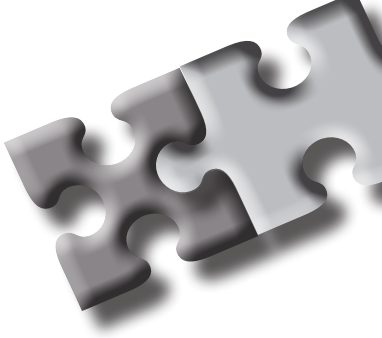

Chapter 1

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

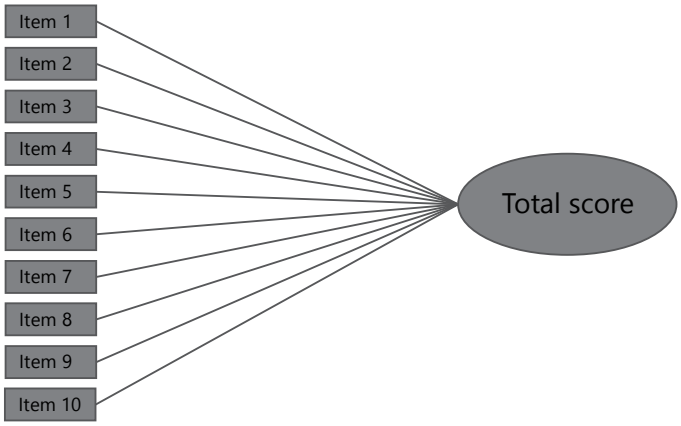

Figure 1.1. Example of a multi-item questionnaire with 10 items that result in a total score.

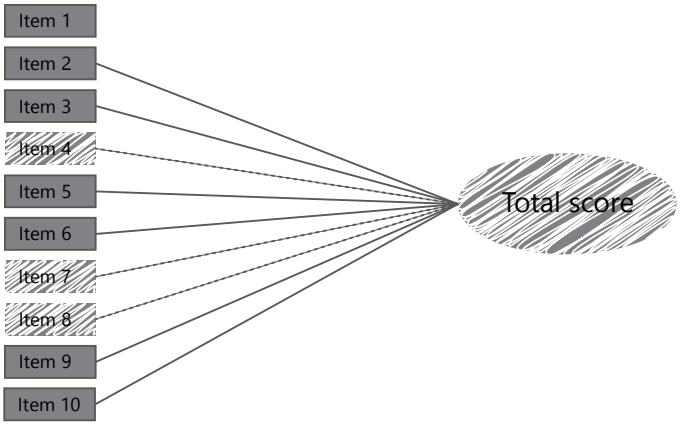

Figure 1.2. Example of a multi-item questionnaire with 3 out of 10 item scores missing that result in an incomplete total score.

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.