ELSEVIER

Journal of Clinical Epidemiology 59 (2006) 1102−1109

Journal of Clinical Epidemiology

Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: A clinical example

Geert J.M.G. van der Heijden[a,b,*], A. Rogier T. Donders[a,c,d], Theo Stijnen[e], Karel G.M. Moons[a]

[a]*Julius Center for Health Sciences and Primary Care, University Medical Center, P.O. Box 80035, 3508 GA Utrecht, The Netherlands*
[b]*Heart Lung Centre Utrecht, University Medical Center, Utrecht, The Netherlands*
[c]*Department of Biostatistics, Utrecht University, Utrecht, The Netherlands*
[d]*Department of Innovation Studies, Copernicus Institute, Utrecht University, Utrecht, The Netherlands*
[e]*Department of Epidemiology and Biostatistics, Erasmus University Medical School, Rotterdam, The Netherlands*

Accepted 10 January 2006

Abstract

Background and Objectives: To illustrate the effects of different methods for handling missing data—complete case analysis, missing-indicator method, single imputation of unconditional and conditional mean, and multiple imputation (MI)—in the context of multivariable diagnostic research aiming to identify potential predictors (test results) that independently contribute to the prediction of disease presence or absence.

Methods: We used data from 398 subjects from a prospective study on the diagnosis of pulmonary embolism. Various diagnostic predictors or tests had (varying percentages of) missing values. Per method of handling these missing values, we fitted a diagnostic prediction model using multivariable logistic regression analysis.

Results: The receiver operating characteristic curve area for all diagnostic models was above 0.75. The predictors in the final models based on the complete case analysis, and after using the missing-indicator method, were very different compared to the other models. The models based on MI did not differ much from the models derived after using single conditional and unconditional mean imputation.

Conclusion: In multivariable diagnostic research complete case analysis and the use of the missing-indicator method should be avoided, even when data are missing completely at random. MI methods are known to be superior to single imputation methods. For our example study, the single imputation methods performed equally well, but this was most likely because of the low overall number of missing values. © 2006 Elsevier Inc. All rights reserved.

Keywords: Missing data; Complete case analysis; Single imputation; Multiple imputation; Indicator method; Bias; Precision

1. Introduction

Missing observations are frequently encountered and occur in all types of studies, no matter how strictly designed or how hard investigators try to prevent them. In diagnostic studies, as in other type of epidemiological studies including clinical trials and repeated measurement surveys, missing data often occur in a selective pattern. Patient referral for subsequent measurements, here diagnostic procedures, is commonly based on prior measurements, here prior test results, certainly when data are obtained from routine care. In diagnostic research this leads to the well-known referral (verification or work-up) bias [1]. Consider, for example, a study among children with neck stiffness. The aim was to quantify which diagnostic test results from patient history and physical examination predict the presence or absence of bacterial meningitis and which blood tests, e.g., leukocyte count or c-reactive protein level, have additional predictive value [2]. Patients who presented with severe signs, such as convulsions and high fever, were more often and quicker referred for additional blood testing, before full completion of patient history and physical examination. On the other hand, for patients presenting with very mild or no symptoms, additional tests were less often done because the physician already ruled out a serious disease early in the diagnostic process. Accordingly, the sample of study subjects with complete data did not represent the group as a whole, and subjects with missing data carried important information on the associations studied.

* Corresponding author. Tel.: +31-30-2509377; fax: +31-30-2505485.
E-mail address: g.vanderheijden@umcutrecht.nl (G.J.M.G. van der Heijden).

0895-4356/06/$ − see front matter © 2006 Elsevier Inc. All rights reserved.
doi: 10.1016/j.jclinepi.2006.01.015

There are three types of missing data [3,4]. When subjects with missing data form a random subset of the study sample (e.g., because a tube with blood material was accidentally broken), missing data are denoted as missing completely at random (MCAR). Whether missing data are MCAR can easily be tested in the data. When missing data occur in relation to observed covariables (such as selective work-up in diagnostic studies) or the outcome, the subjects with missing data are a selective rather than a completely at random subset of the total study population. This pattern of missingness is confusingly called missing at random (MAR). When the reason for a missing value depends on unknown or unobserved information, they are denoted as missing *not* at random (MNAR). Unfortunately, it is impossible to determine from the data whether missing data are MAR or MNAR; this can only be reasoned or speculated [3,4].

Analysis of epidemiological data typically concerns associations between several predictors and an outcome variable using multivariable regression techniques. Most softwares by default exclude every subject from the analysis with at least one missing value on any of the predictors or outcome analyzed. This is called complete case analysis, and it is the most common form of epidemiological analysis. When missing data are MCAR, complete case analysis obviously is inefficient but leads to unbiased associations. However, when missing data are not MCAR, which commonly is the case, it has extensively been argued and shown that complete case analysis is not only inefficient but commonly leads to biased results as well [3–7].

Various methods have been proposed to deal with missing data. Among them is the missing-indicator method, which uses a dummy variable as an indicator for missing data [5,8]. For multilevel and repeated measurement analysis with missing values, maximum likelihood methods as for example in the expectation maximization method, have been proposed. When predictors and outcomes are measured only once (as is common for diagnostic studies), *imputation* of missing values is the advocated approach. In this, missing data are replaced (filled in) by a reasonable estimated value of that variable, commonly a mean value. One may use an unconditional and conditional mean imputation [3,5–7,9]. Unconditional imputation replaces the missing by, for example, the overall variable mean or median from the observed data, or a random value drawn from the subjects with observed data on that variable. Conditional mean imputation replaces the missing by the mean that is estimated from the specific subgroup to which the subject with missing belongs. Conditional mean imputation can be done once (single imputation) or more than once (multiple imputation [MI]). By means of MI, a random component is added to the imputed value, representing uncertainty because the imputed value was not observed but estimated. Single imputation methods are considered to result in unbiased study results (i.e., associations between predictors and outcome) but in an overestimation of the precision (too small standard errors), whereas MI is assumed to yield unbiased results and appropriate standard errors. This notion, however, appears not fully recognized by researchers, because most epidemiological studies still perform complete case analysis. There are only few studies, and certainly no (multivariable) diagnostic studies, that used empirical data in which the various methods to handle missing data have been applied and the results compared [10,11].

Using empirical data from a study among patients suspected of pulmonary embolism (PE), we evaluated which diagnostic test results (predictors) contribute to predicting the presence or absence of PE by handling the missing values on the predictors in five different ways. These included complete case analysis, the indicator method, the unconditional and conditional single imputation, and MI. Our goal was not to provide a technical overview of different methods for dealing with missing data. For this we refer to the literature [3–7,9,10,12,13]. The goal was only to show the effects of the five ''missing data methods'' when applied to an empirical multivariable diagnostic study.

2. Methods

2.1. Design of the example study

For the present analyses we used data from a study on diagnosis of PE for which methods and results have been described [14–16]. In brief, the study included 398 consecutive patients of 18 years or older who were referred to a Dutch hospital because acute PE was clinically suspected. From all patients, first medical history and physical examination were documented. Additional tests included blood gas analysis, chest radiography, and compression ultrasound of the lower extremities. Finally, perfusion–ventilation lung scanning and pulmonary angiography were executed (reference tests) to determine the ''true'' presence or absence of PE, blinded for the other test results. In total, 170 of the 398 patients had PE, a prevalence of 43%. The study protocol was approved by the institutional review boards of the participating hospitals and informed consent was obtained from all patients.

2.2. Diagnostic tests (predictors) under study

We analyzed whether the following predictors independently contributed to the prediction of PE presence or absence: age (years), gender (male/female), duration of symptoms (number of days), period confined to bed (number of days), respiratory rate (breaths/min), cardiac rate (beats/min), Quetelet index (kg/m^2), arterial oxygen pressure (mmHg), arterial carbon–oxygen pressure (mmHg), the presence (yes/no) of leg paresis, leg pain, a family history of deep venous thrombosis (DVT)/PE, crepitations, fever (temperature above $>38.5\,^{\circ}$C), dyspnea, pleura rub, wheezing, palpitations, collapse with or without

unconsciousness, surgery in past 3 months, malignancy, signs of DVT, and of previous history of PE or DVT, leg ultrasound of lower limbs (positive/negative for DVT), and the chest x-ray result (positive/negative for PE). The a priori selection of these predictors was based on previous diagnostic studies [17–19].

2.3. Missing data

There were no missing data for the presence of PE (yes/no; diagnostic outcome variable), whereas there were missing data on various tests or predictors. The frequency of the number of missing data per subject is shown in Table 1. For 246 (62%) subjects there were no missing data and in 152 (38%) there were one or more predictors missing. Most frequently there were two missing predictor values per subject (18%). The frequency of missing data per predictor is shown in Table 2. For 12 predictors no missing data are observed. For the additional tests, leg ultrasound, PaO_2, and $PaCO_2$, respectively, 21%, 22%, and 14% of the data were missing. For cardiac rate, respiratory rate, and Quetelet index, respectively, 5%, 6%, and 7% of the data were missing. For the remainder, less than 5% of the data were missing. It might well be that for subjects with very mild or no symptoms of PE or DVT, additional diagnostic tests were less often done, because the physician did not consider these additional tests to be informative or perhaps even ruled out a serious disease (notably PE) based on previous diagnostic test results obtained earlier in the diagnostic process. Hence, it is not likely that the missing values (certainly not of the additional tests) are MCAR.

2.4. Analyses

2.4.1. Methods for handling missing data

Complete case analysis: In this analysis only the complete cases, i.e., subjects with observed values on all tests or predictors, were included. So, data were analyzed irrespective of missing data, i.e., with case-wise deletion of subjects in which one or more values of the predictors under study were missing.

Missing-indicator method: For each test with a missing value, we assigned a dummy or indicator variable with value 1 if the test result was missing and 0 if the result was present (although one could as well assign any other value than 0) [5,8]. Missing values of the test (predictor variable) were then given the value zero. In the multivariable model, a test with a missing value was always accompanied by its corresponding missing-indicator variable. This prevents that the subject with a missing on any test is deleted from the entire analysis.

Single unconditional mean imputation: We used the Replace Missing Value function in SPSS (version 11.7) to replace missing data by the overall predictor mean derived from the available data on that particular predictor. Thereby, a missing value is replaced by the mean of the observed values for particular variable.

Single conditional mean imputation: We used the Missing Value Analysis function in the analyze menu in SPSS. Using the observed subjects, this method fits a prediction model for each test with a missing value, in which the test with a missing value is the outcome and all the variables (i.e., all predictors plus diagnostic outcome [20]) are included as predictors. In this process, subjects with missing predictors other than the one imputed, were handled according to the method of case-wise deletion that is default in SPSS. The final fitted prediction model was then (automatically) applied to the subjects in which that test result was missing, to estimate a value for the missing given the subjects' covariable and outcome pattern. The missing value was then replaced by this estimate. For dichotomous test results, the imputed values were rounded to 0 or 1. For continuous test results, we constrained the imputed values to the lowest and highest values actually observed.

MI: To multiple impute missing test results we used the *mice* procedure (MI by chained equations), which has been described in detail [13]. In brief, the *mice* procedure assumes that the distribution of each variable (here: test result) with missing data can be modeled on the basis of the other tests plus the diagnostic outcome with logistic regression model if the variable is dichotomous, polytomous logistic regression if it is categorical with three or more categories, and with linear regression if it is continuous. A Gibbs sampler, which has been shown to converge after five iterations, is used to generate MIs by sampling from the posterior predictive distributions of the missing values [13]. The method in fact estimates a distribution of the missing variable, taking all aspects of uncertainty in the imputations into account. From this distribution values are sampled (with predictive mean matching for both continuous and dichotomous predictors) and filled in for the missing test result. This imputation was repeated 10 times.

Table 1
Number (%) of subjects with missing data for predictors

Number of missings	*n*	%
0	246	61.8
1	42	10.6
2	72	18.1
3	24	6.0
4	6	1.5
5	6	1.5
6	2	0.5
Total	398	100

2.4.2. Identifying diagnostic predictors of the target disease

For each of the five methods of handling missing data we aimed to determine which of the potential predictors contribute to estimating the probability of PE presence, using multivariable logistic regression analysis. To do so, we applied the most frequent method encountered in

Table 2
Number (%) of missing values per variable, and distribution of predictors among subjects without and with missing values (100%: $n = 398$)

Variables		Missings *n*, (%)	No missings, $n = 246$ (62%)	⩾1 missing[a], $n = 152$ (38%)	*P*-value
PE (diagnostic outcome variable)	%	0	47	36	0.02[a]
Leg paresis	%	0	6	5	0.53
Leg pain	%	0	14	16	0.59
Dyspnea	%	0	80	66	<0.01[a]
Family history of DVT/PE	%	0	9	13	0.14
Fever	%	0	19	16	0.46
Malignancy	%	0	28	16	<0.01[a]
Gender	%	0	46	40	0.29
Surgery in previous 3 months	%	0	24	16	0.04[a]
Palpitations	%	0	18	16	0.66
Prior DVT	%	0	6	10	0.17
Wheezing	%	0	18	11	0.09
Previous PE	%	0	5	12	0.02[a]
Collapse with or without loss of consciousness	%	1 (0.3)	10	5	0.06
Crepitations	%	1 (0.3)	31	25	0.19
Signs of DVT	%	1 (0.3)	11	7	0.15
Pleural rub	%	1 (0.3)	16	15	0.65
Age (years)	*X* (SD)	2 (0.5)	57 (17)	54 (18)	0.19
Duration of symptoms (days)	*X* (SD)	5 (1.3)	8 (16)	6 (12)	0.22
Period confined to bed (days)	*X* (SD)	8 (2.0)	5 (8)	4 (9)	0.38
Chest x-ray	%	9 (2.3)	43	36	0.17
Cardiac rate (beats/min)	*X* (SD)	19 (4.8)	93 (17)	92 (21)	0.58
Respiratory rate (breaths/min)	*X* (SD)	23 (5.8)	20 (7)	18 (6)	<0.01[a]
Quetelet index	*X* (SD)	29 (7.3)	25 (4)	25 (4)	0.47
Echography of legs	%	55 (13.8)	15	19	0.48
Arterial oxygen pressure (mmHg)	*X* (SD)	84 (21.1)	36 (8)	35 (7)	0.24
Arterial carbon–oxygen pressure (mmHg)	*X* (SD)	86 (21.6)	73 (20)	76 (19)	0.36

X = mean; SD = standard deviation.
[a] Only observed values for variables with ⩾1 missing.

multivariable diagnostic research. We first entered all predictors at once in an overall model, without univariable preselection and without interaction terms or transformations (e.g., using cubic splines or fractional polynomials) of continuous variables. We then used backwards (manual) selection based on the likelihood ratio test for elimination of predictors, with $P_{in} \leqslant 0.157$, $P_{out} > 0.157$ (the Akaike's Information Criterion [AIC] criterion) as thresholds. For each of the five methods this modeling yielded a final model including only the predictors of the outcome that had a (multivariable) *P*-value $\leqslant 0.157$. When using the missing-indicator method, a predictor with a missing value is only meaningful in a model if its corresponding missing indicator is also included. Although there are various analytical approaches proposed for the missing-indicator method, we analyzed pairs of ''predictor-missing indicator variables,'' which were removed together when the *P*-value of the predictor—thus not of the indicator—exceeded the *P*-value threshold of 0.157. Note that the missing indicator itself—i.e., without the original predictor—has no diagnostic importance.

We stress that data were used for illustration purposes only and by no means to develop the optimal diagnostic model or strategy for diagnosis of PE. Our aim was solely to compare the effects of five common methods of handling missing data when using the most popular approach in prediction research, i.e., stepwise (backwards) selection modeling without, e.g., variable transformations, internal validation, or shrinkage techniques. Using the data at hand for selecting the most important predictors commonly leads to an overoptimistic final model, particularly when the number of predictors is much higher than the number of events divided by 10, low alpha levels (e.g., 0.05) for including predictors in the final model, and no internal validation or shrinkage is used [12,21–27]. Unfortunately, a backwards selection process using a (too) large number of predictors and without using internal validation or shrinkage techniques is the commonest approach in multivariable diagnostic and prognostic research. Hence, we explicitly chose this approach—not to support it but—to illustrate the (additional) effects when using improper methods for dealing with missing predictor values. Nevertheless, we refrained from univariable preselection and used a relatively higher *P*-value for inclusion in the final model.

2.4.3. Measures of comparison

To compare the effect of the five methods for handling missing data on the corresponding multivariable prediction model we evaluated the five models on the predictors eventually included in the final model; the direction and

magnitude of the regression coefficients; the standard errors of the regression coefficients (for MI, the standard error per coefficient was estimated using standard formulas, described by Rubin and Little [28,29]); and the area under the receiver operating characteristic curve (ROC area) of the final reduced model [30].

3. Results

Of the152 subjects with at least one missing value 36% ($n = 54$) had PE and of the 246 subjects without a missing value 47% ($n = 116$) had PE. The difference in prevalence of PE for subjects with and without missing data was statistically significant ($P = 0.02$), indicating that the missing data were not MCAR. This was confirmed by comparing the observed values of the predictors for the subjects with at least one missing value to the subjects without any missing values (completely observed subjects). Five predictors were significantly different (at the P-value level of 0.05) across the two subgroups (Table 2). Hence, the missing data were clearly not MCAR. Based on various previous studies [3–7,31], the complete case analysis is expected to produce biased results when data are not MCAR.

3.1. Selected predictors and standard errors

Table 3 shows the regression coefficients plus standard errors for the predictors that were eventually retained in the final model obtained after each studied method for handling the missing predictor values. The following predictors were included in neither of the models. *Predictors without missing data*: leg paresis, leg pain, dyspnea, family history of DVT/PE, malignancy, gender, surgery in previous 3 months, and previous PE. *Predictors with missing data*: crepitations, pleural rub, respiratory rate, Quetelet index, arterial oxygen pressure, and arterial carbon–oxygen pressure.

The main difference existed between the model derived after complete case analysis and the other models. This was to be expected because the missing data were clearly not MCAR. The predictors bed rest, and fever, were included in all models except in the complete case analysis model, whereas signs of DVT and palpitations were only included in the complete case analysis model. In addition, the predictors included in the model fitted with the complete case analysis produced somewhat larger standard errors for most predictors, which was to be expected because of the smaller number of analyzed subjects.

With the exception of previous DVT, which was not included in the model fitted after using MI, the indicator method and the three imputation methods selected the same predictors. The standard errors of these predictors were quite similar, but for the single conditional mean imputation the standard errors were somewhat smaller than for the other imputation methods but this was to be expected [3–7,28,29,31]. For the ''indicator method model,'' the regression coefficients of nearly all the ''missing-indicator variables'' also reached statistical significance (data not presented).

The Pearson correlation coefficients between the different predictors were generally very low (<0.3). Only for the association between leg pain and signs of DVT the correlation coefficient exceeded 0.4 (0.432). Because there were 14 predictors with missing values, the R^2 for the 14 conditional mean imputation models ranged between 0.130 and 0.361 (mean 0.214). Therefore, it was unlikely that collinearity between predictors influenced their inclusion in the final model.

3.2. Direction and value of the regression coefficients

In each model, the directions of the associations of the selected predictors and the outcome were the same. Compared to the other methods, the complete case model

Table 3
The effect of different methods for handling missing data on the regression coefficients (β) and standard errors (SE) of predictors included in the respective diagnostic models

Predictors included in the regression models	Complete case analysis, β (SE)	Indicator method, β (SE)	Unconditional mean imputation, β (SE)	Conditional mean imputation, β (SE)	MI, β (SE)
Echography of legs	2.840 (0.643)	2.355 (0.454)	2.493 (0.445)	1.572 (0.347)	2.302 (0.422)
Cardiac rate	0.054 (0.024)	0.069 (0.021)	0.068 (0.021)	0.058 (0.019)	0.066 (0.021)
Chest x-ray	0.718 (0.317)	0.896 (0.255)	0.788 (0.247)	0.794 (0.239)	0.824 (0.249)
Duration of symptoms	−0.019 (0.013)	−0.020 (0.011)	−0.020 (0.011)	−0.020 (0.011)	−0.019 (0.011)
Age	0.018 (0.009)	0.021 (0.007)	0.020 (0.007)	0.019 (0.007)	0.019 (0.007)
Collapse with or without loss of consciousness	1.164 (0.571)	1.377 (0.506)	1.231 (0.471)	1.260 (0.458)	1.221 (0.467)
Wheezing	−1.061 (0.437)	−0.595 (0.359)	−0.590 (0.349)	−0.614 (0.341)	−0.573 (0.351)
Period confined to bed		0.032 (0.015)	0.037 (0.015)	0.040 (0.014)	0.031 (0.015)
Fever		−0.649 (0.332)	−0.618 (0.328)	−0.519 (0.316)	−0.546 (0.336)
Prior DVT		0.867 (0.519)	0.726 (0.497)	0.673 (0.341)	
Signs of DVT	1.018 (0.507)				
Palpitations	0.769 (0.436)				
Area under the Receiver Operator Characteristic Curve	***0.794***	***0.813***	***0.775***	***0.792***	***0.787***

resulted in slightly smaller regression coefficients for predictors with few (<5%) or no missing data and an inflated coefficient for leg ultrasound (13% missing data). The regression coefficients of most predictors were somewhat larger in the model obtained when using the missing-indicator method for imputation. This was expected because this model included many more predictors (i.e., the extra missing-indicator variables) in the final model [21].

3.3. ROC area

For all models the ROC area was above 0.75 (Table 3). Because of inclusion of the (in itself clinically meaningless) indicator variables, the model fitted after using the missing-indicator method reached the largest ROC area (0.813). The models after single conditional imputation and MI reached the most conservative estimate.

4. Discussion and conclusion

Missing data provide a challenge in design and analyses of (clinical) epidemiological studies. In multivariable diagnostic research the aim is often to determine the predictors that independently contribute to predicting the presence or absence of a particular disease in patients suspected of this disease. We illustrated the practical consequences of five well-known methods for handling missing data when using the popular stepwise (backwards) selection approach in multivariable prediction research.

We found that the final model derived after complete case analyses—the most common approach in epidemiological (diagnostic) research—very much deviated from the models obtained after using imputation of missing data in terms of predictors selected, regression coefficients, and corresponding standard errors. In addition, the model fitted when using the—also popular—missing-indicator method, showed higher regression coefficients and higher predictive accuracy than the models derived after using the imputation methods. This particular empirical study did not show large differences between the models obtained after use of single unconditional, single conditional, and MI of missing data. Collinearity between predictors can influence the inclusion of predictors in a final model, and thus lead to different sets of predictors with equal predictive accuracy. However, we believe that this phenomenon cannot explain the observed differences between the models obtained after using complete case analysis or the missing-indicator method vs. the models obtained after using the imputation methods, because Pearson correlation coefficients between the studied predictors were too low to induce collinearity.

Compared to most other (simulation) studies on the value of different methods to handle missing data, we did not start with a complete data set—i.e., without missing data—to subsequently assign missing data to the studied variables. We rather used an existing data set, which already included missing data on various variables. Accordingly, the true underlying value of the missing data was unknown and also the true regression coefficients and predictive accuracy of each variable (uni- nor multivariable) were unknown: there is no ''reference criterion'' available. This can be considered as a disadvantage of the study. However, the study clearly illustrates the differences in results and inference that may occur when different methods of handling missing data are used to empirical diagnostic research data when using stepwise (backwards) multivariable modeling. Moreover, because most published studies are based on complete case analysis, reanalysis of existing data by applying other methods to handle missing data may change (inferences of) previously published study results.

Based on simulation studies and profound theoretical reasoning, it is widely advocated that imputation of missing data is better than ignoring missing data, that the indicator method often provides biased results, that conditional imputation is better than unconditional imputation, and that multiple is better than single imputation [3–7,9,12,28,29,31–34]. This empirical study confirms that the missing-indicator method provides an overestimated ROC area. This is because of the inclusion of significant but clinically meaningless missing-indicator variables in the final model. The complete case analysis, i.e., ignoring missing data by case-wise deletion, yields different results compared to the models obtained after imputation of missing data. It was also expected that the model obtained after unconditional single imputation would yield different (biased) regression coefficients compared to the model obtained after conditional single imputation, and that single conditional imputation would yield similar regression coefficients as MI but too small standard errors. Many statistical reports have convincingly shown that MI should be preferred over single imputation. In our example study, however, there were no striking differences in direction, magnitude, and precision of the regression coefficients for the prediction models fitted after using the three methods of imputation. This is most likely because most of our predictors revealed a relatively limited number of missing values (Tables 1 and 2). Hence, the different imputation methods contained a majority of subjects with identical data. More empirical studies are needed to illustrate that with larger numbers of missing data the differences will become more profound and in favor of MI. Apparently, MI does not provide superior results under all circumstances. Limited differences in the impact of unconditional and conditional single imputations were also found in previous studies [35]. To aid applied medical researchers, methodologists may focus more on definitions of more general guidelines on when to use multiple rather than single (conditional) imputation.

Development and applicability of a multivariable prediction model using a backwards variable selection process and without proper handling of missing data deserve some comments. First, prediction models derived by means of the missing-indicator method or complete case analysis

complicate the interpretation and applicability of such models in practice [20]. They can only be interpreted validly in view of the mechanism that produced the missing values and applied to settings with a similar mechanism. However, the mechanism or reason of missingness is usually unknown. Thereby, the clinical value of prediction models developed with use of the missing-indicator method is particularly limited. Because the reason for missing predictor values is commonly related to other predictors and—though often indirectly—to the outcome, the missing indicator itself is often a significant predictor of the outcome under study and needs to be retained in the final prediction model to ensure model interpretation, as was the case in our study. But the missing indicator has no clinical meaning. Applying such prediction model to practice merely implies that waiving a particular predictor paradoxically will contribute significantly to the prediction of the presence of the disease. This, of course, hampers the use of the missing-indicator method in prediction research, in contrast to the use of single or multiple conditional imputation of missing data [20]. We note that the missing-indicator method was developed in the context of etiologic studies, where inclusion of a confounder together with its indicator for being missing increased statistical power without compromising the interpretation of the regression coefficient of the causal determinant under study [5,8].

Second, when using the missing-indicator method we tested each predictor plus missing indicator as a pair. We repeated this analysis in which we did not include the missing indicator in the test for inclusion in the model but only the predictor in question using the partial chi-square test—always adjusted for the corresponding missing indicator. This yielded similar results in terms of final models and did not change our inferences.

Finally, we applied the backwards selection approach—starting from a relatively large number of predictors, and without using transformations of continuous variables, internal validation, or shrinkage techniques—to determine which predictors independently contribute to the prediction of the outcome. As said, this was not done to support such analytical approach—because it commonly yields incorrect and too optimistic prediction models [12,21−27]. We simply used it because of its popularity in multivariable prediction research and to empirically illustrate the additional problems when not using imputation methods for handling missing predictor values.

From this single empirical study we cannot yet draw general conclusions on the most proper method of handling missing data in multivariable prediction research, such as prognostic or diagnostic research. But, when missing data occur in such research, imputation is better than ignoring—i.e., complete case analyses—and better than the use of the missing-indicator method. MI methods are known to have superior properties. For our empirical example data, single imputation performed equally well, most likely because of the low overall number of missing values.

Acknowledgment

We gratefully acknowledge the support by The Netherlands Organization for Scientific Research (ZON-MW 904-10-006 and 917-46-360).

References

[1] Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med 1978;299:926−30.
[2] Oostenbrink R, Moons KG, Bleeker SE, Moll HA, Grobbee DE. Diagnostic research on routine care data: prospects and problems. J Clin Epidemiol 2003;56:501−6.
[3] Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods 2002;7:147−77.
[4] Rubin DB. Inference and missing data. Biometrika 1976;63:581−92.
[5] Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol 1995;142:1255−64.
[6] Vach W. Some issues in estimating the effect of prognostic factors from incomplete covariate data. Stat Med 1997;16:57−72.
[7] Little RJA. Regression with missing X's: a review. J Am Stat Assoc 1992;87:1255−64.
[8] Miettinnen OS. Theoretical epidemiology. New York: Delmar Publishers; 1985.
[9] Schafer JL. Analysis of incomplete multivariate data. London: Chapman & Hall; 1997.
[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] Twisk J, de Vente W. Attrition in longitudinal studies. How to deal with missing data. J Clin Epidemiol 2002;55:329−37.
[12] Harrell FE. Regression modeling strategies. New York: Springer-Verlag; 2001.
[13] van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. Stat Med 1999;18:681−94.
[14] van Beek EJ, Kuyer PM, Schenk BE, Brandjes DP, ten Cate JW, Buller HR. A normal perfusion lung scan in patients with clinically suspected pulmonary embolism. Frequency and clinical validity. Chest 1995;108:170−3.
[15] Turkstra F, Kuijer PM, van Beek EJ, Brandjes DP, ten Cate JW, Buller HR. Diagnostic utility of ultrasonography of leg veins in patients suspected of having pulmonary embolism. Ann Intern Med 1997;126:775−81.
[16] van Beek EJ, Kuijer PM, Buller HR, Brandjes DP, Bossuyt PM, ten Cate JW. The clinical course of patients with suspected pulmonary embolism. Arch Intern Med 1997;157:2593−8.
[17] Miniati M, Prediletto R, Formichi B, Marini C, Di Ricco G, Tonelli L, et al. Accuracy of clinical assessment in the diagnosis of pulmonary embolism. Am J Respir Crit Care Med 1999;159:864−71.
[18] Stollberger C, Finsterer J, Lutz W, Stoberl C, Kroiss A, Valentin A, et al. Multivariate analysis-based prediction rule for pulmonary embolism. Thromb Res 2000;97:267−73.
[19] Wells PS, Ginsberg JS, Anderson DR, Kearon C, Gent M, Turpie AG, et al. Use of a clinical model for safe management of patients with suspected pulmonary embolism. Ann Intern Med 1998;129:997−1005.
[20] Moons KG, Donders AR, Stijnen Th, Harrell FE. Imputation of missing predictor values using the outcome was preferred. J Clin Epidemiol 2006.
[21] Van Houwelingen JC, le Cessie S. Predictive value of statistical models. Stat Med 1990;9:1303−25.
[22] Steyerberg EW, Eijkemans MJ, Harrell FE Jr, Habbema JD. Prognostic modelling with logistic regression analysis: a comparison of

selection and estimation methods in small data sets. Stat Med 2000;19:1059−79.

[23] Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15:361−87.

[24] Sun GW, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. J Clin Epidemiol 1996;49:907−16.

[25] Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KG. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. J Clin Epidemiol 2003;56:441−7.

[26] Moons KG, Donders AR, Steyerberg EW, Harrell FE. Penalized maximum likelihood estimation to directly adjust diagnostic and prognostic prediction models for overoptimism: a clinical example. J Clin Epidemiol 2004;57:1262−70.

[27] Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. J Clin Epidemiol 1995;48:1503−10.

[28] Rubin DB. Multiple imputation for non response in surveys. New York: Wiley; 1987.

[29] Little RJA, Rubin DB. Statistical analysis with missing data. New York: Wiley; 1987.

[30] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29−36.

[31] Donders AR, Van der Heijden GJMG, Stijnen Th, Moons KGM. A gentle introduction to imputation of missing values. J Clin Epidemiol 2006.

[32] Crawford SL, Tennstedt SL, McKinlay JB. A comparison of analytic methods for non-random missingness of outcome data. J Clin Epidemiol 1995;48:209−19.

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

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

[35] Vach W, Blettner M. Biased estimation of the odds ratio in case−control studies due to the use of ad hoc methods of correcting for missing values for confounding variables. Am J Epidemiol 1991; 134:895−907.