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.

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).

missing values in medical research are related to observed subject information [7,8]. We, therefore, generated missing values according to a MAR mechanism.

We generated missing values in one variable, the D-dimer level. The probability that a D-dimer level was missing depended on the presence of DVT and difference in calf circumference. Missing values were generated with proportions of missing ranging between 10% and 90%. For each proportion of missing values, we simulated 500 datasets that were based on the original study sample (804 patients). More information about the mechanism that was used to create missing values is provided in the Appendix.

2.3. Methods to deal with missing values

We studied the following three methods to deal with missing values before we developed the multivariable logistic regression model.

1. Complete case analysis: analyzing only the data of patients without missing values.
2. Dropping the variable with missing values (D-dimer level): analyzing only the remaining covariates (a difference in calf circumference of 3 cm or more and a history of leg trauma).
3. Multiple imputation using the default settings of Mice, in which regression models are estimated [21]: multiple imputation replaces each missing value with *m* (here: 10) values drawn from an appropriate estimated distribution. Per simulation, the imputation model was estimated, and 10 imputed datasets were created, which were all analyzed using the same standard method, that is, fitting the previously mentioned multivariable logistic model. The models were combined using proposed methods that reflect the extra variability because of missing values [5].

Each method was applied to all simulations resulting in 500 fitted models per method. Per method, the 500 estimated regression coefficients and standard errors were summarized.

2.4. Outcomes of interest

We compared the results of the three methods with the results of the analyses on the original study sample. The outcomes of interest were the possible bias in the regression coefficients of the variables, coverage of the confidence intervals (CIs) of the regression coefficients, and the discriminative ability.

1. Bias. We compared the true value of the regression coefficients of the three variables (Table 1) with the corresponding mean of the 500 estimations of these regression coefficients in the simulated data.
2. Coverage of CIs. We estimated the percentage of the 90% CIs that included the true value of the regression coefficient of each variable. Values near 90% represent adequate coverage, and values lesser than 90% indicate that the 90% CI is too narrow. This implies that in studies with a valid null hypothesis (‘‘variable has no effect’’), a significant effect will be too often found. Values greater than 90% indicate that the 90% CI is too wide, which implies that the power of the study is suboptimal and more type II errors may occur.
3. Discriminative ability, expressed by the area under the receiver operating characteristic (ROC) curve (ROC area) or the c-statistic [22,23]. The ROC area is a commonly used measure to indicate the overall discrimination, that is, the ability of a prediction model to distinguish between patients with and without DVT. An ROC area ranges from 0.5 (no discrimination; same as flipping a coin) to 1.0 (perfect discrimination). For each method, 500 models could be fitted, and the ROC area was estimated and averaged. We compared the true value of the ROC area (0.88) with the average ROC areas.

2.5. Etiologic perspective

Obviously, none of the three variables in our data were (potential) etiologic factors of DVT. However, statistically we could analyze our data as if it was an etiologic study with missing values on a confounder variable, a frequently encountered situation. To do so, we considered difference in calf circumference of 3 cm or more as the hypothetical etiologic factor, with a history of a leg trauma and the D-dimer level as hypothetical confounders. The confounder D-dimer level had missing values ranging from 10% to 90% as described above. We used the same data to illustrate the effects on the regression coefficient of the etiologic factor under study of complete case analysis, analyses where confounders with missing values are dropped, and multiple imputation.

3. Results

3.1. Bias

Figure 1 shows the mean of the 500 simulation estimates and the true value of the regression coefficients of the three variables. Complete case analysis resulted in a severely biased (underestimated) regression coefficient of a difference in calf circumference of 3 cm or more. The bias was much larger when the proportion of missing values in the D-dimer level was also large. Dropping the D-dimer level from the analysis led to a biased (systematically overestimated) regression coefficient of difference in calf circumference, irrespective of the proportion of missing values in the D-dimer level. Note that any variability in the estimates of the regression coefficients when the predictor with missing values is dropped across different proportions of missing

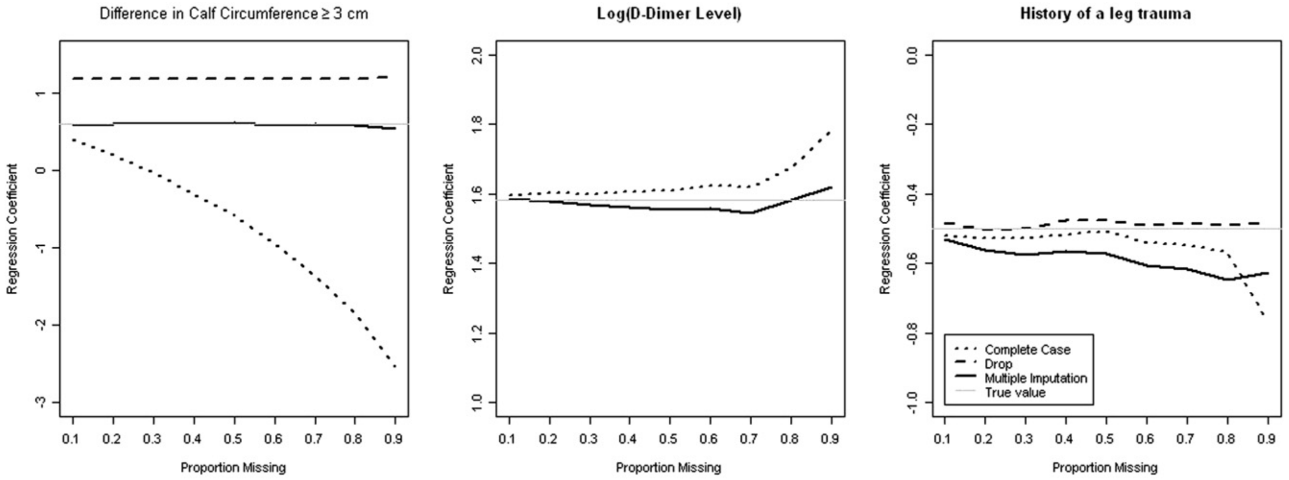

Fig. 1. Estimates of the regression coefficients of the difference in calf circumference of 3 cm or more, the natural logarithm of D-dimer level, and a history of a leg trauma after applying complete case analysis (dotted line), dropping the D-dimer level from the analysis (dashed line), and multiple imputation (black solid line). The proportion of missing data in D-dimer level increased from 10% to 90% in steps of 10%. The gray solid line represents the true value of the regression coefficients.

values is because of sampling variation in the 500 simulations. Multiple imputation led to an unbiased regression coefficient of difference in calf circumference. For the D-dimer level itself, complete case analysis resulted in a biased regression coefficient only when the proportion of missing values was high (more than 70%). However, this bias was small because the true value was −1.58 (Table 1), and the most severely biased estimate was −1.78. Multiple imputation led to an unbiased regression coefficient of the D-dimer level. Logically, the regression coefficient of the D-dimer level could not be estimated when it was dropped from the analyses.

Complete case analysis resulted in a biased regression coefficient of a history of a leg trauma only when the proportion of missing values was about 60% or higher. Dropping the D-dimer level did not lead to a biased estimate of a history of a leg trauma. Multiple imputation led to a biased estimate of a history of a leg trauma when the proportion of missing values was large. However, this bias was small because the true value was −0.50, and the most severely biased estimate was −0.65.

3.2. Coverage of the 90% CIs

Figure 2 shows the coverage of the 90% CI of the regression coefficients. Complete case analysis resulted in coverage of the regression coefficient of a difference in calf circumference that was too low. The coverage was substantially lower when the proportion of missing values in the D-dimer level was higher. Dropping the D-dimer level from the analysis resulted in coverage of almost zero irrespective of the proportion of missing values. Multiple imputation resulted in good coverage (around 90%). Complete case analysis and multiple imputation resulted in good coverage of the 90% CI of the regression coefficient of the D-dimer level. All three methods resulted in good coverage of the 90% CI of the regression coefficient of a history of a leg trauma.

3.3. Discriminative ability

Figure 3 shows the ROC areas of the multivariable (prediction) models. The ROC area after complete case analysis was similar to the true value of 0.88 when only few values were missing. High proportions of missing values resulted in a ROC area of only 0.77. The ROC area was very low (0.65) when the D-dimer level was dropped from the analysis. The ROC area of the model after multiple imputation was equal to the true discriminative ability even with high proportions of missing values.

3.4. Etiologic perspective

When we would interpret the same results as if it was an etiologic study with the missing values occurring in a confounder (here: D-dimer level), we would draw the same conclusions. Complete case analysis and dropping the confounder from the analyses resulted in a biased regression coefficient of the etiologic factor (difference in calf circumference) with substantially lower coverage across all missing confounder value proportions. In contrast, multiple imputation would yield limited bias and appropriate coverage.

4. Discussion

We compared three methods to deal with missing data in a study on the diagnosis of DVT that considers three covariates (predictors). Complete case analysis resulted in two slightly biased regression coefficients (only when the

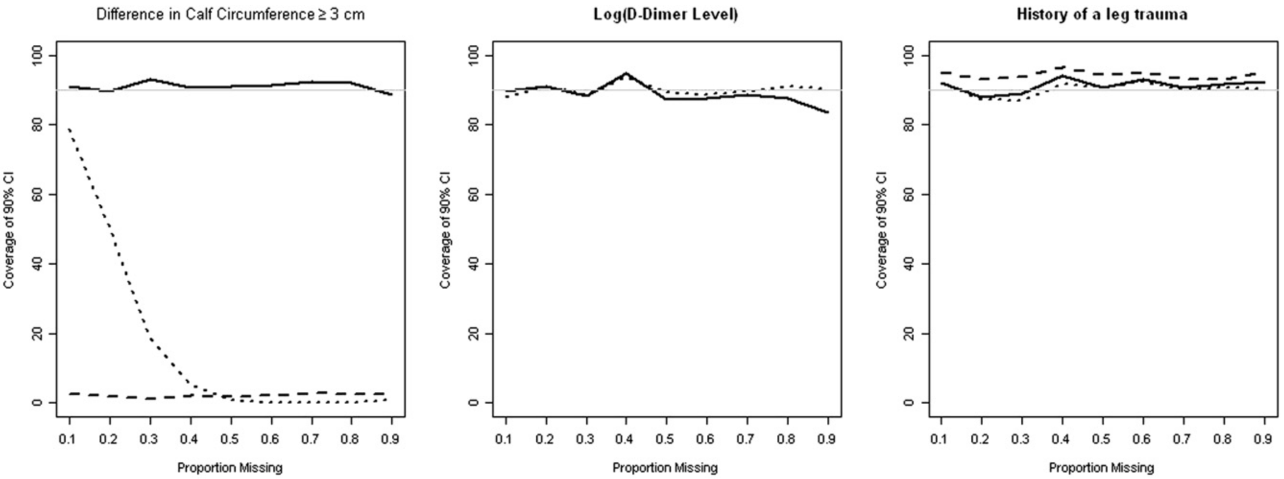

Fig. 2. Coverage of the 90% confidence intervals (CI) of the regression coefficients of a difference in calf circumference of 3 cm or more, the natural logarithm of D-dimer level, and a history of a leg trauma after applying complete case analysis (dotted line), dropping the D-dimer level from the analysis (dashed line), and multiple imputation (black solid line). The proportion of missing data in the D-dimer level increased from 10% to 90% in steps of 10%. The gray solid line represents the true value of the coverage.

covariate had many missing values) and one severely biased regression coefficient (irrespective of the amount of missing covariate values). Dropping the covariate with missing values from the analysis resulted in one severely biased regression coefficient and one unbiased regression coefficient of the two remaining covariates. Multiple imputation resulted in two unbiased and one slightly biased regression coefficient. Complete case analysis resulted in good coverage of the 90% CI for two regression coefficients but for one regression coefficient coverage was too low. Dropping the covariate with missing values from the analysis resulted in good coverage of the 90% CI of one regression coefficient but to a coverage close to zero for another. Multiple imputation resulted in good coverage of the 90% CI of all three regression coefficients irrespective of the proportion of missing values.

Estimates of another important outcome measure in multivariable prediction research, the ROC area, were highly variable for the three methods. Compared with the true ROC area, complete case analysis resulted in a lower ROC area when the proportion of missing values was high. Dropping the predictor with missing values resulted in a much lower ROC area irrespective of the proportion of missing values. The discriminative ability after multiple imputation was similar to the true discriminative ability.

We studied a broad range of proportions of missing data, that is, from 10% to 90%. Results for the extremely high proportions are shown for illustration and should be interpreted with care. We surely do not want to suggest that it is legitimate to analyze datasets using multiple imputation with up to 90% missing values for a particular covariate. In any situation with this much missing covariate values, researchers should question the quality of their remaining data for this variable. However, it does suggest that the proportion of missing covariate data not necessarily determines whether multiple imputation can or cannot be used, but rather the number of observations that is left. In our example, if 90% of the values were missing, complete observations for 80 patients remained available for the analyses. Our results might have been different when, for instance, only 20 patients had observations for all variables in the case of 90% missing values. Of course, the reason for missing values is all important.

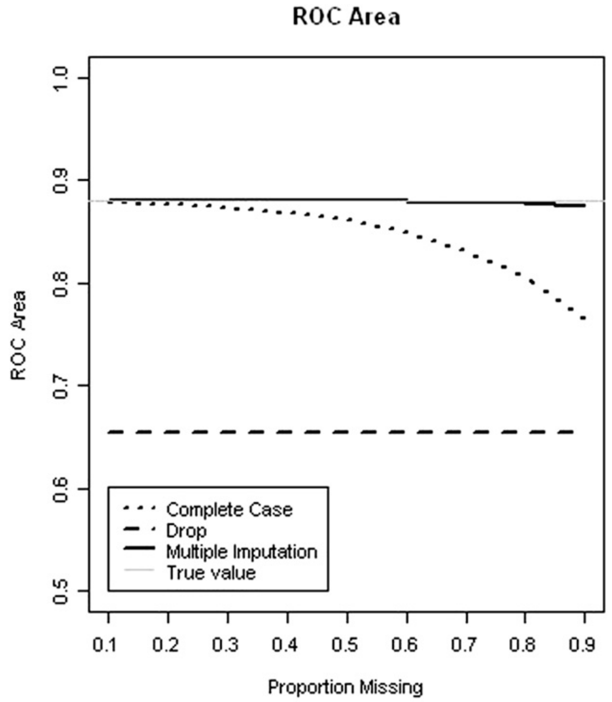

Fig. 3. Receiver operating characteristic (ROC) area of the prediction model after applying complete case analysis (dotted line), dropping the predictor with missing data from the analysis (dashed line), and multiple imputation (black solid line) when the proportion of missing data in the predictor increased from 10% to 90% in steps of 10%. The gray solid line represents the ROC area when there were no missing data (true value).

In our hypothetical exercise, we considered calf circumference as the etiologic factor with the D-dimer level (in which missing values were assigned) and history of a leg trauma as confounders. Similar inferences could be made. Multiple imputation resulted in a less biased regression coefficients of the etiologic factor. Complete case analysis and dropping the confounder D-dimer resulted in a biased regression coefficient of the etiologic factor (complete case analysis even resulted in a negative coefficient). This indicates that when a confounder with missing values is dropped from the analysis, this may (severely) bias the effect estimate of the etiologic factor. Multiple imputation resulted in proper coverage of the corresponding 90% CI, also when the proportion of missing values in the confounder was high. The substantially lower coverage after complete case analysis or dropping the confounder with missing values indicates that too often a significant effect will be found for the etiologic factor in situations where it, in reality, has no effect.

The effect of the different methods on the coverage can be easily explained. When the D-dimer level is dropped from the analysis, the regression coefficient of the difference in calf circumference changes. As a result, the true regression coefficient of the difference in calf circumference is hardly ever in the 90% CI, and, therefore, the coverage is extremely low. The same phenomenon occurs when complete case analysis is applied and the percentage of missing data increases; the regression coefficient of the difference in calf circumference becomes more biased, and the number of times the true regression coefficient of the difference in calf circumference in the 90% CI decreases. The reason why the coverage of history of a leg trauma remains stable when the percentage of missing values increases is that the D-dimer level is independent of a history of a leg trauma. Therefore, the regression coefficient of a history of a leg trauma does not become biased when the percentage of missing values in the D-dimer level increases.

When we would have considered history of a leg trauma as the etiologic factor (and a difference in calf circumference and the D-dimer level as the confounders), dropping the D-dimer level from the analysis gave unbiased results because history of a leg trauma and D-dimer level were not correlated ($r = 0.04$). Strictly speaking, the D-dimer level would not be a confounder of the ‘‘causal’’ association between a history of a leg trauma and the presence of DVT because of the absence of correlation between both. The bias after multiple imputation was again relatively small: a deviation of the regression coefficient from -0.50 (true value) to -0.65 (most severely biased estimate).

When we repeated the analyses and simulated the missing values in calf circumference instead of the D-dimer level, multiple imputation again showed better results than complete case analysis or dropping the variable with missing values from the analysis (data not shown).

We simulated the missing values in our dataset according to a MAR mechanism. It is known from the literature that missing values in medical or social sciences often occur on a MAR mechanism [5,7,8]. Further, it has been shown that even when missing covariate values are not precisely MAR (but also slightly MNAR), multiple imputation still tends to do better than ad hoc methods, such as dropping the covariate with missing values or performing a complete case analysis [7]. In many realistic cases, an erroneous assumption of MAR will have minor impact on the results. When datasets contain many detailed patient characteristics, the missing covariate values can still be reliably imputed [8,24]. Yet, there are also situations in which an erroneous assumption of MAR will have an impact on the results [8].

We used multiple imputation to deal with the missing covariate values in our study. Another advocated method to deal with missing values is the maximum likelihood estimation (eg, using the expectation-maximization [EM]-algorithm). However, maximum likelihood estimations are particularly applied in multilevel or repeated-measurement analysis in which variables are documented more than once. In this study, we focus on a situation where the covariates and outcome are measured once for which multiple imputation is the advocated method [2,3,5–8,13,25,26]. We used the software package Mice for multiple imputation in this study [21]. Several other packages exist. For a comparison of frequently used software packages for multiple imputation, we refer to the literature [27,28].

In conclusion, we showed that complete case analysis can lead to biased study results, which is in agreement with previous studies. Our results also show that simply dropping a covariate or confounder with missing values from the analysis can lead to seriously biased regression coefficients of the remaining covariates, which is particularly harmful in etiologic studies. Further, the discriminative ability of a multivariable prediction model can become much lower when a predictor with missing values is dropped from the analysis. We do not want to suggest that researchers can put less effort to collect as many data as possible. However, every researcher faces the problem of missing values, irrespective of these efforts. To ‘‘provide’’ data according to the strict methodology of multiple imputation seems a better alternative than to give up valuable observed data. We emphasize that the purpose of multiple imputation is not to make up or gain data but to preserve real, observed data. Therefore, we advise to consider multiple imputation when dealing with missing covariate data in medical research.

Appendix

Missing data were generated using a MAR strategy: the probability that an observation was missing was dependent

Table A
Odds ratios of a missing value in the D-dimer test result

	DVT not present	DVT present
Difference in circumference <3 cm	16	1
Difference in circumference ⩾3 cm	5	10

on the other observed patient characteristics. The introduction of missing data in the D-dimer level depended on the presence of DVT and a difference in calf circumference of 3 cm or more. Therefore, we could distinguish four groups; patients without DVT with a difference in circumference of 3 cm or more, patients without DVT without a difference in circumference of 3 cm or more, patients with DVT with a difference in circumference of 3 cm or more, and patients with DVT without a difference in circumference of 3 cm or more. The number of observations in each group depended on the simulated dataset because the presence of DVT was newly generated in each simulation. The odds ratios of a missing response are presented in Table A. The advantage of using odds ratios is that these can be scaled up and down based on the number of missing data that need to be generated.

In the 500 simulated datasets, the values of the independent variables remained identical to the values in the original study sample; only the value of the outcome variable, DVT present (yes/no), was simulated. This approach was chosen because regression analysis is conditional on the values of the independent variables and not on the outcome or dependent variable. The true regression coefficients (Table 1) of the variables and the patient's values were used to calculate the probability of DVT (P) for each patient. Next, for each patient a random number from a uniform distribution in the interval [0,1] was sampled. An event status of 1 (DVT present) was assigned when P was larger than the random number and an event status of 0 otherwise. As a result, from all patients with a DVT probability (P) of 0.3, 30% of those have been assigned an outcome value of 1. We applied the mechanism to introduce missing values as described above in each simulated dataset. All simulations were performed using R2.8.1. The simulation scripts are available on request.

References

[1] Burton A, Altman DG. Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. Br J Cancer 2004;91:4–8.

[2] Little RJ. Regression with missing X's: a review. J Am Stat Assoc 1992;87:1227–37.

[3] Little RJ, Rubin DB. Statistical analysis with missing data. Hoboken, NJ: John Wiley & Sons; 1987.

[4] Little RJ. Methods for handling missing values in clinical trials. J Rheumatol 1999;26:1654–6.

[5] Rubin DB. Multiple imputation for nonresponse in surveys. Hoboken, NJ: John Wiley & Sons; 1987.

[6] Rubin DB. Multiple imputation after 18+ years. J Am Stat Assoc 1996;91:473–89.

[7] Schafer JL. Analysis of incomplete multivariate data. Boca Raton, FL: Chapman & Hall/CRC; 1997.

[8] Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods 2002;7:147–77.

[9] Barnes SA, Lindborg SR, Seaman JW Jr. Multiple imputation techniques in small sample clinical trials. Stat Med 2006;25: 233–45.

[10] Clark TG, Altman DG. Developing a prognostic model in the presence of missing data: an ovarian cancer case study. J Clin Epidemiol 2003;56:28–37.

[11] Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 2006;59:1087–91.

[12] Faris PD, Ghali WA, Brant R, Norris CM, Galbraith PD, Knudtson ML. Multiple imputation versus data enhancement for dealing with missing data in observational health care outcome analyses. J Clin Epidemiol 2002;55:184–91.

[13] Harrell FE Jr. Regression modelling strategies. New York, NY: Springer-Verlag; 2001.

[14] Kmetic A, Joseph L, Berger C, Tenenhouse A. Multiple imputation to account for missing data in a survey: estimating the prevalence of osteoporosis. Epidemiology 2002;13:437–44.

[15] Moons KG, Donders RA, Stijnen T, Harrell FE Jr. Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol 2006;59:1092–101.

[16] Janssen KJ, Vergouwe Y, Donders AR, Harrell FE Jr, Chen Q, Grobbee DE, et al. Dealing with missing predictor values when applying clinical prediction models. Clin Chem 2009;55:994–1001.

[17] Klebanoff MA, Cole SR. Use of multiple imputation in the epidemiologic literature. Am J Epidemiol 2008;168:355–7.

[18] Oudega R, Moons KG, Hoes AW. Ruling out deep venous thrombosis in primary care. A simple diagnostic algorithm including D-dimer testing. Thromb Haemost 2005;94:200–5.

[19] Oudega R, Moons KG, Hoes AW. Limited value of patient history and physical examination in diagnosing deep vein thrombosis in primary care. Fam Pract 2005;22:86–91.

[20] Oudega R, Hoes AW, Moons KG. The Wells rule does not adequately rule out deep venous thrombosis in primary care patients. Ann Intern Med 2005;143:100–7.

[21] van Buuren S, Oudshoorn CGM. What is MICE? Available at. http://web.inter.nl.net/users/S.van.Buuren/mi/hmtl/mice.htm 2007.

[22] Hanley J, McNeil B. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983;148:839–43.

[23] Harrell FE Jr, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. Stat Med 1984;3:143–52.

[24] Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychol Methods 2001;6:330–51.

[25] Rubin DB, Schenker N. Multiple imputation in health-care databases: an overview and some applications. Stat Med 1991;10:585–98.

[26] Schafer JL. Multiple imputation: a primer. Stat Methods Med Res 1999;8:3–15.

[27] Horton NJ, Lipsitz SR. Multiple imputation in practice: comparison of software packages for regression models with missing variables. Am Stat 2001;55:244–54.

[28] Horton NJ, Kleinman KP. Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. Am Stat 2007;61:79–90.