated and were therefore merged together as well. Three multiple imputation strategies were compared and are described below:

MI-PMM

In the first multiple imputation strategy we performed multiple imputation with predictive mean matching on the raw data.

Log MI-PMM

In the second multiple imputation strategy, we applied the predictive mean matching algorithm to the log transformed cost data. This was done by first adding a constant to the raw cost data in order to circumvent problems when transforming zero values, and next the log was taken. After imputation, the complete data were transformed back to their original scale prior to any analyses being performed.

Two-step MI

The third multiple imputation strategy was a conditional two-step approach. We recoded cost variables to dummy variables where subjects were coded as 1 if they had costs and a 0 for no costs. Missing values were left to be multiply imputed with either a 0 or 1 using a logit function. Next, multiple imputation with the PMM algorithm was performed for missing cases with a value 1 on the dummy variables. Only cases with cost estimates higher than zero were used for this imputation step. For variables that did not have a sufficient amount of zeroes to perform the conditional imputation, we chose to apply only the second step on the raw cost variable.

Statistical Analysis

We used a generalized linear regression model with a gamma distribution and an identity link to estimate mean differences in total costs. The gamma distribution was

chosen to take into account the right skewness of the cost data. A generalized linear regression model was used to estimate the incremental effect in quality adjusted life years (QALYs) adjusted for baseline utility estimates. Mean differences and standard errors were pooled using Rubin's rules (Rubin, 1976).

We estimated the correlation between the incremental total costs and the incremental QALYs in the reference dataset and the imputed datasets. In the multiple imputation strategies, the covariance between total costs and QALYs was calculated based on the Fisher z transformation and was then pooled using Rubin's rules (Schafer, 1997; van Buuren, 2012).

Incremental Cost-Effectiveness Ratios (ICERs) were calculated using the pooled cost and effect estimates. The ICER is calculated as

$$ICER=\hat{\Delta}_c/\hat{\Delta}_e$$

where $\hat{\Delta}_c$ is the difference in total costs between the two intervention groups and $\hat{\Delta}_e$ is the difference in QALYs between the two intervention groups. Incremental net benefit (INB) estimates were calculated using the following formula:

$$INB=\hat{b}(\lambda)=\hat{\Delta}_e\,\lambda-\hat{\Delta}_c$$

where $\hat{\Delta}_e$ is the difference in QALYs between the two intervention groups, λ is the willingness to pay, and $\hat{\Delta}_c$ is the difference in costs (Nixon, Wonderling, & Grieve, 2010; Willan & Briggs, 2006). The variance of INB was calculated using:

$$V[\hat{b}(\lambda)]=\hat{V}(\hat{\Delta}_e)\lambda^2+\hat{V}(\hat{\Delta}_c)-2\hat{C}(\hat{\Delta}_e,\hat{\Delta}_c)\lambda$$

where $\hat{C}$ is the covariance between the differences in total costs and QALYs (Nixon et al., 2010; Willan & Briggs, 2006). We set the willingness-to-pay at EUR 30,000 because this is roughly equivalent to the cut-off value mentioned in the Standard National Institute of Clinical Excellence guidelines (20,000-30,000 pounds per QALY) for economic evaluations (NICE, 2010)

Cost-effectiveness acceptability curves (CEAC) were estimated to quantify the uncertainty due to sampling and measurement errors and because lambda is generally unknown. The CEAC is a plot of the probability that co-prescribed heroin compared to methadone maintenance only is cost-effective (y-axis) as a function of the money society might be willing to pay for one additional QALY (x-axis). The pooled coefficients and variance parameters from the regression models were used for the CEACs.

Comparison of strategies

The estimates from the reference dataset were considered the "true values" and we compared the estimates from the different multiple imputation strategies with these true values. We evaluated the percentage of bias from the reference analysis

(RA) in the different imputation strategies for cost and effect differences, standard error estimates, p-values and t-values. The primary outcomes of interest were the value of INB at a willingness to pay of 30,000 EUR per QALY, the standard error of INB and the probability that co-prescribed heroin compared to methadone maintenance at a willingness to pay of 30,000 EUR per QALY. The strategies that gave the closest estimates to the reference dataset were considered the best.

Sensitivity analysis

Research has shown that it is better to impute at the item and not the total level. (Eekhout et al., 2013; Lambert, Billingham, Cooper, Sutton, & Abrams, 2008). In order to confirm the benefit of imputing at sub-cost level, we imputed the total cost variable directly as a sensitivity analysis using all missing data strategies.

Results

Costs

Table 7.2 presents the cost estimates for the reference case, the CCA, and the different imputation strategies for 17%, 35% and 50% missing data. The difference in costs of -12,792 euro in the RA fell within the confidence intervals of all multiple imputation strategies for all rates of missing data. The CCA deviated the most from the RA compared to all other strategies specifically with regard to the cost differences and the associated standard errors in all scenarios. For 17% of missing data, the CCA showed a statistically significant difference in costs just as in the reference analysis. However, for 35% or 50% of missing data the cost difference was no longer statistically significant. The multiple imputations strategies gave similar results to each other in the 17% and 35% missing datasets which were smaller differences in costs and larger standard errors when the amount of missing data increased compared to the reference analysis. The log transformed-PMM deviated the least from the RA in the 50% missing dataset for the cost difference, standard error and p-values. The two-step MI deviated the most from the RA with regard to cost differences, the standard errors and power in the dataset with 50% missing data.

QALYs

Table 7.3 provides the QALY results for the 17%, 35% and 50% missing data. In the 17% missing dataset, all strategies deviated roughly by the same amount for the difference in QALYs, standard error and power. All imputation strategies, including the CCA showed a statistically significant difference ($p < 0.001$) in QALYs between the two intervention groups.

Table 7.2.
Overview of cost estimates for the missing data methods.

	RA		CCA (% Bias)		MI-PMM (% Bias)		Log-MI-PMM (% Bias)		Two-step MI (% Bias)	
17% missing data	M	M+H	M	M+H	M	M+H	M	M+H	M	M+H
n	237	193	201	154	237	193	237	193	237	193
mean	50,560	37,767	53,148	38,933	51,369	37,935	51,966	38,137	52,685	38,482
SE mean	5,359	3,063	6,056	3,744	5,650	3,268	5,652	3,309	5,642	3,394
treatment cost difference	-12,792		-14,215 (11)		-13,434 (5)		-13,829 (8)		-13,203 (3)	
SE cost difference	6,086		7,077 (14)		6,440 (5)		6,459 (5)		6,506 (6)	
z for CCA and t for MI	-2.100		-2.010		-2.090		-2.140		-2.030	
p-value	0.036		0.045		0.037		0.032		0.042	
95% CI lower limit	-24,720		-28,085		-26,057		-26,489		-25,954	
95% CI upper limit	-865		-345		-810		-1,169		-452	
35% missing data	M	M+H	M	M+H	M	M+H	M	M+H	M	M+H
n	237	193	163	122	237	193	237	193	237	193
mean	50,560	37,767	52,255	43,176	50,810	39,851	51,434	40,408	51,195	40,426
SE mean	5,359	3,063	6,953	4,560	5,989	3,448	5,975	3,601	6,052	3,551
treatment cost difference	-12,792		-9,080 (29)		-10,959 (14)		-11,026 (16)		-10,769 (16)	
SE cost difference	6,086		8,463 (39)		6,853 (13)		6,988 (15)		6,954 (14)	
z for CCA and t for MI	-2.100		-1.070		-1.600		-1.580		-1.550	
p-value	0.036		0.283		0.110		0.115		0.122	
95% CI lower limit	-24,720		-25,667		-24,393		-24,725		-24,400	
95% CI upper limit	-865		7,508		2,475		2,673		2,862	
50% missing data	M	M+H	M	M+H	M	M+H	M	M+H	M	M+H
n	237	193	132	91	237	193	237	193	237	193
mean	50,560	37,767	50,160	42,794	48,711	38,335	49,180	38,527	49,110	39,454
SE mean	5,359	3,063	7,447	5,336	5,875	3,513	5,857	3,501	5,913	3,683
treatment cost difference	-12,792		-7,366 (42)		-10,376 (19)		-10,653 (17)		-9,656 (25)	
SE cost difference	6,086		9,496 (56)		6,852 (13)		6,764 (11)		6,954 (14)	
z for CCA and t for MI	-2.100		-0.780		-1.510		-1.570		-1.390	
p-value	0.036		0.438		0.130		0.115		0.165	
95% CI lower limit	-24,720		-25,978		-23,810		-23,912		-23,289	
95% CI upper limit	-865		11,246		3,058		2,607		3,977	

Note: M refers to the methadone maintenance treatment group; M+H refers to the group that had medical co-prescription of heroin. SE means standard error; CI means confidence interval.

Table 7.3.
Overview clinical effect estimates of QALY model for the missing data methods

	RA		CCA (% Bias)		MI-PMM (% Bias)		Log-MI-PMM (% Bias)		Two-step MI (% Bias)	
17% missing data	M	M+H	M	M+H	M	M+H	M	M+H	M	M+H
n	237	193	201	154	237	193	237	193	237	193
mean	0.730	0.798	0.722	0.798	0.728	0.792	0.727	0.791	0.728	0.792
SE mean	0.015	0.016	0.017	0.016	0.016	0.016	0.016	0.016	0.020	0.016
QALY difference	0.054		0.060 (11)		0.061 (12)		0.061 (12)		0.061 (12)	
SE QALY difference	0.018		0.020 (12)		0.020 (10)		0.020 (10)		0.020(11)	
z for CCA and t for MI	2.970		2.950		3.020		3.020		3.000	
p-value	0.003		0.003		0.003		0.003		0.003	
95% CI lower limit	0.018		0.020		0.021		0.021		0.021	
95% CI upper limit	0.090		0.100		0.100		0.100		0.100	
35% missing data	M	M+H	M	M+H	M	M+H	M	M+H	M	M+H
n	237	193	163	122	237	193	237	193	237	193
mean	0.730	0.790	0.715	0.790	0.718	0.790	0.717	0.791	0.718	0.790
SE mean	0.015	0.018	0.020	0.018	0.017	0.016	0.017	0.016	0.017	0.016
QALY difference	0.054		0.068 (24)		0.069 (27)		0.071 (30)		0.069 (27)	
SE QALY difference	0.018		0.023 (27)		0.021 (17)		0.022 (18)		0.022 (20)	
z for CCA and t for MI	2.970		2.910		3.230		3.260		3.150	
p-value	0.003		0.004		0.001		0.001		0.002	
95% CI lower limit	0.018		0.022		0.027		0.028		0.026	
95% CI upper limit	0.090		0.113		0.111		0.113		0.112	
50% missing data	M	M+H	M	M+H	M	M+H	M	M+H	M	M+H
n	237	193	132	91	237	193	237	193	237	193
mean	0.730	0.782	0.717	0.782	0.705	0.785	0.708	0.784	0.706	0.784
SE mean	0.015	0.021	0.021	0.021	0.018	0.017	0.018	0.018	0.018	0.018
QALY difference	0.054		0.047 (13)		0.077 (41)		0.074 (36)		0.075 (38)	
SE QALY difference	0.018		0.026 (43)		0.024 (29)		0.024 (30)		0.024 (31)	
z for CCA and t for MI	2.970		1.820		3.260		3.110		3.140	
p-value	0.003		0.069		0.001		0.002		0.002	
95% CI lower limit	0.018		-0.004		0.031		0.027		0.028	
95% CI upper limit	0.090		0.098		0.123		0.120		0.122	

Note: M refers to the methadone maintenance treatment group; M+H refers to the group that had medical co-prescription of heroin. SE means standard error; QALY means Quality of life years gained; CI means confidence interval.

In the dataset with 35% missing data, the QALY coefficient in the CCA deviated the least and the most deviation occurred in the Log-MI-PMM, but the reference coefficient was still contained in all confidence intervals. The standard error of the CCA deviated the most from the standard error in the RA while the MI-PMM deviated the least. All strategies still showed co-prescribed heroin was associated with higher QALY scores compared to methadone maintenance. In the 50% missing dataset, the QALY coefficient deviated the most in the MI-PMM and the least in the CCA but the regression coefficient from the RA was still within all 95% confidence intervals. The standard error for the CCA deviated the most from the reference analysis, but the deviation in all MI strategies was similar. The CCA was the only strategy where the difference in QALYs was no longer statistically significant.

Cost-utility analysis

Figure 7.2 and Table 7.4 show the ICERs, INB, its variance, and the probability that co-prescribed heroin compared to methadone maintenance is cost-effective at a threshold value of 30,000 €/QALY for the 17%, 35% and 50% missing datasets. The CCA showed the largest deviation from the RA for the INB and for the standard error in the 17% missing data scenario. The INBs in the two-step MI strategy deviated the least from the INB in the reference analysis. The standard error deviated similarly for all imputation strategies. The probability of co-prescribed heroin compared to methadone maintenance being cost-effective was 99 percent for a willingness-to-pay threshold value of EUR 30,000 for one-unit gain in QALY score regardless of the imputation strategy. The ICER deviated the least from the RA in the CCA and the

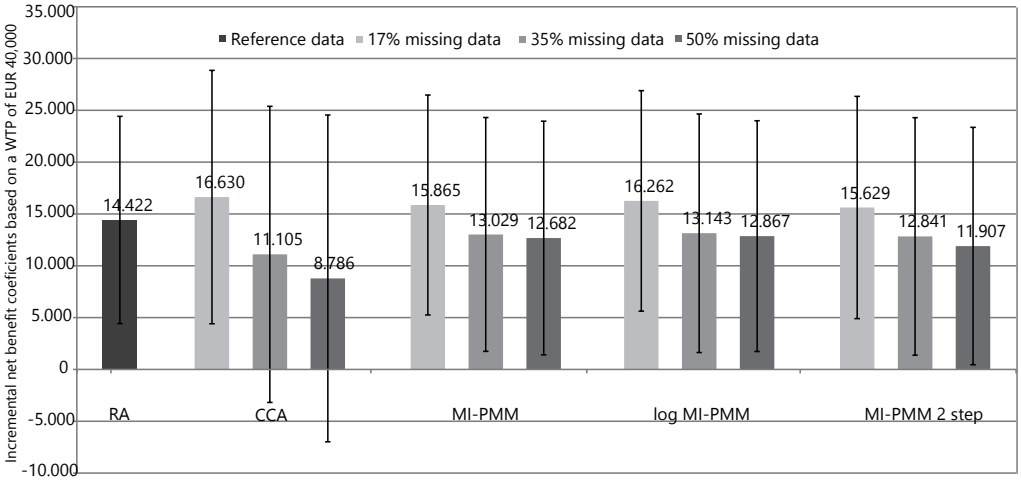


Figure 7.2. INB coefficients for a willingness-to-pay of EUR 30,000 based on the amount of missing data and imputation method

Table 7.4.
Cost effectiveness analysis estimates for the missing data methods

	RA	CCA (% Bias)	MI-PMM (% Bias)	Log-MI-PMM (% Bias)	Two-step MI (% Bias)
17% Missing data					
Correlation utility and costs	0.0507	0.0591	0.0517	0.0509	0.0487
Covariance	5.6	8.6	6.7	6.6	6.4
INB	14,422	16,026 (11)	15,257 (6)	15,654 (9)	15,023 (4)
SE INB	6,083	7,270 (20)	6,438 (6)	6,457 (6)	6,504 (7)
95% CI lower limit	4,417	4,069	4,669	5,034	4,324
95% CI upper limit	24,427	27,983	25,846	26,274	25,721
Prob C-E	0.99	0.99 (0)	0.99 (0)	0.99 (0)	0.99 (0)
ICER	-235,472	-235,448 (0)	-220,988 (6)	-227,410 (3)	-217,656 (8)
35% Missing data					
Correlation utility and costs	0.0507	0.0251	0.0300	0.0292	0.028
Covariance	5.6	4.9	4.4	4.4	4.4
INB	14,422	11,105 (23)	13,029 (10)	13,143 (9)	12,841 (11)
SE INB	6,083	8,685 (43)	6,864 (13)	7,000 (15)	6,966 (15)
95% CI lower limit	4,417	-3,181	1,738	1,629	1,383
95% CI upper limit	24,427	25,390	24,319	24,656	24,299
Prob C-E	0.99	0.90 (9)	0.97 (2)	0.97 (2)	0.97 (2)
ICER	-235,472	-134,488 (43)	-158,857 (33)	-156,289 (34)	-155,935 (34)
50% Missing data					
Correlation utility and costs	0.0507	0.0223	0.0433	0.0436	0.0406
Covariance	5.6	5.5	7.0	7.0	6.7
INB	14,422	8,786 (39)	12,682 (12)	12,867 (11)	11,907 (17)
SE INB	6,083	9,584 (58)	6,858 (13)	6,770 (11)	6,962 (14)
95% CI lower limit	4,417	-6,978	1,401	1,731	456
95% CI upper limit	24,427	24,551	23,962	24,003	23,358
Prob C-E	0.99	0.82 (17)	0.97 (2)	0.97 (2)	0.96 (3)
ICER	-235,472	-155,561 (34)	-134,979 (43)	-144,317 (39)	-128,670 (45)

Note: SE means standard error; INB mean Incremental net benefit; CI means confidence interval; Prob C-E means probability of cost-effectiveness; ICER mean Incremental cost effectiveness ratio.

most in the two-step MI. The reference value of INB was contained in the confidence intervals of all imputation strategies.

In the 35% missing data scenario, the CCA deviated the most from the RA for the ICER INB coefficient, the INB standard error, the probability that the intervention was cost effective. The MI-PMM deviated least from the RA for the INB standard error compared to the other imputation strategies. The probability of methadone plus heroin being cost-effective compared with methadone alone was 97% for a willingness-to-pay threshold value of EUR 30,000 for one-unit gain in QALY score for all multiple imputation strategies versus 99% for the RA (CCA was 90%).

In the scenario with 50% missings, the INB was no longer statistically significant for the CCA. The Log-MI-PMM showed the least deviation from the RA in the INB coefficient, the INB standard error and the probability that the intervention was cost effective. The probability of methadone plus heroin being cost-effective compared with methadone alone at 30,000 €/QALY was 97%.

For all imputation strategies, the reference INB was within the 95% confidence intervals (see Figure 7.2). In all strategies, the INB decreased with higher rates of missing data and the uncertainty was larger as evidenced by the larger standard errors and wider confidence intervals. This was most pronounced for the CCA where the INB was not statistically significant anymore with 35% and 50% of missing data.

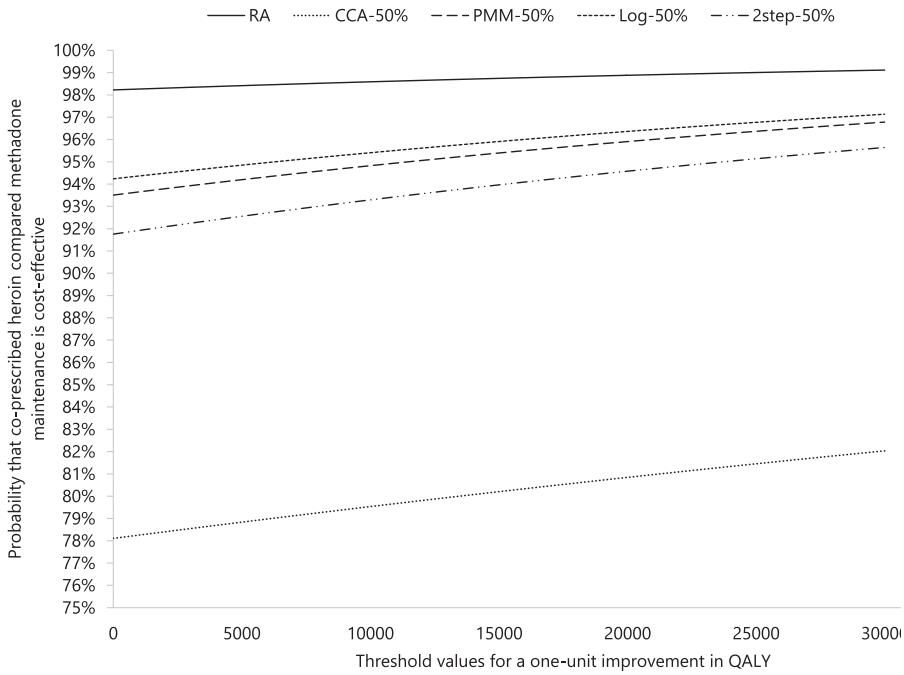


Figure 7.3. CEAC curves for a willingness to pay of EUR 30,000 in the 50% missing data condition.

The Log-MI-PMM showed the least uncertainty around INB in all missing data scenarios. Figure 7.3 presents the CEAC curves for the different strategies with 50% of missing data. This figure shows that there are pronounced differences between the strategies in this scenario. It shows that the probability that heroin plus methadone is cost-effective when the threshold value is zero is 98% for the reference analysis, 94% for the Log-MI-PMM and MI-PMM, 92% for the two-step-MI and 78% for the CCA. This increases to 99%, 97%, 97%,96% and 82% for the RA, MI-PMM, Log MI-PMM, two-step MI and CCA, respectively at a threshold value of 30,000 €/QALY.

Sensitivity analysis

For the MI-PMM and the log-MI-PMM the imputation procedure was applied to the total costs directly as well. The results confirmed that the precision was better when the imputation were applied to the sub-costs. This was reflected in smaller standard errors and decreased bias of the cost difference when applying multiple imputation to the sub-costs compared with imputation of the total costs variable (data not shown).

Discussion

Main findings

In this study, we evaluated the performance of different multiple imputation strategies and CCA for scenarios with varying rates of missing data in costs and effects in a pragmatic economic evaluation. We found that for all rates of missing data, multiple imputation strategies performed better than CCA. The results of the CCA, MI-PMM and the two-step MI were all influenced by the amount of missing data. With a larger amount of missing data, the Log MI-PMM deviated the least from the RA for the cost difference, cost standard error, INB estimate, the INB standard error and the probability that the heroin plus methadone treatment was cost effective in comparison with methadone only all at a willingness to pay of 30,000 euro per QALY. Therefore the Log MI-PMM is considered most appropriate to use to impute missing cost and effect data. However, when looking at QALYs the MI-PMM performed best since it deviated the least from the RA with increasing amounts of missing data. In general, the Log-MI-PMM was least affected by the amount of missing data.

Our results imply that addressing only the right-skewness of the data by using a log transformation in combination with PMM is enough and that strategies to deal with zero inflation such as our two-step PMM are not needed. Beforehand, we expected that the two-step MI strategy would have performed better because it controls for the large amount of zeroes and the skewness in the data. However,

in practice there were no relevant differences with the other multiple imputation strategies and the two-step MI was more difficult to apply than the Log MI-PMM. Not all software packages have incorporated a comprehensive way to apply the two-step MI strategy, whereas the Log MI-PMM is easily applied and available in software packages like SPSS, Stata, SAS and R.

Comparison with existing literature

Our study adds to the findings from other studies that multiple imputation is better than CCA to deal with missing data in economic evaluations (Briggs et al., 2003; Burton, Billingham, & Bryan, 2007; Oostenbrink et al., 2003; Yu, Burton, & Rivero-Arias, 2007). However, in contrast to Briggs et al. (2003), Oostenbrink et al. (2003) and Burton et al. (2007), we had information on the observed values of the missing data, because we created the missing data ourselves using the MAR assumption. This allowed us to estimate the deviation of the different imputation strategies from the original complete dataset.

Yu et al. (2007) showed in a simulation study that predictive mean matching in R and STATA performed reasonably well and maintained the underlying distribution of the resource use data (Yu et al., 2007). However, they did not evaluate the effect of the different imputation strategies on the cost-effectiveness estimates.

Strengths and limitations

Our study adds to these studies by focusing on estimation of both incomplete costs, utilities and cost-effectiveness and by comparing different MI strategies using MICE with PMM in STATA. Additionally we use a correlation after multiple imputation between costs and utilities using Fisher's Z transformation to calculate the cost-effectiveness (Schafer, 1997; van Buuren, 2012). We used MICE with PMM which gave us more flexibility around assumptions of normality (van Buuren, 2007).

The strengths of this study were its systematic and applied approach using real data to examine the performance of different multiple imputation strategies in situations with varying amounts of missing data. To our knowledge, this is one of the first studies to compare the two-step MI strategy with other multiple imputation strategies for cost-effectiveness evaluations.

As we used only one dataset we were limited in our evaluation parameters for direct comparisons to the true coefficients instead of averages over simulations. We did perform a small simulation pilot study repeating the imputation procedures to verify the stability of the methods. This was done by repeatedly drawing samples of 100 cases from each of our incomplete datasets and applying our method to these small samples. We simulated 1000 times and used a smaller number of imputations and iterations: 15 and 20 respectively. For each method and incomplete data condition

the average over the 1000 simulations was taken and compared to the complete reference data results. This simulation confirmed the relative differences between the performance of the methods presented in this study. Future research should perform a larger simulation study. It would be interesting to vary the proportion of zeroes in a future study and see how that affects the performance of the missing data methods. It is possible that with a greater amount of zeroes the two part model could be more beneficial over the other methods. In all of our scenarios, we assumed the same missing mechanism in both treatment arms, and in future simulations this probably should be changed for some simulated data. Future research should test our strategies on other datasets to confirm our results.

Implications for further research

Prospective economic evaluations alongside trials play an important role in providing decision makers cost-effectiveness information to inform reimbursement decisions. Therefore, it is important that economic evaluations provide robust and unbiased information. The consequences of using different imputation strategies can affect policy decisions. In this study, we considered heroin plus methadone treatment to be cost-effective in comparison with methadone alone in all strategies evaluated although the uncertainty increased. However, in situations with smaller differences between groups, the decision may change depending on the imputation procedure chosen.

In conclusion, we recommend the use of the Log MI-PMM because of its ease to use and its reliable results especially with increased amounts of missing data. Log MI-PMM also appears to perform well for zero-inflated data as long as a constant is used in place of the zero in the data.

References

Briggs, A., Clark, T., Wolstenholme, J., & Clarke, P. (2003). Missing... presumed at random: cost-analysis of incomplete data. Health economics, 12(5), 377-392.

Brooks, R. (1996). EuroQol: the current state of play. Health policy, 37(1), 53-72.

Burton, A., Billingham, L. J., & Bryan, S. (2007). Cost-effectiveness in clinical trials: using multiple imputation to deal with incomplete cost data. Clinical Trials, 4(2), 154-161.

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. Statistical Methods in Medical Research, 16(3), 219-242.

van Buuren, S. (2012). Flexible Imputation of Missing data. New York: Chapman & Hall/CRC.

Dijkgraaf, M. G., van der Zanden, B. P., de Borgie, C. A., Blanken, P., van Ree, J. M., & van den Brink, W. (2005). Cost utility analysis of co-prescribed heroin compared with methadone maintenance treatment in heroin addicts in two randomised trials. BMJ, 330(7503), 1297.

Dolan, P. (1997). Modeling valuations for EuroQol health states. Medical care, 35(11), 1095-1108.

Eekhout, I., de Vet, H. C., Twisk, J. W., Brand, J. P., de Boer, M. R., & Heymans, M. W. (2013). Missing data in a multi-item instrument were best handled by multiple imputation at the item score level. Journal of clinical epidemiology.

Enders, C. K. (2010). Applied Missing Data Analysis. New York Guilford Press.

Horton, N. J., & Lipsitz, S. R. (2001). Multiple Imputation in Practice: Comparison of Software Packages for Regression Models with Missing Variables. The American Statistician, 55(3), 244-254.

Javaras, K. N., & Van Dyk, D. A. (2003). Multiple imputation for incomplete data with semicontinuous variables. Journal of the American Statistical Association, 98(463), 703-715.

Klebanoff, M. A., & Cole, S. R. (2008). Use of multiple imputation in the epidemiologic literature. American journal of epidemiology, 168(4), 355-357.

Lambert, P. C., Billingham, L. J., Cooper, N. J., Sutton, A. J., & Abrams, K. R. (2008). Estimating the cost-effectiveness of an intervention in a clinical trial when partial cost information is available: a Bayesian approach. Health economics, 17(1), 67-81.

Lee, K. J., & Carlin, J. B. (2010). Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. American journal of epidemiology, 171(5), 624-632.

NICE. (2010). Measuring effectiveness and cost effectiveness: the QALY. Retrieved 26.09.2013, 2013, from http://www.nice.org.uk/newsroom/features/measuringeffectivenessandcosteffectiveness theqaly.jsp.

Nietert, P. J., Wahlquist, A. E., & Herbert, T. L. (2013). Characteristics of recent biostatistical methods adopted by researchers publishing in general/internal medicine journals.

Statistics in medicine, 32(1), 1-10.

Nixon, R. M., Wonderling, D., & Grieve, R. D. (2010). Non-parametric methods for cost-effectiveness analysis: the central limit theorem and the bootstrap compared. Health economics, 19(3), 316-333.

Oostenbrink, J. B., Al, M. J., & Rutten-van Molken, M. P. (2003). Methods to analyse cost data of patients who withdraw in a clinical trial setting. PharmacoEconomics, 21(15), 1103-1112.

R Core Development Team. (2014). R: A language and environment for statistical computing. Vienna, Austria.

Rubin, D. B. (1976). Inference and missing data. Biometrika, 63(3), 581-592.

Schafer, J. L. (1997). Analysis of Incomplete Multivariate Data. London: Chapman & Hall.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ, 338, b2393.

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. Statistics in Medicine, 30(4), 377-399.

Willan, A. R., & Briggs, A. H. (2006). Statistical Analysis of Cost-effectiveness Data: John Wiley & Sons Ltd.

Yu, L. M., Burton, A., & Rivero-Arias, O. (2007). Evaluation of software for multiple imputation of semi-continuous data. Statistical Methods in Medical Research, 16(3), 243-258.

Acknowledgements

The original trial was commissioned and financially supported by the Netherlands Ministry of Health, Welfare, and Sports.

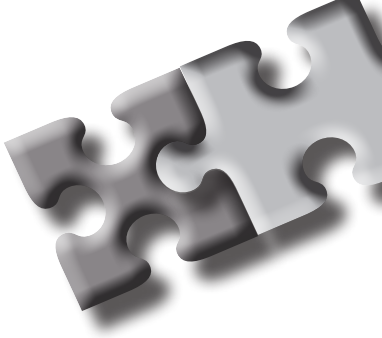
Appendix 7.1| Descriptive statistics of the cost variables for the incomplete datasets

Table 1.
Descriptive statistics of the cost variables for the reference dataset and the datasets with missing values.

	QALY (Dolan, 1997)	Travel costs	Program costs	Out-of-hospital consultations costs	Inpatient hospital consultations costs
Reference					
mean	0.76	350.04	8,692.99	45.97	316.12
sd	0.22	292.79	10,315.04	124.06	1,206.92
% of zeroes	N/A	3.3	0.2	67.4	56.7
Skewness	-1.28	0.90	0.93	4.81	10.39
18% missing					
mean	0.75	350.04	8,266.09	43.89	323.21
sd	0.23	292.79	10,022.88	109.55	1,267.93
% of zeroes	N/A	3.3	0.3	67.2	58.1
Skewness	-1.30	0.90	0.99	3.95	10.01
35% missing					
mean	0.75	350.04	8,218.24	44.99	336.41
sd	0.23	292.79	10,075.05	112.93	1,318.40
% of zeroes	N/A	3.3	0.3	66.0	57.4
Skewness	-1.23	0.90	1.02	4.00	9.69
50% missing					
mean	0.74	350.04	7,907.28	49.09	338.09
sd	0.23	292.79	9,995.16	118.29	1,357.34
% of zeroes	N/A	3.3	0.3	63.7	57.7
Skewness	-1.20	0.90	1.11	3.82	9.60

Table 2.
Descriptive statistics of the cost variables for the reference dataset and the datasets with missing values.

	Inpatient hospital stay costs	Police investigations costs	Convicting criminal offenders costs	Sanctioning criminal offenders costs	Damage to victims costs
Reference					
mean	778.96	6,004.28	3,172.95	1,854.53	23,602.36
sd	3,680.04	14,497.13	12,627.55	5,935.78	59,945.50
% of zeroes	91.6	58.4	88.8	77.2	71.2
Skewness	7.32	7.10	6.14	4.20	3.17
18% missing					
mean	727.68	6,304.52	3,260.05	1,751.09	24,391.48
sd	3,446.62	15,219.27	12,995.47	5,797.58	61,348.19
% of zeroes	91.8	58.9	88.9	77.8	71.0
Skewness	7.36	6.81	6.06	4.45	3.14
35% missing					
mean	794.57	6,462.09	3,027.35	1,731.97	24,611.35
sd	3,711.62	15,665.51	12,569.37	5,783.11	62,221.27
% of zeroes	91.6	58.8	89.6	78.1	70.7
Skewness	7.09	6.75	6.40	4.56	3.17
50% missing					
mean	853.85	6,523.91	2,586.35	1,749.48	23,280.58
sd	3,972.24	15,794.79	11,541.70	5,873.06	60,485.85
% of zeroes	91.0	57.7	90.5	78.2	72.5
Skewness	7.00	6.99	7.31	4.58	3.27



Chapter 8

General discussion

Under review as discussion section of a review article: Eekhout, I., de Vet, H.C.W., de Boer, M.R., Twisk, J.W.R., Heymans, M.W. Missing data in multi-item questionnaires: analyze carefully and don't waste available information. International Journal of Epidemiology.

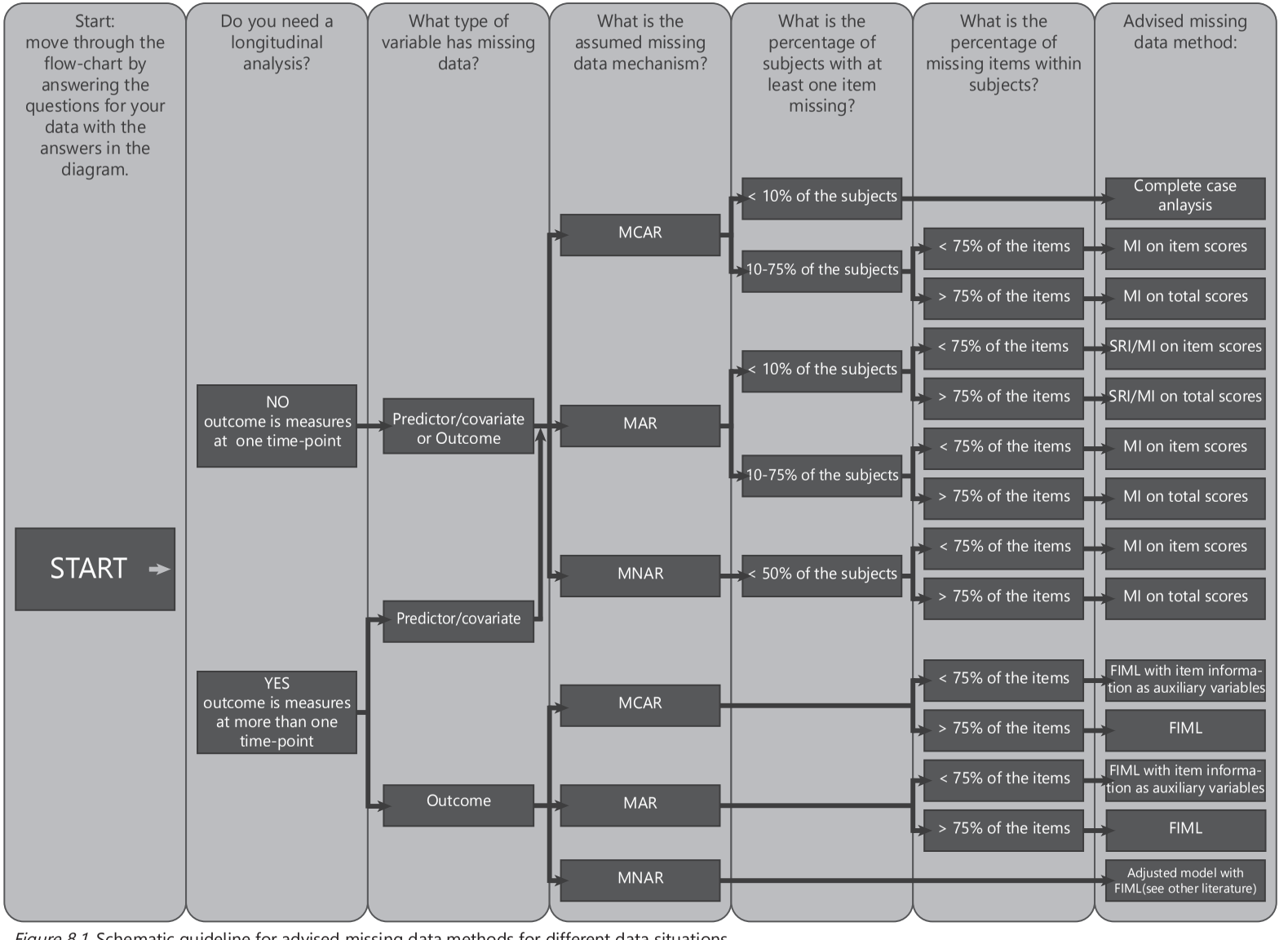


Figure 8.1. Schematic guideline for advised missing data methods for different data situations.

General advice

The best method for dealing with missing data in multi-item questionnaires depends on many aspects, for example the study design, the analysis method and type of the missing data. Missing data in a predictor in a cross-sectional study should be handled differently than missing data in the outcome of a longitudinal study. For that reason there is not one optimal solution for handling missing data in a multi-item questionnaire, but depending on the various facets about the missing data there are different missing data methods to use. An overview of data situations and corresponding missing data methods is given in Figure 8.1. The practical guide aims to help researchers find an optimal solution to their specific missing data problem.

In general there are several characteristics of the data analysis and missing data that have to be taken into account. These are the analysis method that is applied to analyze the dataset (i.e., longitudinal analysis or not), the type of variable with missing data (i.e., predictor/covariate or outcome), the missing data mechanism (i.e., MCAR, MAR or MNAR), the overall percentage of subjects with at least one item missing, and the level of missing data in the questionnaire (i.e., missing item scores or missing total scores). When less than 75% of the item scores are missing within subjects, the missing data is considered to be at the item score level and when more than 75% of the item scores are missing within subjects, the missing data is considered to be at the total score level.

For data analysis of studies with outcomes at only one time-point the type of variable in the analysis model that contains missing data is of minor importance. In general whether the missings are in the outcome, in the predictor or covariate of the analysis, the solution is similar. In these studies with the predictors and outcome data collected at one time-point, it is of primary importance whether the missing data are MCAR, MAR or MNAR.

If the missing data are MCAR, the percentage of missing data plays an important role. If the percentage of subjects that have missing data on at least one item is low (i.e., < 10%), then the missing data will have a minor effect on the study results and a complete-case analysis can be performed. However, when the percentage of subjects who have missing data on at least one item is larger (i.e., 10%-75%), it will be more important to sustain statistical power. In that case it is advised to perform an imputation method to be able to use all the available data. In that situation it is advised to apply multiple imputation to the missing item scores, if <75% of item scores are missing within subjects, or multiple imputation applied to the total scores if >75% of item scores are missing within subjects (Eekhout et al., 2014).

For studies with one outcome that have MAR data for a small percentage of data (i.e., < 10% of the subjects have at least one item missing) applying a single stochastic regression imputation could suffice. Although multiple imputation would be most

optimal in these situations, stochastic regression imputation results in unbiased results when data are MAR for a small percentage of subjects with missing data. In stochastic regression imputation, the uncertainty about the missing data is not incorporated in the missing data method. When the percentage of subjects with at least one item missing is small (i.e., <10%) the effect of not including this uncertainty is negligible (Eekhout et al., 2014). When the percentage subjects with missing data is larger (i.e., 10%-75%), multiple imputation is advised. When the items scores are missing (i.e., <75% of the item scores missing within subjects), the imputation (i.e., single stochastic regression imputation or multiple imputation) should be applied to the item scores first, prior to computing the total score for analysis. When large part of the questionnaire was not filled out by the study participants (i.e., >75% of the item scores missing within subjects), so total score level data are missing, the imputation method should be directly applied to the total scores. In datasets where some study participants have item scores missing and some have total score missings, the imputation method should first be applied to the item scores and subsequently to the total scores for the subjects who have the total score missing data (Eekhout et al., 2014).

When the missing data in studies with one outcome are MNAR and the percentage of subjects with missing data is not too large (i.e., <50%), the missings on the item scores (i.e., <75% of the item scores missing within subjects) should be handled by multiple imputation of the items and the missings on the total scores (i.e., >75% of the item scores missing within subjects) by multiple imputation of the total scores. When data are MNAR and more than 50% of subjects have missing data, multiple imputation is not a reliable solution (Eekhout et al., 2014).

In data that require a longitudinal analysis strategy, so when outcomes at more time-points are analyzed, missings in the predictors or covariates need to be handled differently than missings in the outcome. The missing data in the predictors or covariates should be handled the same way as the missing data in studies that have the outcome measured at one time-point.

The missing data in the outcome of a longitudinal study can be handled within the analysis, when a method based on full information maximum likelihood estimation (e.g., mixed model or structural equation model) is used. However, for participants that have all outcomes missing at all the time-points, the missing data cannot be handled and these participants will not be analyzed. When the missing outcome data are MCAR or MAR, the longitudinal methods based on full information maximum likelihood estimation are advised. In situations where the outcomes are missing due to missing item scores, the observed item information should be included in the model as auxiliary variables. When longitudinal data are MNAR, the missing data can be handled in a MNAR model. In a MNAR model there is a relation between the

probability of missing data and the outcome. Two common MNAR models are the selection model and the pattern mixture model. An explanation and evaluation of these models is beyond the scope of this thesis, but are described in other literature (Enders, 2011b; Molenberghs, Thijs, Kenward, & Verbeke, 2003).

In many empirical data situations, the missing data are neither only at the item score level nor only at the total score level. Mostly a dataset contains missing data at both levels. In the diagram (Figure 8.1) a solution is indicated for each situation. In practice, for the subjects who have missing data at the item level (i.e., <75% of the items missing), the multiple imputation procedure should be applied to the items. Simultaneously, the subjects who have missing total scores (i.e., >75% of the items missing) should have their total score imputed. In many software packages it is quite complicated to do this simultaneously.

As a practical solution, there are four steps that can be taken to do this in an appropriate manner in SPSS. (1) Two copies of the dataset can be used. (2) In one copy all data can be imputed at the item level and after the imputation procedure, the total scores should be calculated using the imputed items. (3) In the second copy, the total scores can be calculated prior to the imputation. These total scores are left incomplete when one or more items are missing. Subsequently, the multiple imputation can be applied to the total scores directly. (4) After the imputation procedures are performed, the total scores from the imputed items and the imputed total scores can be merged into one dataset. Then the total scores from the imputed items should be selected for the subjects who had less than 75% of the item scores missing, and the imputed total scores for the subjects with more than 75% of the items missing. After this procedure the regular MI analysis phase and pooling phase can be performed using the merged total scores in the analysis.

The practical guide presented in Figure 8.1 includes no condition where more than 75% of the subjects have missing data. Generally, when more than 75% of the data contains missing values it might not be wise to analyze the data. Although there are situations imaginable where the missing data can be handled adequately and valid analysis results can be obtained. For example in a dataset in which 85% of the subjects have item scores missings, but in this data only a small percentage of subjects have the total scores missing (i.e., <10% of the subjects have >75% of the item scores missing), the other subjects with missing data have less than 75% of the item scores missing. The total scores that will be calculated will be missing for 85% of the subjects, because the totals score will be missing when one or more item scores are missing. However, the fraction of missing information will be smaller than 85%, because the available item score information contains information about the missing data in the total score (Graham, Olchowski, & Gilreath, 2007). In that situation, multiple imputation applied to the item scores might result in valid study results. This

means that the performance of missing data methods as multiple imputation and full information maximum likelihood is not necessarily directly related to the percentage of subjects who contain missing data, but more so to the fraction of missing information (Schafer, 1997). The fraction of missing information (FMI) represents the amount of missing information available to estimate parameters (Rubin, 1987). Theoretically the FMI is as large as the total percentage of missing data, however, this value is reduced by auxiliary variables that can include additional information about the missing data into the analysis or missing data handling (Graham, 2012).

The guidelines on percentages of missing data in the practical guide are recommendations. However, depending on the missing data patterns and location of missing data on the multi-item questionnaire (i.e., missing item scores or missing total scores), the advanced missing data methods might also perform effectively in situations where more than 75% of the subjects have some missing data.

Methodological considerations

Multiple imputation versus full information maximum likelihood

Both multiple imputation and full information maximum likelihood are currently considered to be the state-of-the-art methods for missing data handling (Schafer & Graham, 2002). Both methods can handle missing data in studies with the outcome measured at one time-point or with the outcome measured more than once (Baraldi & Enders, 2010). However, the advice in this thesis is focused on the most optimal and practical methods for epidemiological researchers and therefore one of the two methods is recommended in each study design. In studies that assess information at one time-point, the analysis method might be preferred to be kept simple and straight forward and therefore handle the missing data with multiple imputation. However, if in these studies the missing data would be handled by full information maximum likelihood, the analysis should be specified in for example a structural equation model. This would require additional knowledge of structural equation modeling to accommodate the missing data, while in multiple imputation the analysis method that was planned at the design stage of the study can be applied after the multiple imputation procedure. Whereas, epidemiologists who perform longitudinal analysis might be familiar with methods based on full information maximum likelihood estimation (Twisk, 2013). In that case, it might be more practical for them to handle missing data in studies with outcomes at multiple time-points with full information maximum likelihood than with multiple imputation.

If the missing data are only in the outcome, handling the missing data by multiple

imputation or by full information maximum likelihood in a longitudinal model will yield similar results when the variables in the imputation model are the same as the variables in the longitudinal model (Collins, Schafer, & Kam, 2001; Schafer, 2003). Handling the missing data in the model directly, as is done in full information maximum likelihood, will be more feasible (Enders, 2011a). However, some researchers point out that multiple imputation is the preferred strategy to include auxiliary variables to make the MAR assumption more plausible and therefore prefer handling missing data via this approach when auxiliary information is available (Bell & Fairclough, 2013). Nevertheless, in structural equation models it is possible to include auxiliary variables, without changing model interpretations. Software programs as Mplus accommodate this method and in these programs the inclusion of auxiliary variables to handle missing data is relatively easy (B. O. Muthén, Asparouhov, Hunter, & Leuchter, 2010; L. K. Muthén & Muthén, 1998-2012).

Multiple imputation in practice

As previously mentioned, in multiple imputation many different methods are available to adapt the imputation algorithm to the assumed distribution of the data. Item scores in a multi-item questionnaire are frequently measured by a Likert scale. These are ordinal items which are not necessarily normally distributed, and incomplete ordinal data might best be imputed with the proportional odds model. However, in a simulation study was found that the distribution of the items did not limit the performance of the linear regression algorithm of multiple imputation (Eekhout et al., 2014). Furthermore, the predictive mean matching procedure is more robust against the deviations from the normal distribution and imputes more realistic values compared with linear regression imputation. For that reason, predictive mean matching might be attractive for imputing categorical item scores; however the linear regression algorithm performed just as well as predictive mean matching when final analysis results were evaluated (Eekhout et al., 2014).

Another example where the distribution of the data deviates from normality is cost-data. The total costs in a study are often the sum of several sub-costs. The relation between the sub-costs and the total costs is not reflective, however also in the cost data it was most feasible to apply the missing data method to the sub-costs instead of to the total costs directly. This is in concordance with the advice with respect to multi-item questionnaires, where the missing data need to be handled at the item level. In studies where costs are measured for economic evaluations, a part of the sample have zero costs, but the participants that actually have costs sometimes have very large cost values. Furthermore, costs cannot become negative. So, the distribution of cost data is not normal and might require different methods for multiple imputation. It is possible to transform the data with a log to obtain a (close

to) normal distribution. Alternatively one can use a method that imputes the missing cost data in two separate steps. In this method first a value for having costs versus not having costs is imputed by a logistic method and in the second step the people that are indicated to have costs will have their costs imputed by predictive mean matching. Another option is to use predictive mean matching as a method without the first step. It might be expected that a method that takes into account all aspects of the distribution would perform best. Nevertheless, in a study that was conducted on multiple imputation of cost data (MacNeil-Vroomen et al., under review) the two step method did not perform better than predictive mean matching without the first step. Moreover, simply log-transforming the data and imputing that distribution was the most stable solution in larger percentages of missing data.

The application of multiple imputation to the item scores can pose some problems in some study designs. The basic rule for the construction of the imputation model is to include all relevant information about the analysis model in the imputation model, together with the information relevant to the missing data handling. This includes all variables used in the main analysis and auxiliary variables. However, when many multi-item questionnaires are administered in one study the imputation model might become extremely large and even make model estimations impossible. As a solution it is possible to construct the imputation model in such a way that for each separate questionnaire the item scores are imputed using the total scores from the other questionnaires as predictors. These total scores from the other questionnaires are calculated from the imputed item scores of that questionnaire. This strategy, called passive imputation, is further explained in a simulation study that evaluated this method (Eekhout, De Vet, De Boer, Twisk, & Heymans, under review).

Missing Not At Random

The assumption about the missing data mechanism is very important in order to select a valid method to handle the missing data. Unfortunately, it is not possible to distinguish between MAR and MNAR mechanisms, because the missing values are unknown. Brand (1999) describes in Chapter 2 of his dissertation two examples that demonstrate how an initially MNAR missing data mechanism can change into MAR by including additional variables that are related to the probability of missing data. In practice, by including variables related to the probability of missing data a MNAR mechanism can get closer to MAR. Accordingly, the MAR assumption can be made more plausible by including auxiliary information in the missing data handling method (Baraldi & Enders, 2010). Furthermore, it is often advised to do sensitivity analysis by applying additional MNAR models (e.g., selection models or pattern mixture models (Enders, 2011b; Molenberghs et al., 2003)) and examine if the conclusions change (Enders, 2011a).

Future research

This thesis focuses on handling missing data in the total score that are caused by missing item scores. Several solutions are offered, however the solutions presented here might not be the only valid options to handle missing item score data. It might be interesting to compare the methods that were proposed here, multiple imputation applied to the items or including the item scores as auxiliary variables in a full information maximum likelihood analyses, to methods that use the item scores in the analysis model directly. This can be accomplished by including the item scores as indicators for a latent variable in a structural equation model or by using another latent variable model that is robust against missing data, such as item response techniques. It might be interesting to study in which situations it is preferred to use the item scores in the analysis. For example, the internal consistency of the multi-item questionnaire (i.e., coherence of the items) might be related to the performance of such methods. Also in this context, the possibilities of using the item scores in the analysis to handle missing data in epidemiological studies that measure at one time-point can be further explored and compared to applying multiple imputation.

The studies about missing data in multi-item questionnaires referred to in this thesis are about questionnaires with a reflective model. In a reflective model the change in the construct (e.g., better physical functioning) is reflected by changes in the items (e.g., higher scores on the items). In this setting the construct is measured indirectly by the items. In questionnaires with a formative model the construct is more like an index score, for example food intake measured by a food frequency questionnaire. In this case the items can be an extensive list of food items that together form tot total food intake. In a formative model the items are mostly not as highly correlated and a change in the construct (e.g., food intake) is not necessarily reflected by a change in all of the items (de Vet, Terwee, Mokkink, & Knol, 2011). The investigated methods for missing data have not been extensively evaluated in multi-item questionnaires with a formative model. However, it may be expected that the advice formulated for questionnaires with a reflective model (i.e., multiple imputation at the item level and full information maximum likelihood with auxiliary item information) will apply to these questionnaires as well. Nevertheless, the distributions of the items may be very skewed or zero-inflated in some of the formative questionnaires due to the count nature of the items. This might require different approaches that take account of this distribution as investigated in some studies (e.g., Fraser et al., 2009; Nevalainen, Kenward, & Virtanen, 2009; Parr et al., 2008). This needs to be investigated in future research.

The relation between the fraction of missing information and the performance of missing data methods has been previously studied in non-questionnaire data. Most studies about fraction of missing information were aimed at the required number of

imputations for multiple imputation (e.g., Bodner, 2008; Graham et al., 2007; Schafer, 1997). The fraction of missing information in multi-item questionnaire data and the consequences for the missing data methods might be a relevant tool to analyze the performance of missing data handling for item scores. This should be explored in future research.

In longitudinal studies, where the outcomes are measured more than once, the missing data in the predictors or covariates can be handled by multiple imputation. It might be interesting to study the possibilities of handling the missing data in the predictors in a structural equation model. It can also be useful to further study the applications of MNAR models when total scores are incomplete due to missing item scores. And further, to what extent the items can be included as auxiliary variables in these models and whether this improves model estimates, might be an interesting and useful focus for future studies.

Conclusion

This thesis provides a practical guide on how to handle missing data in multi-item questionnaires. In Box 8.1 an overview of what is known, what is new and future challenges are summarized as the key messages of this thesis. Overall, it is important to incorporate all available information from the item scores in order to obtain the optimal level of accuracy and precision.

What is known:

- Missing data can cause biased results when the missings are not missing completely at random.
- Missing data in multi-item questionnaires can be handled at the item level or at the total score level.
- Multiple imputation and Full information maximum likelihood estimation methods are the advanced state-of-the-art missing data methods.

What is new:

- Methods that are advised in manuals for multi-item questionnaires are often sub-optimal and should be ignored.
- Missing data in multi-item questionnaire 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.
- 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.
- 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.

Future challenges:

- To investigate the relation between the fraction of missing information and the amount of missing item scores and the performance of missing data methods.
- Comparing the use of missing data methods to handle the missing item scores, to advanced methods that include the item scores in the main analysis, for example when missings are in the predictor.
- Incorporating the observed item scores in longitudinal models that correct for MNAR data.

Box 8.1. Key messages of this thesis

References

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. Journal of School Psycholology, 48(1), 5-37.

Bell, M. L., & Fairclough, D. L. (2013). Practical and statistical issues in missing data for longitudinal patient reported outcomes. Stat Methods Med Res, 19, 19.

Bodner, T. E. (2008). What Improves with Increased Missing Data Imputations? Structural Equation Modeling: A Multidisciplinary Journal, 15(4), 651-675.

Brand, J. P. L. (1999). Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets. . Dissertation, Rotterdam: Erasmus University Rotterdam.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological Methods, 6(4), 330-351.

Eekhout, I., De Vet, H. C., De Boer, M. R., Twisk, J. W., & Heymans, M. W. (under review). Passive imputation of missing values in studies with many multi-item questionnaire outcomes. American Journal of Epidemiology.

Eekhout, I., De Vet, H. C. W., Twisk, J. W. R., Brand, J. P. L., De Boer, M. R., & Heymans, M. W. (2014). Missing data in a multi-item instrument were best handled by multiple imputation at the item score level. Journal of Clinical Epidemiology, 67(3), 335-342.

Enders, C. K. (2011a). Analyzing longitudinal data with missing values. Rehabilitation Psychology, 56(4), 267-288.

Enders, C. K. (2011b). Missing not at random models for latent growth curve analyses. Psychol Methods, 16(1), 1-16.

Fraser, G. E., Yan, R., Butler, T. L., Jaceldo-Siegl, K., Beeson, W. L., & Chan, J. (2009). Missing data in a long food frequency questionnaire: are imputed zeroes correct? Epidemiology, 20(2), 289-294.

Graham, J. W. (2012). Missing data analysis and design. New York: Springer.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. Prev.Sci., 8(3), 206-213.

MacNeil-Vroomen, J., Eekhout, I., Dijkgraaf, M. G., Van Hout, H., De Rooij, S. E., Heymans, M. W., & Bosmans, J. E. (under review). Comparing multiple imputation strategies for zero-inflated cost data in economic evaluations: which method works best? European Journal of Health Economics.

Molenberghs, G., Thijs, H., Kenward, M. G., & Verbeke, G. (2003). Sensitivity Analysis of Continuous Incomplete Longitudinal Outcomes. Statistica Neerlandica, 57(1), 112-135.

Muthén, B. O., Asparouhov, T., Hunter, A., & Leuchter, A. (2010). Growth Modeling with Non-Ignorable Dropout: alternative analysis of the STAR*D antidepressant trial. February version wide.

Muthén, L. K., & Muthén, B. O. (1998-2012). Mplus User's Guide. Seventh Edition. Los Angeles, CA: Muthén & Muthén.

Nevalainen, J., Kenward, M. G., & Virtanen, S. M. (2009). Missing values in longitudinal dietary data: a multiple imputation approach based on a fully conditional specification. Stat.Med., 28(29), 3657-3669.

Parr, C. L., Hjartaker, A., Scheel, I., Lund, E., Laake, P., & Veierod, M. B. (2008). Comparing methods for handling missing values in food-frequency questionnaires and proposing k nearest neighbours imputation: effects on dietary intake in the Norwegian Women and Cancer study (NOWAC). Public Health Nutr, 11(4), 361-370.

Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: J. Wiley & Sons.

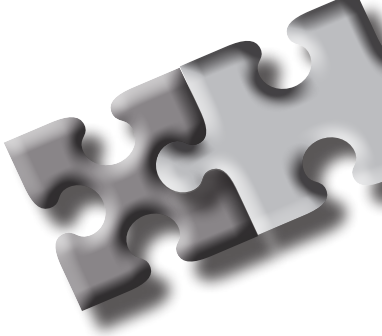
Schafer, J. L. (1997). Analysis of Incomplete Multivariate Data. London, UK: Chapman & Hall.

Schafer, J. L. (2003). Multiple Imputation in Multivariate Problems When the Imputation and Analysis Models Differ. Statistica Neerlandica, 57(1), 19-35.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. Psychol. Methods., 7(2), 147-177.

Twisk, J. W. R. (2013). Applied Longitudinal Data Analysis for Epidemiology: A Practical Guide, Second Edition. New York: Cambridge University Press.

de Vet, H. C. W., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011). Measurement in Medicine. Cambridge: Cambridge University Press.



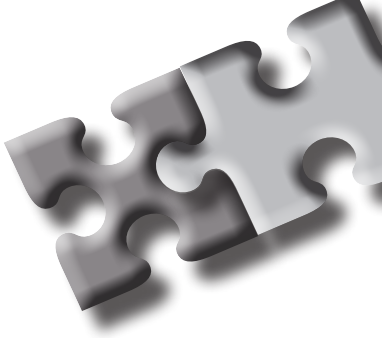
Abbreviations and symbols

Abbreviations

Aux	Auxiliary
CCA	Complete case analysis
CEAC	Cost-effectiveness analysis curve
CES-D	Center for Epidemiological Studies Depression Scale
CI	Confidence interval
FIML	Full Information Maximum Likelihood
FMI	Fraction of missing information
ICER	Incremental cost-effectiveness ratio
INB	Incremental net benefit
MAR	Missing at random
MCAR	Missing completely at random
MI	Multiple imputation
MICE	Multivariate imputation by chained equations
MNAR	Missing not at random
MSE	Mean squared error
PCI	Pain coping inventory
PMM	Predictive mean matching
PO	Proportional odds model
Prob C-E	Probability of cost-effectiveness
QALY	Quality of life years
RA	Reference analysis
RCT	Randomized controlled trial
sd	Standard deviation
SE	Standard error
SRI	Single stochastic regression imputation
STROBE	STrengthening the Reporting of OBservational studies in Epidemiology
TS	Total score
WTP	Willingness to Pay

Symbol description

θ	Parameter
θ_j	Parameter from imputed dataset j
$\theta \text{ or } \bar{\theta}$	Average parameter for the imputed datasets (pooled parameter)
m	Number of imputations in multiple imputation
n	Sample size
Var	Variance
$\beta \text{ or } \beta_i$	Parameter, usually regression coefficient
$\hat{\beta} \text{ or } \hat{\beta}_i$	Parameter estimate
$\bar{\hat{\beta}}$	Average parameter estimate
$\bar{\beta}_c$	Average complete (true) parameter
ζ	Residual variance
$\hat{\Delta}_c$	The difference in total costs between two groups
$\hat{\Delta}_e$	The difference in QALY between two groups
λ	The willingness to pay
$\hat{C}$	Covariance estimate



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.

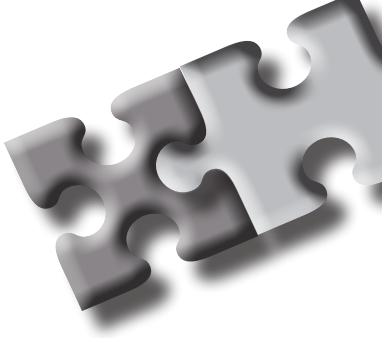
When the outcome is measured at multiple time-points, the analysis method should take the correlation between the multiple measurements into account. Longitudinal analysis methods often use full information maximum likelihood procedures to obtain the parameter estimated and these procedures handle the 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.



Samenvatting

Nederlands

Mis het niet!
Incomplete data kan waardevolle informatie bevatten

In epidemiologisch onderzoek wordt veel gebruik gemaakt van vragenlijsten om data te verzamelen. Deze vragenlijsten meten vaak een bepaald onderliggend construct door de scores op meerdere losse items (i.e., vragen) op te tellen tot een totaal score. Doordat een of meerdere vragen niet zijn ingevuld of doordat de gehele vragenlijst niet is ingevuld, kunnen de vragenlijst gegevens missende waarden bevatten. Missende scores op vragen (i.e., missende item scores) vereisen mogelijk andere statistische methoden dan missende totaal scores.

De onderliggende redenen van missende data kunnen onderverdeeld worden in verschillende mechanismes. De data kan *missing completely at random* (MCAR) zijn, wanneer het missende deel van de data een geheel random sub-sample van de data is. Een voorbeeld hiervan is dat een vragenlijst mist doordat deze in de post is kwijtgeraakt. Het is ook mogelijk dat de kans op missende data gerelateerd is aan andere variabelen die in de studie zijn gemeten. Dit mechanisme heet *missing at random* (MAR). Bijvoorbeeld wanneer fysieke activiteit scores vaker missen voor oudere mensen, dan is de missende data voor fysieke activiteit gerelateerd aan leeftijd. Missende data kan ook *missing not at random* (MNAR) zijn, wanneer de missende data gerelateerd is aan de missende score zelf. Bijvoorbeeld als de mensen met een lage score op fysieke activiteit hun fysieke activiteit score missen. De werking van methoden om met missende data om te gaan is afhankelijk van het onderliggende missing data mechanisme. Daarom is het belangrijk om een valide assumptie over het meest waarschijnlijke missing data mechanisme te maken. Dit kan gedaan worden door de data te onderzoeken en goed over de meest waarschijnlijke redenen voor de missende data na te denken.

Incomplete data in epidemiologische studies worden meestal simpelweg niet gebruikt in de analyse, oftewel een complete-case analyse. Daarnaast adviseren veel handleidingen van vragenlijsten om de missende waarden te vervangen voor een bepaalde waarde, bijvoorbeeld de gemiddelde score. Echter, deze methoden werken niet goed en veroorzaken bias in onderzoeksresultaten. Multipele imputatie en f*ull information maximum likelihood* schatting (FIML) zijn geavanceerde methoden om met missende data om te gaan. Beide methoden werken goed in MCAR en MAR data. In multipele imputatie worden de missende waarden vervangen door meerdere plausibele waarden, waardoor er meerdere kopieën van de dataset ontstaan met in iedere dataset andere geïmputeerde waarden. De plausibele waarden worden geschat met behulp van regressie technieken uit de geobserveerde data. De item scores in een vragenlijst worden vaak gemeten met een Likert schaal. Hierdoor zijn de item variabelen ordinaal en vaak niet normaal verdeeld. De regressie technieken om plausibele geïmputeerde waarden te schatten, zoals lineaire regressie, zijn dan niet altijd optimaal. Een procedure die robuust is tegen afwijkingen

van de normaal verdeling is *predictive mean matching.* Hierbij wordt er een random waarde getrokken uit de geobserveerde data waarden die het dichtst bij de voorspelde waarde uit de regressie schatting ligt. Deze methode gebruikt dus de geobserveerde data en imputeert daardoor meer realistische waarden. Multipele imputatie levert meerdere datasets op en deze worden ieder geanalyseerd met het analysemodel dat zou worden gebruikt als de data compleet was geweest. Vervolgens worden de resultaten van deze analyses gecombineerd voor het eindresultaat van de analyse. In FIML worden de populatie parameters geschat die meest waarschijnlijk het datasample zouden kunnen produceren. In deze methode worden geen waarden geïmputeerd of vervangen, maar alle geobserveerde data wordt gebruikt om de parameter schattingen te verkrijgen. Beide geavanceerde methoden, multipele imputatie en FIML, worden beschouwd als de state-of-the-art missing data methoden.

De beste missing data methode om met missende data om te gaan is afhankelijk van de analyse methode die wordt toegepast om de data the analyseren (i.e., longitudinaal of niet), het type variabele in de analyse dat missende waarden bevat (i.e., de predictor/ covariaat of de uitkomst), het missing data mechanisme (i.e., MCAR, MAR, MNAR), het percentage respondenten in de data met missende waarden en het niveau van de missende data in de vragenlijst (i.e., item scores of totaal score niveau).

Missende data in een vragenlijst moet worden behandeld op het item niveau van de vragenlijst. Als de uitkomst in een studie op één tijdspunt is gemeten, en er dus geen longitudinale analyse wordt uitgevoerd, moet multipele imputatie op de item scores worden toegepast. Dit houdt in dat de incomplete item variabelen worden geïmputeerd en dat na de imputatie de totaal scores van de vragenlijsten worden berekend en gebruikt voor analyse.

In studies met heel veel vragenlijsten of extreem lange vragenlijsten kan het aantal item variabelen te groot worden om betrouwbare imputaties te schatten. Een oplossing hiervoor is passieve imputatie. Passieve imputatie methoden combineren de variabelen in het imputatie model om het aantal variabelen in het model te reduceren. De item scores van een vragenlijst worden geïmputeerd, waarbij de totaal scores van de andere vragenlijsten worden gebruikt als predictor. Deze totaal scores kunnen ook missende waarden bevatten die veroorzaakt zijn door missende item scores, en deze worden dan ook op dezelfde manier geïmputeerd. De totaal scores worden tussen elke imputatie herhaling (i.e., iteratie) geüpdate door de geïmputeerde item scores.

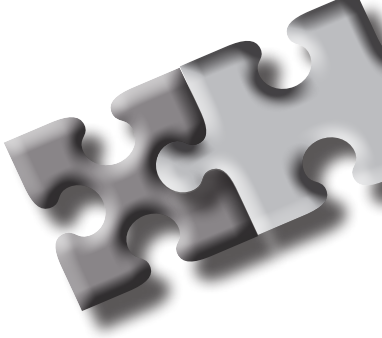
Wanneer de uitkomst in een studie op meerdere tijdspunten wordt gemeten, moet er in de analyse methode rekening gehouden worden met de correlatie tussen de meerdere meetmomenten. Longitudinale analyses maken vaak gebruik van FIML procedures om parameter schattingen te verkrijgen en deze procedures behandelen de missende data in de analyse. Wanneer de uitkomst variabele wordt gemeten aan de hand van een vragenlijst, wordt over het algemeen alleen de totaal score van de vragenlijst in de longitudinale

analyse gebruikt. Desalniettemin moet de missing data op het item niveau aangepakt worden, wanneer de totaal scores incompleet zijn doordat de item scores missende waarden bevatten. De informatie uit de items kan in de analyse worden toegevoegd door de geobserveerde item scores te includeren als hulpvariabelen (i.e., *auxiliary variables*). Op die manier zijn de parameter schattingen meer precies en bevatten ze minder bias. De missende data in de predictor of covariaten in een longitudinale analyse moeten worden behandeld met multipele imputatie.

Het advies met betrekking tot vragenlijsten kan ook worden gebruikt voor andere situaties. Bijvoorbeeld bij kosten-data, waar de totale kosten worden gebruikt in een kosten-effectiviteitsanalyse. Deze totale kosten kunnen incompleet zijn door missende sub-kosten. Het is hier wederom het beste om op het sub-kosten niveau met de missende data om te gaan. Ook is de verdeling van kosten data bijna nooit normaal. Kosten zijn vrijwel altijd positief en vaak scheef naar rechts verdeeld met een overmaat aan nullen. De imputatie strategie kan worden aangepast door het gebruik van *predictive mean matching* op de log-getransformeerde data. Na de imputatie kan de data dan weer terug-getransformeerd worden voor de data analyse.

De belangrijkste conclusies van dit proefschrift zijn dat de methoden die in handleidingen voor vragenlijsten geadviseerd worden vaak niet optimaal zijn en moeten worden genegeerd. Missende waarden in vragenlijsten moeten worden behandeld op item niveau. De informatie uit de geobserveerde item scores verhoogt de accuraatheid en precisie in onderzoeksresultaten. Daarnaast vergroot de inclusie van geobserveerde item informatie als hulpvariabelen in een longitudinaal model om de totaal scores te analyseren de precisie en power van parameter schattingen. Passieve imputatie om missende item scores te imputeren in een dataset met een extreem groot aantal items is een valide methode om met missende item scores om te gaan.

Over het algemeen is het belangrijk om alle beschikbare informatie uit de geobserveerde item scores te betrekken bij het omgaan met missende item scores in vragenlijsten. Dit zorgt voor een optimaal niveau van accuraatheid en precisie in parameter schattingen.



Dankwoord

Acknowledgments

Over het algemeen is het dankwoord het meest gelezen "hoofdstuk" in proefschriften. Ik heb dit zelf niet statistisch onderzocht, maar aan de hand van persoonlijke ervaring en een kleine steekproef onder mijn collega's, durf ik dit wel te concluderen. Ondanks dat ik natuurlijk hoop dat dit proefschrift in dat opzicht uniek is en lezers zich zullen begraven in de inhoudelijke hoofdstukken, vrees ik dat dit dankwoord hetzelfde zal ondervinden. Ik heb even getwijfeld om het dankwoord heel letterlijk te nemen en alleen "Bedankt" op te schrijven, maar dan legt het gros van de lezers mijn proefschrift wel heel snel aan de kant. En natuurlijk heb ik bij het maken van dit proefschrift veel hulp en steun gehad, dus wijd ik met liefde dit hoofdstuk aan mijn collega's, familie en vrienden.

Martijn, bedankt voor de geweldige samenwerking. Onze discussies, die soms over de gang te horen waren, en je enthousiasme heeft dit project voor mij heel succesvol gemaakt. Precies op de momenten dat ik het niet meer voor me zag, wist jij me te overtuigen en te motiveren om het gewoon weer op te pakken. En wanneer ik met een wild idee kwam, was jij altijd een goede sparringpartner. Ik hoop dat we in de toekomst nog in veel projecten kunnen samenwerken. Voor nu in ieder geval nog in onze missing data cursus. Riekie, bedankt voor de fijne samenwerking. Van jou kreeg ik altijd als eerste reacties op mijn papers en vragen. Jouw goed en kritisch commentaar heeft mijn stukken tot een hoger niveau gebracht. En natuurlijk zal ik de gezelligheid op de ISOQoL congressen niet vergeten. Jos, jouw snelle manier van denken is heel prettig in overleggen en werkt heel efficiënt. Ik vind het prettig om met je samen te werken aan onderzoek en ook in onderwijs en ik ben blij dat we dit in ieder geval nog even kunnen voortzetten. Michiel, bedankt voor het meedenken aan mijn projecten en je input bij het schrijven van mijn artikelen.

I would like to thank the members of the reading committee for taking the time and effort to read my thesis and to attend the defence ceremony: Prof dr. Hans Brug, Prof dr. Stef van Buuren, Prof dr. Craig Enders, dr. Joost van Ginkel, Prof dr. Theo Stijnen, dr. Caroline Terwee and Prof dr. Koos Zwinderman.

De leukste collega's, Alette en Rosalie, wat zal ik jullie gaan missen. Vier jaar met elkaar op de werkkamer lief en leed gedeeld over onderzoek en al het andere dat ons bezig hield. Talloze kopjes koffie, thee, ijsjes en taartjes zijn er doorheen gegaan tijdens onze brainstorm sessies en pauzes. Vanaf het begin zijn we samen bezig aan onze onderzoeken en ik ben blij dat jullie er nu aan het einde ook bij zijn als mijn paranimfen.

Tsjitske en Martine, wat hebben wij een toptijd gehad in Miami! En gelukkig bleef het daar niet bij: gezellige etentjes, die we hopelijk blijven voortzetten, en natuurlijk onze schrijfweek in Friesland, die tot de inleiding en discussie van mijn proefschrift heeft geleid.

Natuurlijk wil ik ook al mijn collega's bij de afdeling Epidemiologie & Biostatistiek van het VUmc bedanken. En ook de klinimetrie groep voor de leuke bijeenkomsten en discussies. Ook aan alle collega's van de Methodologie en Toegepaste Biostatistiek van VU gezondheidswetenschappen bedankt voor de leuke tijd. Regelmatig op woensdag even een biertje drinken in de stelling heeft mijn werk en motivatie zeker bevorderd. Paul al zit ik niet meer met je in de trein, bier lust ik nog steeds. En de statistiek meisjes: Trynke, Dagmar en Nienke bedankt voor de gezellige etentjes.

Jaap, ook jou wil ik bedanken voor je hulp in de beginfase van mijn project. Van jou heb ik geleerd om meer wiskundig te denken. Gaandeweg heb je mij alle ins en outs van het programmeren en simuleren in R geleerd en dit heeft me erg geholpen in mijn gehele project.

Janet, thank you for the great collaboration. I quickly got to know your very thorough working method, which got us in a very interesting study that became one of the chapters of my dissertation. I really liked working with you and chatting over coffee.

I also want to thank Craig for the great collaboration and for making it possible for me to spend three months at Arizona State University as a research scholar. I really hope that we will work together again in the future. And also thanks to all colleagues at Arizona State University that I met during my visit that made my stay very memorable and pleasant. Gina, thank you for sharing your workspace with me. Milica thank you for showing me the great music and nightlife scene of Tempe and Phoenix, I hope we will stay in touch. And off course my roomie Becky, thank you for the great time! I hope you are still working on that beer-map!

Vlak na het eerste jaar van mijn promotieonderzoek heb ik helaas afscheid moeten nemen van mijn lieve moeder. Ondanks dat is het gelukt om dit proefschrift helemaal binnen te tijd af te krijgen, maar dat was mij zeker niet gelukt zonder de steun van mijn lieve familie en grote groep vrienden.

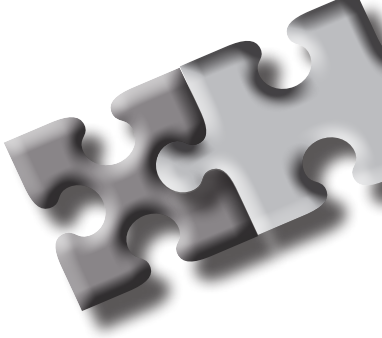
Papa, bedankt voor het overbrengen van een groot doorzettingsvermogen om doelen te bereiken. Maar vooral bedankt voor je liefde en trots en dat je er altijd voor me bent. Amber en Jentel, mijn lieve lieve zusjes, zonder jullie was dit zeker niet gelukt. Bedankt dat jullie er altijd voor mij zijn om mijn verhalen aan te horen en gezellige dingen te doen. Rem, bedankt voor de mooie tijd en voor alles wat nog gaat komen. Oma, bedankt dat u er voor ons bent. Natuurlijk ook dank aan de rest van mijn lieve familie, die altijd voor mij klaar staan.

Lieve oud-huisgenoten van de Hooigracht. Ook al wonen we niet meer bij elkaar zien we elkaar toch nog vaak bij feestjes of in de kroeg. Klaas, Geert, Daan, Daniel, JW, Jurgen, Natalie, Ras, Bram bedankt voor alle mooie tijden. Bas, bedankt voor je belangstelling in mijn onderzoek en de discussies erover in de kroeg en natuurlijk voor de inspiratie voor de laatste alinea van dit dankwoord. Jason, bedankt voor je

gezelschap en lekkere snacks op mijn thuiswerkdagen. Luc, bedankt dat je ook bij dit feestje weer wilt spelen. Raq, Sil en Pauli bedankt voor jullie liefde en trots.

Al mijn andere lieve vrienden en vriendinnen: Ellen, Charlie, Floor, Natas, Gert, Nathalie, Nelleke, Emma, Lotte, Mariam. Bedankt dat jullie er altijd voor mij zijn voor een goed gesprek, een lekker wijntje of een biertje in de kroeg. Karine, bedankt dat je altijd zo een lieve vriendin bent, bedankt voor het altijd aanhoren van mijn verhalen en voor het laten zien van de andere kant van de overweging. Maartje, bedankt dat je er altijd bent en voor mij klaar staat. Nicole en Anne, bedankt voor de gezellige MINA etentjes. Meia, bedankt voor de gezellige concerten, festivals en biertjes. Wen, bedankt voor het redigeren van mijn website teksten en het uittesten van de applicaties. Voetbal meiden, bedankt voor de heerlijke tijd op het veld. Wendy, bedankt voor het momentje rust in de week tijdens de yoga lessen. Ook dank aan de Young Statisticians, en vooral natuurlijk mede bestuursgenoten Nynke, Sanne en Nadia, voor het belichten van de gezellige kant van statistiek. Elise, bedankt voor je hulp bij het ontwerpen van de omslag en de lay-out van dit proefschrift. Ik ben erg blij met het eindresultaat.

In het onwaarschijnlijke geval dat er nog mensen missen in dit dankwoord, kan er altijd gebruik worden gemaakt van multipele imputatie om met deze missende waarden om te gaan. Maar vergeet niet om alle informatie die er wel is hierin mee te nemen! Op www.iriseekhout.com is hierover meer te vinden.



Publication list

Published or accepted for publication

Eekhout, I., De Boer, M.R., Twisk, J.W.R., De Vet, H.C.W., Heymans, M.W. (2012) Reporting on and handling of missing data in epidemiological and medical research: a systematic review. Epidemiology, 23(5), 729-732.

Eekhout, I., de Vet, H.C.W., Twisk, J.W.R., Brand, J.P.L., de Boer, M.R., Heymans, M.W. (2014.) Missing data in a multi-item instrument were best handled by multiple imputation at the item score level. Journal of Clinical Epidemiology, 67(3):335-342.

Eekhout, I., Enders, C.K., Twisk, J.W.R., de Boer, M.R., de Vet, H.C.W., Heymans, M.W. (in press) Analyzing incomplete item scores in longitudinal data. Structural Equation Modeling.

Under review

Eekhout, I., Enders, C.K., Twisk, J.W.R., de Boer, M.R., de Vet, H.C.W., Heymans, M.W. Longitudinal data analysis with auxiliary item information to handle missing questionnaire data. Journal of Clinical Epidemiology.

MacNeil-Vroomen, J., **Eekhout, I.** , Dijkgraaf, M.G., Van Hout, H., De Rooij, S.E., Heymans, M.W., Bosmans, J.E., Comparing multiple imputation strategies for zero-inflated cost data in economic evaluations: which method works best? European Journal of Health Economics.

Eekhout, I., de Vet, H.C.W., de Boer, M.R., Twisk, J.W.R., Heymans, M.W. Missing data in multi-item questionnaires: analyze carefully and don't waste available information. International Journal of Epidemiology.

Eekhout, I., de Vet, H.C.W., de Boer, M.R., Twisk, J.W.R., Heymans, M.W. Passive imputation of missing values in studies with many multi-item questionnaire outcomes. Quality of Life Research.

Halberstadt, J. , de Vet, E., Nederkoorn, C., Jansen, A., van Weelden, O., **Eekhout, I.**, Heymans, M.W., Seidell, J. The association of self-regulation with weight loss maintenance after an intensive combined lifestyle intervention for children and adolescents with severe obesity. International Journal of Obesity.

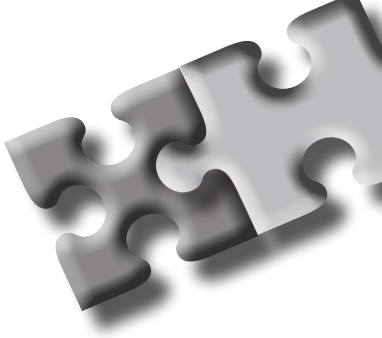
Terluin, B., **Eekhout, I.**, Terwee, C.B., de Vet, H.C.W. Estimating the minimal important change (MIC) using a predictive modelling appoach. Journal of Clinical Epidemiology.

Manuscript in preparation

Eekhout, I., Van de Wiel, M., Heymans, M.W. A new method for significance testing of categorical covariates after multiple imputation.

Spriensma, A.S., **Eekhout, I.**, de Boer, M.R., Twisk, J.W.R., Heymans, M.W. Analyzing longitudinal RCT outcomes with floor effects: a simulation study.

Haanstra, T.M., Kraamwinkel, N., **Eekhout, I.**, Nijpels, G., de Vet, H.C.W., Ostelo, R. The association between cognitive illness perceptions and health outcomes in type 2 diabetes patients.



Curriculum vitae

Iris Eekhout

Iris Eekhout was born on October 6th 1985 in Kwintsheul, the Netherlands. In 2008 Iris finished her Bachelors degree in Psychology at Leiden University and as part of that degree she studied one semester at the University of Montana, USA. At the University of Montana Iris worked on a research project on domestic violence. In 2010 Iris finished her Masters degree in Clinical Psychology and her Masters degree in Methodology and Statistics at Leiden University. During her Masters in Clinical Psychology she did a research internship at Leiden University investigating the relation between worry, problem-solving skills and word use in a student sample under the supervision of Jolijn Drost and Prof. dr. Philip Spinhoven. For the Methodology and Statistics master Iris did a research internship at the Netherlands Forensic Institute (NFI) researching dimension reduction methods in gasoline data to calculate likelihood ratios for forensic comparisons. At the NFI Iris worked with dr. Reinoud Stoel and dr. Annabel Bolck and under supervision of Prof. dr. Willem Heiser at Leiden University.

After her graduation Iris started working as a PhD student at the EMGO Institute for Health and Care Research (EMGO+) and the department of Epidemiology and Biostatistics of the VU University medical center. Under the supervision of dr. Martijn Heymans, Prof. dr. Riekie de Vet, Prof. dr. Jos Twisk and dr. Michiel de Boer she worked on her project about handling missing data in questionnaire items and total scores. As part of her PhD, Iris spend three months at Arizona State University as a research scholar in 2013 to work with Prof. dr. Craig Enders on a project about handling missing item scores in longitudinal data.

Since November 2014 Iris works at the Military Mental Health Care - Research Center and University Medical Center Utrecht as a post-doctoral researcher. The research is about post-traumatic stress disorder in soldiers. Besides her research work, she will remain working as a part-time teacher in the Epidemiology master (EpidM) and consultant at the Epidemiology & Biostatistics department of the VU University medical center.

Iris Eekhout

After finishing a master in Clinical Psychology and a master in Methodology and Statistics at Leiden University, Iris Eekhout worked as a PhD student at the EMGO Institute for Health and Care Research and the department of Epidemiology and Biostatistics of the VU University medical center in Amsterdam on a project about handling missing questionnaire items and total scores.

In research, missing data occur when a data value is unavailable. Many empirical studies encounter missing data. Missing data can occur in many stages of research due to many different causes in many different forms. For example, missing data can take place on one or more of the measured variables that are used as a predictor, covariate or outcome. Missing data can also occur on a multi-item questionnaire due to questions that have not been filled-out by the participant. In that case some items can be missing, or the entire questionnaire might not be filled out. These different forms of missing data can have different underlying causes and might require different solutions. The performance of missing data methods depends on several aspects of the study and the missing data. These are the analysis method that is applied to analyze the dataset (i.e., longitudinal analysis or not), the location of the missings in the 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). This dissertation aims to help researchers find an advised solution to their specific missing data problem.

www.missingdata.nl