Clinical Chemistry 55:5
994–1001 (2009)

Dealing with Missing Predictor Values When Applying Clinical Prediction Models

Kristel J.M. Janssen,[1*] Yvonne Vergouwe,[1] A. Rogier T. Donders,[2] Frank E. Harrell, Jr.,[3] Qingxia Chen,[3] Diederick E. Grobbee,[1] and Karel G.M. Moons[1]

BACKGROUND: Prediction models combine patient characteristics and test results to predict the presence of a disease or the occurrence of an event in the future. In the event that test results (predictor) are unavailable, a strategy is needed to help users applying a prediction model to deal with such missing values. We evaluated 6 strategies to deal with missing values.

METHODS: We developed and validated (in 1295 and 532 primary care patients, respectively) a prediction model to predict the risk of deep venous thrombosis. In an application set (259 patients), we mimicked 3 situations in which (1) an important predictor (D-dimer test), (2) a weaker predictor (difference in calf circumference), and (3) both predictors simultaneously were missing. The 6 strategies to deal with missing values were (1) ignoring the predictor, (2) overall mean imputation, (3) subgroup mean imputation, (4) multiple imputation, (5) applying a submodel including only the observed predictors as derived from the development set, or (6) the "one-step-sweep" method. We compared the model's discriminative ability (expressed by the ROC area) with the true ROC area (no missing values) and the model's estimated calibration slope and intercept with the ideal values of 1 and 0, respectively.

RESULTS: Ignoring the predictor led to the worst and multiple imputation to the best discrimination. Multiple imputation led to calibration intercepts closest to the true value. The effect of the strategies on the slope differed between the 3 scenarios.

CONCLUSIONS: Multiple imputation is preferred if a predictor value is missing.

Clinical prediction models or risk scores are developed to estimate a patient's risk of having (diagnosis) or developing (prognosis) a particular outcome. Well-known examples are the Apgar score *(1)* to estimate the prognosis of newborns and the Framingham risk score *(2)* to predict heart disease. Usually, 3 consecutive phases can be distinguished in clinical prediction research: derivation of the prediction model, validation of the model in new subjects (testing), and application in daily practice *(3–6)*.

Studies aimed at deriving or validating a prediction model commonly are negatively affected by missing values in one or more predictors. Often researchers conduct a so-called complete case analysis, neglecting the data of patients with missing values. Furthermore, predictors with (many) missing values are frequently excluded or replaced by a reference value. These approaches lead not only to loss of power (complete case analysis), but also to biased estimates of diagnostic or prognostic accuracy *(7–14)*. A more advanced method is multiple imputation *(7–16)*. This technique uses all observed patient information to multivariately impute the missing predictor values, which leads to more valid results *(7–12, 15, 16)*.

Physicians who apply prediction models to their patients may also face the problem of a missing predictor value. It is unclear how to deal with missing predictor values in individual patients. For example, a model to predict the presence of a bacterial infection in children with acute fever includes the predictor "duration of fever" *(17)*; however, the parents may not remember the exact duration of the fever. Applying the prediction model without this predictor is not a sound solution, as the relative weights of the other predictors in the model become invalid. We compared 6 strategies to deal with missing values when applying a prediction model to individual patients. We used the empirical data of a

[1] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, the Netherlands; [2] Department of Epidemiology, Biostatistics and Health Technology Assessment, Radboud University Nijmegen Medical Center, the Netherlands; [3] Department of Biostatistics, Vanderbilt University Medical School, Nashville, TN.
* Address correspondence to this author at: Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, P.O. Box 85500, 3508 GA Utrecht, The Netherlands. Fax +31 30 2505480; e-mail k.j.m.janssen@umcutrecht.nl.

Received January 14, 2009; accepted February 3, 2009.
Previously published online at DOI: 10.1373/clinchem.2008.115345

prediction model aimed at predicting the presence of deep vein thrombosis (DVT).

Materials and Methods

CLINICAL EXAMPLE

Timely diagnosis of DVT is important because patients with untreated DVT may develop pulmonary embolism, whereas unjustified therapy with anticoagulants poses a risk for major bleeding *(18)*. Physicians have to decide which patients need to be referred for further workup and which can be safely kept under their surveillance without further workup. A diagnostic prediction model could aid physicians in this decision.

For this analysis, we used data from a large cohort of 2086 primary care patients suspected of DVT as described in previous studies *(19–22)*. Because prediction models are first developed from a so-called derivation data set, then tested in a (usually smaller) validation set, and finally applied in daily practice *(3–6)*, we have split our cohort into a derivation, a validation, and an application set. These 3 datasets have been described in previous studies *(19, 20, 23)*. For the purpose of the current study, we completed the missing values in the data with regression imputation. As a result, there were no missing values in the data sets.

DERIVATION AND VALIDATION OF THE PREDICTION MODEL

The derivation set consisted of 1295 patients included in the period between January 2001 and May 2003 (Table 1). After information was obtained on patient history, physical examination, and D-dimer test, all patients were referred for ultrasonography as a reference standard to document the true presence or absence of DVT. The prediction model was developed with multivariable logistic regression. Model reduction (stepwise backward) was performed with a P value >0.157 according to Akaike Information Criterion *(4, 24, 25)*, and the final model included 7 predictors:

$$\log\left(\frac{risk\ of\ DVT}{1 - risk\ of\ DVT}\right) = \text{linear predictor}$$
$$= -14.84 + 0.81^{*}Absence\ of\ a\ leg\ trauma - 0.02^{*}Age + 0.39^{*}Vein\ distension - 0.02^{*}Duration\ of\ symptoms + 0.34^{*}Immobilization + 0.80^{*}\text{Log}(difference\ in\ calf\ circumference) + 1.72^{*}\text{Log}(D\text{-}dimer\ test) \quad (1)$$

where −14.84 is the so-called intercept and the other numbers the regression coefficients of each predictor.

Table 1. Patient characteristics in the derivation set, the validation set, and the application set.[a]

	Derivation	Validation	Application
n	1295	532	259
Mean age, years (SD)	60 (18)	60 (17)	59 (18)
Male	465 (36)	217 (41)	86 (33)
Oral contraceptive use	129 (10)	46 (9)	29 (11)
Mean duration of symptoms, days (SD)	8 (9)	7 (6)	9 (12)
Leg trauma absent	1104 (85)	438 (82)	213 (82)
Malignancy present	77 (6)	19 (4)	18 (7)
Immobilization	172 (13)	65 (12)	41 (16)
Recent surgery	163 (13)	66 (12)	31 (12)
Swelling whole leg	580 (45)	231 (43)	121 (47)
Vein distension	233 (18)	89 (17)	48 (19)
Log(calf circumference)[a]	1.14 (0.59)	1.13 (0.58)	1.16 (0.57)
Log(D-dimer level)[a]	6.80 (1.21)	6.93 (1.15)	6.66 (1.28)
DVT present	289 (22)	91 (17)	35 (14)

[a] Data are n (%) unless noted otherwise.

The risk of DVT in an individual patient (scale 0%–100%) can be calculated by

$$risk = \frac{1}{1 + e^{-linear\ predictor}} * 100\% \quad (2)$$

We validated this prediction model in the second part of our data, i.e., 532 equally selected and measured patients, included in the period between June 2003 and June 2005 (see also Supplemental Data Section 1, which accompanies the online version of this article at http://www.clinchem.org/content/vol55/issue5).

APPLICATION OF THE PREDICTION MODEL

The application set consisted of the last 259 consecutive patients (Table 1). This application set did not contain any missing predictor values and served as the reference situation. We then mimicked 3 scenarios in which predictor values were missing. First, the D-dimer test (strongest predictor, see Table 3) was missing for all patients. Second, the difference in calf circumference (weaker predictor) was missing for all patients. Third, both predictors were missing for all patients.

STRATEGIES TO DEAL WITH MISSING VALUES

We compared 6 strategies (Table 2) that can deal with missing predictor values when a prediction model is applied to individual patients. The first 4 strategies impute the missing value, in which case the original prediction model can be applied. The last 2 strategies use a

Table 2. Strategies to handle missing predictor values when a prediction model is applied to individual patients.

Missing predictor values imputed	
1. Zero imputation	The missing predictor value is imputed with zero, implying that the predictor is ignored and the prediction model is applied with the unadjusted regression coefficients of the original model.
2. Mean imputation	The missing predictor value is imputed with the mean value, estimated in the derivation set.
3. Subgroup mean imputation	The missing predictor value is imputed with a subgroup mean value, estimated in the derivation set. The subgroups were determined by sex and 5 age categories.
4. Multiple imputation	The missing predictor value is imputed with multiple imputation techniques. Each patient record with missing predictor values was individually merged to the derivation set to apply multiple imputation.
Prediction model adjusted	
5. Submodel without predictors with missing values, derived in the derivation set	The submodel contains adjusted regression coefficients that are estimated in the derivation set.
6. Submodel without predictor(s) with missing values, derived with one-step-sweep	The submodel contains adjusted regression coefficients that are estimated with the original regression coefficients and covariance matrix.

modified prediction model (a submodel without the unobserved predictors). In these submodels, the intercept and regression coefficients of the remaining (observed) predictors are adjusted for the exclusion of the unobserved predictors. The submodels are derived either from the data of the derivation set or by a method called one-step-sweep.

1. Imputation of the value zero. The missing predictor value was imputed with the value zero. For example, in the first scenario, this means that the D-dimer test is neglected, whereas the intercept and regression coefficients of the remaining predictors in Formula 1 are used without adjustments.

2. Overall mean imputation. The missing predictor value was imputed with the mean value of the predictor, estimated from the derivation set. For example, if the D-dimer test was missing (first scenario), the mean log(D-dimer test) of the patients in the derivation set was imputed.

3. Subgroup mean imputation. The missing predictor value was imputed with a subgroup mean value, estimated from the derivation set. Subgroups were determined by sex and 5 age categories. For example, if the D-dimer test was missing for a male patient of 44 years old, the mean log(D-dimer test) of male patients between 40 and 50 years of age in the derivation set was imputed.

4. Multiple imputation. Multiple imputation (see also online Supplemental Data Section 2) is a more advanced method that uses regression models to estimate multiple values of the missing predictor, based on the observed predictors or characteristics of that patient *(7–16)*. Multiple imputation is straightforward and feasible when analyzing a whole dataset. To use this method when applying a prediction model to an individual patient, however, is less straightforward. One needs to have access to the data of the derivation set, for example via a website. Hence, the individual patient with a missing value is added to the derivation set, and the missing predictor value is (multiple) imputed. In this study, we imputed 10 values of the missing predictor for each patient. Then we calculated 10 linear predictors (Formula 1) for each patient, which we subsequently averaged to obtain the patient's risk of DVT presence (Formula 2).

5. Submodel derived from the derivation set. The submodel, including only the observed predictors, was derived in the derivation set.

6. Submodel derived by one-step-sweep. The submodel, including only the observed predictors, was derived with a noniterative 1-step approximation called the one-step-sweep that has been proposed recently *(26)* (see also online Supplemental Data Section 3). This method can be applied without using the individual patient data of the derivation set. The regression coefficients of the submodel are based on the regression coefficients of the original model (Formula 1) and the covariance matrix obtained from the derivation set.

Table 3. Intercept and regression coefficient (SE) of the predictors of the original prediction model (applied in strategies 1–4) and the submodels without the predictor(s) with missing values, derived in the derivation set (strategy 5) or with one-step-sweep (strategy 6).

	Missing predictor(s)						
	None	D-dimer level		Calf circumference		D-dimer level and calf circumference	
	Original model	Derivation set	One-step-sweep	Derivation set	One-step-sweep	Derivation set	One-step-sweep
Predictor	Strategies 1–4	Strategy 5	Strategy 6	Strategy 5	Strategy 6	Strategy 5	Strategy 6
Intercept	−14.84 (0.95)	−2.89 (0.38)	−2.66 (0.46)	−13.83 (0.93)	−13.50 (0.93)	−2.12 (0.33)	−1.65 (0.41)
Absence of trauma	0.81 (0.27)	0.41 (0.23)	0.54 (0.27)	0.64 (0.26)	0.63 (0.27)	0.28 (0.22)	0.39 (0.27)
Age	−0.02 (0.01)	0.001 (0.004)	−0.005 (0.005)	−0.013 (0.005)	−0.02 (0.005)	0.005 (0.004)	−0.002 (0.005)
Vein distension	0.39 (0.21)	0.39 (0.17)	0.46 (0.21)	0.38 (0.21)	0.38 (0.21)	0.42 (0.16)	0.48 (0.21)
Duration of symptoms	−0.02 (0.01)	−0.009 (0.01)	−0.01 (0.01)	−0.017 (0.01)	−0.02 (0.01)	−0.01 (0.01)	−0.01 (0.01)
Immobilization	0.34 (0.27)	0.02 (0.22)	−0.07 (0.27)	0.33 (0.26)	0.33 (0.27)	0.04 (0.21)	−0.08 (0.27)
Log(calf circumference)	0.80 (0.16)	0.73 (0.13)	0.85 (0.16)				
Log(D-dimer level)	1.72 (0.12)			1.68 (0.12)	1.68 (0.12)		

PREDICTIVE ACCURACY MEASURES

We estimated the accuracy of the 6 strategies by quantifying the discrimination and calibration, and compared it with the reference situation (no missing values in the application set). Discrimination is the ability of the model to distinguish between patients with and without DVT, quantified with the area under the ROC curve *(27)*. An ROC area can range from 0.5 (no discrimination) to 1.0 (perfect discrimination) *(28)*. Calibration refers to the agreement between the predicted probabilities and observed frequencies of DVT. It can be graphically assessed with a calibration plot with the predicted probabilities on the *x* axis and the observed frequencies on the *y* axis. The plot shows a line that can be described with a so-called calibration slope and calibration intercept (estimated by fitting the linear predictor of the model as the only covariate in a logistic model with DVT as the outcome) *(29, 30)*. The calibration slope and calibration intercept are ideally 1 and 0, respectively. A slope <1 indicates too-optimistic predictions (low predicted probabilities are too low and high predicted probabilities are too high); a slope >1 indicates that predictions are not extreme enough (low predicted probabilities not low enough and high predicted probabilities not high enough). A calibration intercept close to 0 indicates good calibration in the large and means that the mean predicted DVT probability equals the mean observed DVT frequency. A positive calibration intercept indicates (on average) underestimated risks, whereas a negative value indicates overestimated risks. Because the interpretation of this calibration intercept is difficult if the calibration slope is unequal to 1, the calibration intercept is estimated with the slope fixed at 1, implying that the calibration intercept equals the difference between the observed DVT prevalence and the mean predicted risk *(29, 30)*.

Results

STRATEGY 1: ZERO IMPUTATION

This is the only strategy we could apply without additional estimations.

STRATEGY 2: OVERALL MEAN IMPUTATION

The overall mean log(D-dimer test) and log(difference of calf circumference) in the derivation set were 6.83 and 1.14, respectively.

STRATEGY 3: SUBGROUP MEANS

The subgroup means for log(D-dimer test) and log(difference of calf circumference) in the derivation set are presented in online Supplemental Table 1. For both predictors, the subgroup means differed from the overall means (strategy 2) and were higher for men and older patients.

STRATEGY 4: MULTIPLE IMPUTATION

We used the derivation data set and a multiple imputation script (available on request).

STRATEGY 5: SUBMODELS ESTIMATED IN THE DERIVATION SET

Three submodels were derived (Table 3).

Table 4. Effect of the 6 strategies on the discriminative ability of the prediction model in the application set, expressed by the ROC area (95% CI) when the D-dimer value is missing (scenario 1), when differences in calf circumference are missing (scenario 2), and when the 2 predictors are simultaneously missing (scenario 3).[a]

	Predictor with missing data		
	D-dimer	Difference in calf circumference	D-dimer and difference in calf circumference
1. Zero imputation	0.70 (0.61–0.79)	0.89 (0.83–0.96)	0.62 (0.53–0.71)
2. Mean imputation	0.70 (0.61–0.79)	0.89 (0.83–0.96)	0.62 (0.53–0.71)
3. Subgroup mean imputation	0.69 (0.61–0.78)	0.90 (0.83–0.96)	0.64 (0.55–0.74)
4. Multiple imputation	0.77 (0.69–0.84)	0.90 (0.84–0.96)	0.78 (0.71–0.86)
5. Model without predictor(s) with missing values, estimated in the derivation set	0.70 (0.62–0.79)	0.89 (0.83–0.96)	0.64 (0.54–0.74)
6. Model without predictor(s) with missing values, estimated by one step sweep	0.70 (0.61–0.78)	0.89 (0.83–0.96)	0.66 (0.57–0.75)

[a] The ROC area when no data were missing in the application set (reference situation) was 0.90 (95% CI 0.84–0.96).

STRATEGY 6: SUBMODELS ESTIMATED BY ONE-STEP-SWEEP

Three submodels were derived (Table 3). Online Supplemental Table 2 shows the covariance matrix (needed for this strategy) of the regression coefficients of the original model.

ACCURACY OF THE 6 STRATEGIES

Discrimination. In the reference situation (no missing predictor values in the application set), the ROC area of the original prediction model was 0.90 (95% CI 0.84–0.96). With only the D-dimer test missing (scenario 1), the ROC area decreased to approximately 0.70 in all strategies (Table 4), except with multiple imputation (ROC area 0.77). If the difference in calf circumference was missing (scenario 2), the ROC did not decrease in any of the strategies (ROC area 0.89 or 0.90). If both predictors were missing (scenario 3), the ROC area decreased to 0.66 or lower for all strategies, except with multiple imputation (ROC area 0.78). Zero imputation and mean imputation resulted in the largest decrease (ROC area 0.62).

Calibration. In the reference situation, the calibration slope was 1.06. If the D-dimer test was missing (scenario 1), the subgroup mean imputation resulted in a calibration slope (1.02) closest to the reference slope (Table 5). Multiple imputation resulted in a calibration slope <1, indicating too-extreme predictions. The other 4 strategies led to calibration slopes >1, where strategy 5 and 6 (submodels without the predictor with missing values) resulted in the largest deviation from the reference slope. If the difference in calf circumference was missing (scenario 2), all slopes were similar to the reference slope. If the 2 predictors were missing simultaneously (scenario 3), none of the strategies led to calibration slopes close to the reference slope, though subgroup imputation resulted in the smallest deviation (slope 0.94) from the reference situation. All imputation methods (strategies 1–4) resulted in calibration slopes <1.

The intercept of the calibration line in the reference situation was -0.10. In scenario 1, all strategies led generally to insufficient calibration in the large (intercept not equal to 0), apart from multiple imputation (intercept -0.06) (Table 5). Strategy 1, simply neglecting the predictor, led to the worst calibration (intercept 12.48). In scenario 2, calibration was most similar to the reference situation for multiple imputation (intercept -0.04) and for the submodel estimated in the derivation set (intercept -0.03). Also in this case, neglecting the predictor with missing values led to the largest deviation (intercept 0.97). For scenario 3, all strategies generally resulted in insufficient calibration, apart from multiple imputation (intercept 0.01).

Discussion

When applying a prediction model to individual patients, often a particular predictor may not be measured. The question arises how to use the prediction model in such situations. We compared 6 strategies, of which multiple imputation of the missing values led to the most accurate model predictions.

DISCRIMINATION OF THE MODEL

If the strong predictor D-dimer was missing, multiple imputation resulted in a ROC area closest to the reference value, whereas all other methods led to highly underestimated ROC areas. We expected that this would occur if the predictor with missing values was ignored

Table 5. Effect of the 6 strategies on the calibration of the prediction model in the application set, expressed by the slope (95% CI) and the intercept (95% CI) when the calibration slope was fixed at 1, when the D-dimer value is missing (scenario 1), differences in calf circumference are missing (scenario 2), or the two predictors are simultaneously missing (scenario 3).[a]

	Predictor with missing data		
	D-dimer	Difference in calf circumference	D-dimer and difference in calf circumference
1. Zero imputation			
Slope	1.13 (0.52–1.75)	1.05 (0.74–1.36)	0.89 (0.18–1.61)
Intercept[b]	12.48 (12.11–12.84)	0.97 (0.52–1.41)	13.46 (13.10–13.82)
2. Mean imputation			
Slope	1.13 (0.52–1.75)	1.05 (0.74–1.36)	0.89 (0.18–1.61)
Intercept	0.73 (0.36–1.10)	0.05 (−0.39 to 0.50)	0.80 (0.44–1.17)
3. Subgroup mean imputation			
Slope	1.02 (0.47–1.57)	1.06 (0.74–1.37)	0.94 (0.27–1.61)
Intercept	0.72 (0.35–1.09)	0.06 (−0.38 to 0.50)	0.82 (0.46–1.18)
4. Multiple imputation			
Slope	0.76 (0.40–1.12)	1.07 (0.75–1.39)	0.83 (0.47–1.19)
Intercept	−0.06 (−0.38 to 0.40)	−0.04 (−0.49 to 0.41)	0.01 (−0.38 to 0.40)
5. Model without predictor(s) with missing values, estimated in the derivation set			
Slope	1.47 (0.70–2.24)	1.07 (0.76–1.39)	2.14 (0.72–3.55)
Intercept	−0.30 (−0.66 to 0.07)	−0.03 (−0.47 to 0.41)	−0.28 (−0.64 to 0.08)
6. Model without predictor(s) with missing values, estimated by one-step-sweep			
Slope	1.27 (0.62–1.92)	1.04 (0.73–1.34)	2.11 (0.80–3.42)
Intercept	−0.42 (−0.79 to −0.06)	0.11 (−0.34 to 0.55)	−0.42 (−0.78 to −0.06)

[a] The slope and intercept with the slope fixed at 1 when no data were missing in the application set (reference situation) were 1.06 and −0.10, respectively.
[b] Intercept when the calibration slope was fixed at 1.

without adjusting the regression coefficients of the remaining predictors. For imputation of the overall mean, this was also expected, because it does not change the rank order of patients (since every patient receives the same imputed value). Yet this solution is frequently used in medical research. Imputation of subgroup mean can hypothetically improve the model's discrimination, although this was not found in our results. Apparently, the variability in imputed subgroup means (compared to the overall mean) was not large enough. Furthermore, using submodels (derived either from the derivation set or with the one-step-sweep method) that contain only the predictors with observed values can hypothetically better discriminate than strategies that impute the same value for all patients. However, this is less likely if a strong predictor is missing, as for example the D-dimer test in our study. Any submodel without this predictor substantially loses discriminative ability. Indeed, if the value of a relatively weak predictor was missing (scenario 2), all strategies led to ROC areas similar to the reference situation. If both predictors were missing (scenario 3), we found similar or even worse results compared to scenario 1. Apparently, the discriminative ability of our prediction model was largely based on the strong predictor, the D-dimer test.

CALIBRATION OF THE MODEL

In the case of a missing D-dimer test (scenario 1), ignoring the predictive effect of this strong predictor (i.e., imputing the value zero) led to the worst calibration in the large. If this risk-increasing predictor was ignored, all predicted risks were too low. As expected, multiple imputation led to a calibration intercept closest to the reference situation, as this strategy best approached the missing predictor values. Application of the submodels, derived either from the derivation set or by the one-step-sweep method, showed too-high

predicted probabilities (negative intercept), although closer to the reference value than the (subgroup) mean imputation. The effect of these strategies probably depended on the data at hand and may be different in other situations. As expected, imputation of overall mean and subgroup mean improved this calibration intercept, as the missing predictor value is to some extent incorporated in the risk estimation. If the difference in calf circumference was missing (scenario 2), we found better results for all methods, though ignoring the predictor again led to the worst results. If both predictors were missing (scenario 3), we found results similar to those in scenario 1.

The same inferences can be drawn for the calibration slope (Table 5). If the D-dimer test was missing (scenario 1), ignoring the predictor and overall mean imputation resulted in a slope >1, indicating that the predicted probabilities were not extreme enough. This is expected, as the predicted probabilities become more alike (less extreme). Indeed, subgroup mean imputation improved the slope to a value close to 1, as it allows for more variation between patients. Application of the submodels resulted in calibration slopes >1 and with the largest deviation from the reference situation. Again, this probably depended on the data at hand and may be different in other situations. If difference in calf circumference was missing (scenario 2), all slopes were (nearly) equal to the reference slope. If both predictors were missing (scenario 3), largely the same inferences can be drawn as for scenario 1.

METHODOLOGICAL CONSIDERATIONS

First, our results are based on one empirical example. Other datasets with other prediction models predicting other outcomes may show different results. For example, applying the submodels (without the predictor with missing values) may lead to better results if the remaining predictors of the model have a predictive strength similar to the one that is missing. In our study, the D-dimer test was such a strong predictor that estimating a submodel without this predictor inevitably led to less accurate predictions.

Second, to our knowledge, this is the first time that multiple imputation has been studied when a prediction model is applied to individual patients. We calculated 10 linear predictors (Formula 1) for each patient, averaged these, and transformed this average to the probability of presence of DVT. Another option would be to first transform the 10 linear predictors to 10 probabilities, and average these to a probability. Because risks are not normally distributed, we chose the first strategy. Yet elaborate simulation studies in which all potential scenarios can be mimicked may be necessary to choose the ultimate strategy.

Third, multiple imputation resulted in a calibration slope smaller than 1, indicating predicted probabilities that were too extreme, which are often caused by overfitted models. This suggests that the imputation model may have been overfitted. Shrinkage of the imputation model (i.e., adjusting the model for overfitting) may be a possible solution. More research should be conducted on these methodological issues.

Fourth, the 6 strategies vary in applicability. Mean imputation and subgroup mean imputation are easily applicable in daily clinical practice, as these values can easily be added to the appendix of the manuscript presenting the prediction model. Additionally, the submodels derived from the derivation set and the covariance matrix necessary for the one-step-sweep can be presented in a manuscript. However, this can become quite complex if many predictors have missing values. Hypothetically, all 7 predictors of our prediction model can be missing in practice. Accordingly, $2^7 = 128$ submodels would have to be developed. The one-step-sweep can more easily estimate these submodels without the need to develop all the submodels *(26)*. Multiple imputation is the most complex strategy to apply, as the original derivation set and the multiple imputation models need to be stored in such a way that they are publicly available. Storing the data at Internet sites is a good option. Owing to the increasing introduction of electronic patient records in primary and secondary care, with its potential for built-in algorithms, these strategies may be more easily implemented and applied.

Fifth, there may be more strategies to deal with missing values. For example, we could have imputed the missing values with regression models, in which the predictor with missing values is the dependent variable and the other predictors the independent variables. We did not apply this single regression imputation approach, as it is less feasible in practice. For a prediction model with 7 predictors like ours, one would need to develop and store the $6 * 2^7 = 768$ potential regression models.

Sixth, the gain of multiple imputation over single regression imputation is in the correct estimation of the standard errors of the predicted probabilities. We did not take full profit of this advantage, as in our study the interest was not in the confidence intervals of the predicted probabilities but in the predictive accuracy of the model. However, in situations where the confidence intervals of predicted probabilities are of interest, this will be an extra advantage of multiple imputation.

Finally, we could have split our cohort into a derivation set and an application set (ignoring the validation phase), which would have resulted in a larger derivation set. In our study, however, we explicitly wanted to use an in-between validation set to test the accuracy of the newly developed model. Although model validation is always

highly important, it is still rarely applied. We would like to stress that before any clinical prediction model is applied in practice, it needs to be tested in new patients *(5, 6)*.

In conclusion, if a prediction model is applied in individual patients and a predictor is missing, ignoring that predictor is the worst strategy, as the weights of the remaining predictors become incorrect. Imputation of the overall mean does not improve the discrimination, and the estimated risks may be incorrect. Imputation of a subgroup mean may improve the discrimination, although the predicted risks are not necessarily correct. Using a submodel without the predictor can result in a poor discrimination if the predictor with missing values was a strong predictor. We found that multiple imputation resulted in the best discrimination, and, the predicted risks were on average correct. The question of why the models derived by multiple imputation seemed overfitted needs to be addressed in future research.

Author Contributions: *All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.*

Authors' Disclosures of Potential Conflicts of Interest: *Upon manuscript submission, all authors completed the Disclosures of Potential Conflict of Interest form. Potential conflicts of interest:*

Employment or Leadership: F.E. Harrell, Jr., Department of Biostatistics, Vanderbilt University.
Consultant or Advisory Role: F.E. Harrell, Jr., Pfizer, Amgen, Becker Consulting, GlaxoSmithKline, Novartis, and Merck.
Stock Ownership: None declared.
Honoraria: F.E. Harrell, Jr., Johnson & Johnson, Statistics Society of Canada, and American Statistical Association.
Research Funding: F.E. Harrell, Jr., NIH. We gratefully acknowledge the support by the Netherlands Organization for Scientific Research (ZonMw 016.046.360).
Expert Testimony: None declared.

Role of Sponsor: The funding organizations played no role in the design of study, choice of enrolled patients, review and interpretation of data, or preparation or approval of manuscript.

Acknowledgments: We gratefully acknowledge that part of this work has been conducted in the Department of Statistics, Harvard University (Prof. D.B. Rubin), and in the Department of Biostatistics, Vanderbilt University Medical School (Prof. F.E. Harrell, Jr.).

References

1. Auld PA, Rudolp AJ, Avery ME, Cherry RB, Drorbaugh JE, Kay JL, Smith CA. Responsiveness and resuscitation of the newborn: the use of the Apgar score. Am J Dis Child 1961;101:713–24.
2. Kannel WB, McGee D, Gordon T. A general cardiovascular risk profile: the Framingham Study. Am J Cardiol 1976;38:46–51.
3. Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med 2000; 19:453–73.
4. 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.
5. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med 1999;130:515–24.
6. Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. Ann Intern Med 2006; 144:201–9.
7. Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol 1995; 142:1255–64.
8. Little RJ. Regression with missing X's: a review. J Am Stat Assoc 1992;87:1227–37.
9. Little RJ, Rubin DB. Statistical analysis with missing data. Hoboken (NJ): John Wiley & Sons; 1987. 278 p.
10. Rubin DB. Multiple imputation for nonresponse in surveys. Hoboken (NJ): John Wiley & Sons; 1987. 285 p.
11. Rubin DB. Multiple Imputation after 18+ years. J Am Stat Assoc 1996;91:473–89.
12. Schafer JL. Analysis of incomplete multivariate data. London: Chapman & Hall/CRC; 1997. 420 p. (Monographs on statistics and applied probability 72).
13. 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.
14. 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.
15. Rubin DB, Schenker N. Multiple imputation in health-care databases: an overview and some applications. Stat Med 1991;10:585–98.
16. Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods 2002;7: 147–77.
17. Bleeker SE, Moons KG, Derksen-Lubsen G, Grobbee DE, Moll HA. Predicting serious bacterial infection in young children with fever without apparent source. Acta Paediatr 2001;90:1226–32.
18. Hirsh J, Hoak J. Management of deep vein thrombosis and pulmonary embolism: a statement for healthcare professionals. Council on Thrombosis (in consultation with the Council on Cardiovascular Radiology), American Heart Association. Circulation 1996;93:2212–45.
19. 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.
20. Toll DB, Oudega R, Bulten RJ, Hoes AW, Moons KG. Excluding deep vein thrombosis safely in primary care. J Fam Pract 2006;55:613–8.
21. 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.
22. 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.
23. Toll DB, Oudega R, Vergouwe Y, Moons KG, Hoes AW. A new diagnostic rule for deep vein thrombosis: safety and efficiency in clinically relevant subgroups. Fam Pract 2008;25:3–8.
24. Harrell FE, Jr. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York: Springer-Verlag; 2001. 568 p.
25. 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.
26. Marshall G, Warner B, MaWhinney S, Hammermeister K. Prospective prediction in the presence of missing data. Stat Med 2002;21:561–70.
27. Harrell FE, Jr, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. Stat Med 1984;3:143–52.
28. 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.
29. Miller ME, Langefeld CD, Tierney WM, Hui SL, McDonald CJ. Validation of probabilistic predictions. Med Decis Making 1993;13:49–58.
30. Cox DR. Two further applications of a model for binary regression. Biometrika 1958;45:562–5.