ELSEVIER

Journal of Clinical Epidemiology 67 (2014) 335–342

Journal of Clinical Epidemiology

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

Iris Eekhout[a,b,c,*], Henrica C.W. de Vet[a,b], Jos W.R. Twisk[a,b,c], Jaap P.L. Brand[d], Michiel R. de Boer[c,e], Martijn W. Heymans[a,b,c]

[a]*Department of Epidemiology and Biostatistics, VU University Medical Center, P.O. box 7057, 1007 MB Amsterdam, The Netherlands*
[b]*EMGO Institute for Health and Care Research, VU University Medical Center, Van der Boechorststraat 7, 1081 BT Amsterdam, The Netherlands*
[c]*Department of Methodology and Applied Biostatistics, Faculty of Earth and Life Sciences, Institute for Health Sciences, VU University, De Boelelaan 1085, 1081 HV Amsterdam, The Netherlands*
[d]*Skyline Diagnostics, Marconistraat 16, 3029 AK Rotterdam, The Netherlands*
[e]*Department of Public Health, University Medical Center Groningen, PO box 196, 9700 AD Groningen, The Netherlands*

Accepted 13 September 2013; Published online 2 December 2013

Abstract

Objectives: Regardless of the proportion of missing values, complete-case analysis is most frequently applied, although advanced techniques such as multiple imputation (MI) are available. The objective of this study was to explore the performance of simple and more advanced methods for handling missing data in cases when some, many, or all item scores are missing in a multi-item instrument.

Study Design and Setting: 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.

Results: 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 items (>25%), MI methods applied to the items outperformed methods applied to the total score.

Conclusion: We recommend applying MI to the item scores 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

1. 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 [1], for example, the Pain Coping Inventory assesses active coping skills of people with pain complaints by 12 items [2]. 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 [3]. 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) [4]. 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

Funding: This work was financially supported by EMGO Institute of Health and Care Research.

Conflict of interest: None.

* Corresponding author. Department of Epidemiology and Biostatistics, VU University Medical Center, Room MF D439, Van der Boechorststraat 7, 1081 BT Amsterdam, The Netherlands. Tel.: +31-204446040; Fax: + 31 20 444 8181.

E-mail address: i.eekhout@vumc.nl (I. Eekhout).

0895-4356/$ - see front matter
http://dx.doi.org/10.1016/j.jclinepi.2013.09.009

What is new?

Key findings

- Mean imputation methods result in highly biased estimates in all missing data situations when more than 10% of subjects data missing data. Furthermore, single stochastic regression turns out to be the best working single-imputation method, but standard errors are underestimated because missing data uncertainty is not incorporated.
- Multiple imputation (MI) repeats the imputation process multiple times to incorporate missing data uncertainty; accordingly, MAR item data are best handled by applying MI based on predictive mean matching or stochastic regression to the item scores.

What this adds to what was known?

- Complete-case analysis (CCA) is still used in 80% of epidemiological studies; CCA results in unbiased estimates for the regression coefficient; however, this method overestimates error and decreases power.
- User manuals of widely used multi-item questionnaires advise item or person mean imputation, but this method yields highly biased estimates when more than 10% of cases have missing data and is therefore not recommended.

What is the implication and what should change now?

- Missing item score data should be handled by applying MI to the items. MI is now available in many software packages, which makes it accessible for all researchers. When only small amount of item scores are missing (<25%) in only a small amount of cases (<10%), CCA or single stochastic regression imputation can be preferred purely for practical reasons.

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) [4]. In contrast to traditional methods such as CCA and mean imputation, more advanced methods such as multiple imputation (MI) produce reliable and unbiased results under the MAR mechanism and take missing data uncertainty into account [5,6]. 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 [3]. Other simulation studies have researched the performance of missing data methods applied to nonquestionnaire data [7,8] or only studied methods applied to the item scores of a multi-item instrument [9–13]. For example, Burns et al. [13] studied the performance of MI of missing item scores but did not compare this with 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 present study aimed to explore the performance of different missing data-handling methods designed for missing item scores and missing total scores in a multivariate regression model. This objective is considered in the following two aspects: (1) which missing data methods should be used to handle missing (item) data and (2) should this missing data method be applied to the item scores or to the total scores.

2. Methods

2.1. Simulation set up

To investigate the differences between several imputation methods, we used a simulation procedure comparable with the study performed by Marshall et al. [7]. We based our simulation on an empirical data set, which was previously used in a prospective cohort study investigating the prognosis of low back pain [14]. 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) [2]. The PCI-active consists of 12 items with four ordered response categories that result in a total sum score, which we considered as a continuous scale. Additionally, five other covariates were selected to be included in this data set: gender, health status, job demands, number of working years, and absenteeism, and the outcome variable was lower 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 (R Core Development Team) [15]. 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 data sets. The implications of these different techniques were compared by fitting