ELSEVIER

Journal of Clinical Epidemiology 63 (2010) 721−727

Journal of Clinical Epidemiology

ORIGINAL ARTICLES

Missing covariate data in medical research: To impute is better than to ignore

Kristel J.M. Janssen[a,*], A. Rogier T. Donders[b], Frank E. Harrell Jr.[c], Yvonne Vergouwe[a], Qingxia Chen[c], Diederick E. Grobbee[a], Karel G.M. Moons[a]

[a]*Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands*
[b]*Department of Epidemiology, Biostatistics and Health Technology Assessment, Radboud University Nijmegen Medical Center, Nijmegen, The Netherlands*
[c]*Department of Biostatistics, Vanderbilt University School of Medicine, Nashville, TN, USA*

Accepted 14 December 2009

Abstract

Objective: We compared popular methods to handle missing data with multiple imputation (a more sophisticated method that preserves data).

Study Design and Setting: We used data of 804 patients with a suspicion of deep venous thrombosis (DVT). We studied three covariates to predict the presence of DVT: D-dimer level, difference in calf circumference, and history of leg trauma. We introduced missing values (missing at random) ranging from 10% to 90%. The risk of DVT was modeled with logistic regression for the three methods, that is, complete case analysis, exclusion of D-dimer level from the model, and multiple imputation.

Results: Multiple imputation showed less bias in the regression coefficients of the three variables and more accurate coverage of the corresponding 90% confidence intervals than complete case analysis and dropping D-dimer level from the analysis. Multiple imputation showed unbiased estimates of the area under the receiver operating characteristic curve (0.88) compared with complete case analysis (0.77) and when the variable with missing values was dropped (0.65).

Conclusion: As this study shows that simple methods to deal with missing data can lead to seriously misleading results, we advise to consider multiple imputation. The purpose of multiple imputation is not to create data, but to prevent the exclusion of observed data. © 2010 Elsevier Inc. All rights reserved.

Keywords: Missing data; Complete case analysis; Multiple imputation; Bias; Coverage; DVT

1. Introduction

No matter how hard researchers try to prevent it, missing data occur frequently in medical research [1]. Commonly, researchers simply neglect all the data of patients with missing values because this is what standard software packages do when the data are analyzed (complete case analysis). Because this leads to a smaller dataset, it comes at least at the price of loss of power. Complete case analysis not necessarily leads to biased results. Under the condition that the missing values are missing completely at random (MCAR), meaning that the cause of missingness is pure coincidence, complete case analysis will not lead to biased results. As an alternative to complete case analysis, researchers tend to drop a variable from the analysis when it has missing values. However, both methods neglect valuable observed data.

Multiple imputation is a statistical technique that uses all observed data to fill in plausible values for the missing values [2−8]. This method receives increasing attention in the medical literature [9−16]. Nevertheless, many researchers seem unaware or uncertain about this approach to deal with missing values and still perform a complete case analysis or drop variables with missing values from the analysis [17]. The extent and sort of bias related to these approaches depend on the type of study. Diagnostic or prognostic studies often study the contribution of covariates (eg, patient characteristics and test results) in the prediction of a particular outcome by estimating the predictors' regression coefficients. For example, one may study the predictive effect of body mass index (BMI), age, gender, the intake of saturated fat, and other life style factors on the risk of cardiovascular diseases (CVD). Sometimes, these studies are aimed at developing a multivariable prediction model or risk score and estimate the ability of such a model to distinguish between patients at high and low risk of CVD. In etiologic studies, usually the effect of a specific

* Corresponding author. Tel.: +0031-8875-51752; fax: +0031-8875-55485.

E-mail address: k.j.m.janssen@umcutrecht.nl (K.J.M. Janssen).

0895-4356/$ − see front matter © 2010 Elsevier Inc. All rights reserved.
doi: 10.1016/j.jclinepi.2009.12.008

What is new?

- Dropping a variable with missing data from the analyses or conducting a complete case analysis more often leads to biased effect estimates, decreased coverage of the confidence intervals, and a decreased discriminative ability of the multivariable model, compared with multiple imputation.
- To ''provide'' data according to the strict methodology of multiple imputation seems a better alternative than to give up or delete valuable observed data.

etiologic factor on the outcome of interest is studied corrected for the influence of other covariates (confounders). Following the previous example, the regression coefficient of BMI could be the parameter of interest, corrected for the confounders age, gender, intake of saturated fat, and other life style factors.

We used empirical data of a previous study on deep venous thrombosis (DVT) to quantify the effect of different analyses in the presence of missing covariate data both for prediction and etiologic research purposes. We studied the effect of complete case analysis, dropping covariates with missing values, and multiple imputation on individual regression coefficients and on the predictive ability of a multivariable model for various proportions of missing covariate values.

2. Methods

2.1. Empirical data

Data were obtained from a large cross-sectional study among adult patients with a suspicion of DVT. For specific details and main results of the study, we refer to the literature [18–20]. In brief, patients with a suspicion of DVT were consecutively included when they visited one of 110 participating primary care physicians in The Netherlands. Suspicion of DVT was primarily based on the presence of at least one of the following symptoms or signs of the lower extremities: swelling, redness, or pain in one of the legs. After informed consent, the primary care physician systematically documented the patient's history and physical examination. Subsequently, venous blood was drawn to measure the D-dimer level. Finally, all patients were referred to the hospital to undergo the reference test (repeated compression ultrasonography of the lower extremities) to determine the presence or absence of DVT.

For our illustration, we specifically selected two dichotomous variables (difference in calf circumference of 3 cm or more and history of a leg trauma) and one continuous variable (D-dimer level) with different mutual correlations. A difference in calf circumference of 3 cm or more was correlated with the D-dimer level (eta = 0.28), whereas history of a leg trauma was neither correlated with the D-dimer level (eta = 0.04) nor with a difference in calf circumference of 3 cm or more (Pearson product-moment correlation coefficient = 0.06). We included 804 patients with completely observed data on any of the three variables, including the outcome. This will be referred to as the ''true'' original study sample. Thirty-eight percent had a difference in calf circumference of 3 cm or more, 17% had a leg trauma in the past 4 weeks, and the prevalence of DVT was 20% (Table 1).

We fitted a multivariable logistic regression model with these three independent variables and DVT presence (yes/no) as the outcome. D-dimer level was included by a natural logarithm transformation. All three were predictors of DVT presence or absence. The estimated regression coefficients in this original study sample were considered as the ''true'' values (Table 1) with which all subsequent estimations were compared.

2.2. Missing values

Missing values can be caused by several mechanisms. When, for example, a tray with blood samples drops from a table and the samples can, therefore, not be analyzed, the missing values are completely random (MCAR) [5]. Missingness is, however, often related to other observed patient characteristics. For example, patients who are relatively healthier might be less likely to undergo subsequent, more invasive tests, leading to more missing values on those tests for these patients. Such missing values are called missing at random (MAR) [5]. The missing values are random *conditional* on the other available information. If no information exists on the reason for missingness, these missing values are called missing not at random (MNAR) or nonignorable missing. This means that the probability that an observation is missing depends on unobserved subject information. Usually, it is plausible to assume that the

Table 1
Distribution of the studied predictors: the (natural logarithm of) D-dimer level, history of a leg trauma (yes/no) and difference in calf circumference of 3 cm or more (yes/no), and the true values of the logistic regression coefficients

Predictors	Distribution, % (*n*)	True regression coefficients[a]
Intercept	—	−13.24
Difference in calf circumference of 3 cm or more	38 (306)	0.60
Natural logarithm of the D-dimer level[b]	6.83 (1.49)[b]	1.58
History of a leg trauma	17 (136)	−0.50

[a] No 95% confidence interval is given because these regression coefficients are considered to be the truth.
[b] Mean (standard deviation).

Download English Version:

https://daneshyari.com/en/article/1082579

Download Persian Version:

https://daneshyari.com/article/1082579

Daneshyari.com