Circulation: Cardiovascular Quality and Outcomes

METHODS PAPER

Effect of Variable Selection Strategy on the Performance of Prognostic Models When Using Multiple Imputation

Peter C. Austin, PhD
Douglas S. Lee, MD, PhD
Dennis T. Ko, MD, MSc
Ian R. White, PhD

BACKGROUND: Variable selection is an important issue when developing prognostic models. Missing data occur frequently in clinical research. Multiple imputation is increasingly used to address the presence of missing data in clinical research. The effect of different variable selection strategies with multiply imputed data on the external performance of derived prognostic models has not been well examined.

METHODS AND RESULTS: We used backward variable selection with 9 different ways to handle multiply imputed data in a derivation sample to develop logistic regression models for predicting death within 1 year of hospitalization with an acute myocardial infarction. We assessed the prognostic accuracy of each derived model in a temporally distinct validation sample. The derivation and validation samples consisted of 11 524 patients hospitalized between 1999 and 2001 and 7889 patients hospitalized between 2004 and 2005, respectively. We considered 41 candidate predictor variables. Missing data occurred frequently, with only 13% of patients in the derivation sample and 31% of patients in the validation sample having complete data. Regardless of the significance level for variable selection, the prognostic model developed using only the complete cases in the derivation sample had substantially worse performance in the validation sample than did the models for which variables were selected using the multiply imputed versions of the derivation sample. The other 8 approaches to handling multiply imputed data resulted in prognostic models with performance similar to one another.

CONCLUSIONS: Ignoring missing data and using only subjects with complete data can result in the derivation of prognostic models with poor performance. Multiple imputation should be used to account for missing data when developing prognostic models.

Key Words: death ■ hospitalization ■ incidence ■ myocardial infarction ■ probability

https://www.ahajournals.org/journal/circoutcomes

Prognostic models are mathematical or statistical models that combine information on patient characteristics to produce predictions about future patient outcomes (eg, subsequent mortality or future incidence of heart disease).[1] Prognostic models permit informed clinical decision-making. They permit effective risk stratification so that effective therapies and interventions are targeted at the patients most likely to benefit. Examples of prognostic models include the Framingham Risk Score for predicting cardiovascular disease,[2] the GRACE score for predicting mortality following hospitalization for acute coronary syndromes,[3] and the EFFECT-HF mortality risk score for predicting mortality in patients hospitalized with congestive heart failure.[4]

Selection of variables for inclusion in a prognostic model is an important issue. Clinical knowledge and expertise combined with the existing literature often provide investigators with a lengthy list of candidate predictor variables. To increase use of a prognostic model by clinicians and to reduce the data collection burden on future users, it is often necessary to develop a parsimonious prediction model that uses only a subset of the candidate predictor variables. Despite their limitations, variable selection methods such as backward variable elimination and forward variable selection are popular with applied analysts.[5]

The occurrence of missing data is an important issue when using clinical data. Missing data occur when some variables are only measured on a subset of the subjects. Rubin developed a framework for addressing missing data.[6] The framework is easiest to describe for a single incomplete variable. Data are said to be missing completely at random if the probability that a given variable is missing for a specific subject is unrelated to the value of that variable or of any other variable. Data are said to be missing at random if the probability that a given variable is missing for a specific subject is unrelated to the value of that variable, conditional on the observed values of other variables. Finally, data are said to be missing not at random if the probability that a given variable is missing for a specific subject is related to the value of that variable itself, conditional on the observed values of other variables. Developing a prognostic model using only subjects with complete data (ie, excluding those subjects with any missing data) can have at least 2 possible adverse consequences. First, the estimated standard errors for the regression coefficients would be unnecessarily large, as information would be lost by excluding subjects with any missing data. Second, if the data were missing at random and not missing completely at random, then the estimated regression coefficients could be biased. To address the problems presented by missing data, Rubin developed multiple imputation, which entails the creation of M (M>1) copies of the original sample in which missing data have been filled-in using a model for the missing data.[6] Each of the M imputed datasets is complete, in that missing data are not present. In each of the M imputed datasets, a conventional statistical analysis is conducted. Estimated regression coefficients and their standard errors can be combined using Rubin's Rules, which account for both within- and between-imputation variability.

An important issue when developing prognostic models is the validation of their performance. Validation refers to assessing the performance of the prognostic model in samples other than the one used for model development or derivation. Justice described different types of model validation or transportability.[7] A model is described as displaying geographic transportability if it performs well in geographic locations different from the one in which it was developed. A model is described as displaying temporal transportability if it performs well in time periods different from the one in which it was developed. Before their widespread adoption in clinical practice, it is important that prognostic models undergo validation.

Despite the frequency with which missing data occur in clinical research and the need to develop parsimonious prognostic models, there is paucity of information on how to conduct variable selection when using multiple imputation. The issue is not straightforward because the variables selected by a given variable selection procedure could differ across the different imputed datasets (eg, the variable denoting systolic blood pressure could be selected for inclusion in the first imputed dataset, but not in the second imputed dataset). Wood et al described 9 different methods to conduct variable selection when using multiple imputation and evaluated the performance of these methods (we briefly describe these methods in the following section). Wood et al used Monte Carlo simulations to assess the performance of the 9 different variable selection schemes. They evaluated the variable selection methods in terms of their ability to correctly select variables from the true model and to exclude variables not in the true model.[8]

While the ability to correctly identify variables in the true model is important, equally important is the ability to develop prognostic models that display good performance when validated in independent samples that were not used for model development. The objective of the current study is to evaluate the performance of different variable selection methods for use with multiply imputed data when the evaluation criterion is the prognostic accuracy of the derived models when applied to independent validation samples. The paper is structured as follows: First, we review previously described methods for variable selection when using multiply imputed data. Second, we describe a case study used to compare the prognostic ability of models developed using different variable selection methods. Third, we report the results of our analyses. Finally, we summarize our findings and place them in the context of the existing literature.

STATISTICAL METHODS FOR VARIABLE SELECTION WHEN USING MULTIPLE IMPUTATION

Wood et al conducted a simulation study to examine the performance of different methods for variable selection using backwards variable elimination in multiply imputed data.[8] We describe the variable-selection methods described in their paper, using their terminology when referring to each method, and briefly summarize them in the appendix. While Wood et al examined the use of backwards variable selection, similar approaches could be used with other variable selection methods (eg, forward variable selection or shrinkage-based methods such as the least absolute shrinkage and selection operator[1]).

Variable Selection Using Complete Cases (Complete)

This approach restricts the analytic sample to those subjects with complete data on all candidate variables. Conventional variable selection methods (eg, backwards variable selection) are applied in the sample consisting of all subjects with complete data.

Single Stochastic Imputation (Single)

This approach uses a single imputed dataset for variable selection. For instance, variable selection can be conducted using the first imputed dataset.

Separate Imputations (S1, S2, and S3)

This approach is a modification of the previous approach. Variable selection is conducted separately in each of the M imputed datasets. The analyst notes the variables that were selected for inclusion in each of the M imputed datasets. Once this has been done, there are 3 different approaches to selecting the variable for inclusion in the final prediction model. Approach S1 selects those covariates that were selected in at least one of the M imputed datasets. Approach S2 selects those covariates that were selected in at least half of the M imputed datasets. Approach S3 selects those covariates that were selected in all of the M imputed datasets.

Stacked Imputed Datasets With Weighted Regressions (W1, W2, and W3)

This approach entails stacking the M imputed datasets into 1 large dataset and then conducting variable selection in this single stacked dataset. To account for the multiple observations for each subject, weights are incorporated into the regression model when conducting variable selection. Wood et al proposed 3 different sets of weights that could be used: W1: w=1/M, in which each subject is weighted using the reciprocal of the number of imputed datasets; W2: w=(1−f)/M, where f denotes the proportion of missing data across all variables; W3: w_j=(1−f_j)/M, where f_j denotes the proportion of missing data for variable X_j. Using the third approach, a different set of weights is used when assessing the statistical significance of a given candidate predictor variable.

Application of Rubin's Rules for Variable Selection (RR)

The final approach involves using Rubin's Rules at each stage of variable selection to determine the statistical significance of each candidate predictor variable included in the regression model at a given step in variable selection. Thus, when using backward variable selection, the full model is fit in each of the M imputed datasets and the estimated regression coefficients and their standard errors are pooled using Rubin's Rules. The variable with the largest P value is then excluded from the regression model, and the process is repeated until all retained variables meet a prespecified level of statistical significance (eg, P≤0.05).

Estimation of Regression Coefficients for the Selected Variables

Once the variables have been selected using a given variable selection method, the associated regression coefficients can be estimated in each of the imputed datasets and the regression coefficients and their standard errors can be combined using Rubin's Rules to produce the final model (this step is obviously superfluous in the last variable selection approach, which explicitly applied Rubin's Rules when conducting variable selection).

METHODS

The use of data in this project was authorized under section 45 of Ontario's Personal Health Information Protection Act, which does not require review by a Research Ethics Board. The first author had full access to all the data in the study and takes responsibility for its integrity and the data analysis. The data sets used for this study were held securely in a linked, de-identified form and analyzed at ICES. While data sharing agreements prohibit ICES from making the data set publicly available, access may be granted to those who meet prespecified criteria for confidential access, available at www.ices.on.ca/DAS.

Data Sources

The EFFECT (Enhanced Feedback for Effective Cardiac Treatment) Study was designed to improve the quality of care for patients with cardiovascular disease in Ontario.[9] During the first phase (referred to as the EFFECT Baseline sample), detailed clinical data were collected on patients hospitalized with acute myocardial infarction or congestive heart failure between April 1, 1999 and March 31, 2001 at 85 hospital corporations in Ontario, Canada, by retrospective chart review. During the second phase (referred to as the EFFECT follow-up sample), data were abstracted on patients hospitalized with these conditions between April 1, 2004 and March 31, 2005 at 81 Ontario hospital corporations. Data on patient demographics, vital signs, and physical examination at presentation, medical history, and results of laboratory tests were collected for these samples.

For the current study, we restricted our sample to patients hospitalized with acute myocardial infarction. Data were available on 11 524 and 7889 patients hospitalized with a diagnosis of acute myocardial infarction during the first and second phases of the study, respectively. For the current study, the outcome was a binary variable denoting whether the patient died within 1 year of hospital admission. Candidate predictor variables were those 41 binary and continuous variables listed in Table 1; no categorical variables had >2 levels. The mean/prevalence of each of the 41 covariates along with that of the binary outcome are reported in Table 1 for each of the 2 phases of the study, along with the percentage of subjects with missing data for each variable. The outcome variable was not subject to missing data as it was obtained by deterministic linkage to a population-based registry of the vital status of all residents of Ontario. In the derivation sample, 13% of subjects had complete data (and 87% of subjects were missing information on at least one variable), while in the validation sample 31% of subjects had complete data (and 69% of subjects were missing information on at least one variable).

In the derivation sample, 2310 (20.0%) subjects died within 1 year of hospital admission, while 1590 (20.2%) subjects in the validation sample died within 1 year of hospital admission. Given the use of 41 candidate predictor variables, the number of events per variable was 56 in the derivation sample and 39 in the validation sample.

Statistical Methods

The EFFECT baseline sample was used as the derivation sample for model selection and estimation. The EFFECT follow-up sample was used as the validation sample for assessing the performance of the models estimated in the derivation sample.

Our aim was to examine ideal model performance in settings without missing data.[10] As such, all of the imputation models included the outcome variable.

Multiple imputation was conducted separately in the derivation and validation samples. Imputation was conducted using a fully conditional specification approach using PROC MI in SAS (SAS/STAT version 14.1). Logistic regression models were used as the imputation models for the binary variables, while linear regression models were used as the imputation models for the continuous variables. All variables (including the binary outcome variable) were included in each imputation model (with the exception of the variable that was being imputed). No interactions or nonlinear terms were included. For each of the 2 samples, we set the number of imputed datasets (M) to be equal to the percentage of subjects with any missing data in the given sample.[11] Thus, we created 87 imputed datasets for the derivation sample and 69 imputed datasets for the validation sample.

Variable selection was conducted using the 9 approaches described above (complete, single, S1, S2, S3, W1, W2, W3, and RR). For each approach, we considered 2 different significance levels for variable retention. First, backward variable selection was used with the criterion that the statistical significance of retained variables had to be ≤0.05. Second, backward variable selection was used with the criterion that the statistical significance of the retained variables had to be <0.157, which, for continuous or binary variables, is equivalent to the use of the Akaike Information Criterion.[1] The second criterion was used as the first may be overly restrictive for developing prognostic models. Once the variables were selected using 1 of the 9 variable selection approaches, the coefficients for the final model were estimated in each of the 87 imputed versions of the derivation sample and the regression coefficients combined using Rubin's Rules.

After a final regression model had been selected using each of the 9 approaches and its coefficients estimated, the estimated regression model was applied to each of the 69 imputed versions of the validation sample. The performance of the logistic regression model developed in the derivation sample was assessed in each of the 69 imputed versions of the validation sample. We used 4 different quantitative metrics for assessing the performance of the selected models: the c-statistic (equivalent to the area under the receiver operating characteristic curve), Nagelkerke's generalized R^2 statistic, the scaled Brier score, and the calibration slope.[1,12,13] The Brier score is the mean squared prediction error (with smaller values of the Brier score denoting more accurate prediction). The scaled Brier score is defined as $\text{Brier}_{\text{scaled}} = 1 - \frac{\text{Brier}}{\text{Brier}_{\text{max}}}$, where $\text{Brier}_{\text{max}}$ denotes the maximum possible Brier score. The scaled Brier score ranges from 0% to 100%. In each imputed version of the validation sample, we regressed the observed binary outcome on the linear predictor computed using the regression coefficients estimated in the derivation sample. The calibration slope is the regression coefficient for the estimated linear predictor.

As our aim was to examine ideal model performance in settings without missing data, we evaluated the performance of the derived model in each of the imputed versions of the validation sample, rather than pooling predictions for each subject across the imputed datasets and evaluating performance based on these pooled predictions.[10,14] For each of the 9 variable selection methods and for each of the 4 quantitative measures of model performance, there was no evidence that the distribution of the measure of model performance was non-normal across the imputed datasets (P>0.11 for the 36 applications of the Shapiro-Wilk test of normality when using P=0.05 for the variable selection criterion and when using P=0.157 for the variable selection criterion). Thus, we applied Rubin's Rules and computed the mean of the estimates of model performance (eg, the c-statistic) across the 69 imputed versions of the validation sample.

Loess-based graphical methods were used to assess the calibration of each derived model when applied to the imputed versions of the validation sample.[15] Each model

Table 1. Description of Derivation and Validation Samples

Variable	Mean/Prevalence (%) (Derivation)	Mean/Prevalence (%) (Validation)	*P* Value	% Missing (Derivation)	% Missing (Validation)
Demographic characteristics					
Age	67.62	68.49	<0.001	0.1	0.0
Female	36%	37%	0.288	0.1	0.0
Presenting signs and symptoms					
Cardiogenic shock	2%	0%	<0.001	1.1	0.0
Acute pulmonary edema	6%	7%	0.008	0.9	1.0
Vital signs on admission					
Systolic blood pressure	146.20	142.86	<0.001	0.5	1.6
Diastolic blood pressure	82.58	79.97	<0.001	0.9	1.8
Heart rate	84.45	85.33	0.014	0.8	1.6
Respiratory rate	21.23	20.48	<0.001	7.7	4.8
BMI	27.87	28.00	0.282	48.0	29.7
Cardiac risk factors					
Diabetes mellitus	26%	28%	0.005	0.6	0.2
Hypertension	46%	58%	<0.001	1.2	0.6
Current smoker	38%	31%	<0.001	14.7	11.5
Dyslipidemia	31%	45%	<0.001	4.0	1.9
Family history of heart disease	38%	38%	0.898	20.5	18.0
Comorbid conditions and vascular history					
History of stroke/TIA	10%	12%	<0.001	0.7	0.1
Angina	33%	30%	<0.001	1.5	1.2
Cancer	3%	2%	<0.001	1.6	1.0
Dementia	4%	6%	<0.001	1.1	0.5
Peptic ulcer disease	5%	5%	0.114	1.3	0.8
Previous AMI	24%	24%	0.151	1.8	1.3
Asthma	6%	6%	0.14	1.1	0.5
Depression	7%	10%	<0.001	1.2	1.2
Peripheral arterial disease	8%	9%	0.01	3.0	0.2
Previous revascularization	9%	12%	<0.001	0.5	0.2
Congestive heart failure	5%	6%	0.008	1.1	0.7
Hyperthyroidism	1%	0%	<0.001	1.1	0.0
Aortic stenosis	2%	2%	0.032	1.3	0.5
Initial laboratory tests					
Hemoglobin	137.49	136.53	<0.001	1.6	0.6
White blood count	10.54	10.66	0.12	1.6	0.6
Sodium	138.98	138.71	<0.001	1.7	0.6
Potassium	4.11	4.11	0.862	1.8	0.7
Glucose	9.63	9.21	<0.001	3.7	1.5
Urea	7.86	8.18	<0.001	8.1	4.3
Creatinine	108.78	113.81	<0.001	2.5	0.7
International normalized ratio	1.13	1.15	0.046	16.9	9.8
Total cholesterol	4.93	4.55	<0.001	48.1	29.5
HDL cholesterol	1.10	1.08	0.023	55.9	31.4
LDL cholesterol	3.04	2.68	<0.001	57.9	33.9
Triglycerides	1.98	1.86	<0.001	48.8	30.0

(Continued)

Table 1. Continued

Variable	Mean/Prevalence (%) (Derivation)	Mean/Prevalence (%) (Validation)	*P* Value	% Missing (Derivation)	% Missing (Validation)
Characteristics of AMI					
Elevated cardiac enzymes	94%	98%	<0.001	0.7	0.0
ST-segment elevation MI	65%	49%	<0.001	1.0	1.0
Outcomes					
Death within 1 y	20%	20%	0.852	0.0	0.0

The columns for mean/prevalence report the mean value of the variable for continuous variables and the proportion of subjects with the condition for binary variables. A standard *t* test was used to compare means of continuous variables between the derivation and validation sample, while a χ^2 test was used to compare proportions between samples. AMI indicates acute myocardial infarction; BMI, body mass index; HDL, high-density lipoprotein; LDL, low-density lipoprotein; and TIA, transient ischemic attack.

developed in the derivation sample (and whose coefficients were subsequently estimated using Rubin's Rules) was applied to each of the 69 imputed versions of the validation sample. A predicted probability of the outcome was obtained for each subject in each of these 69 samples. Loess-based graphical methods were used to assess the calibration of the derived model when applied to each of the imputed versions of the validation sample. The resultant 69 calibration curves were then averaged to obtain a final calibration curve.

In clinical practice, simple mean imputation may be used rather than multiple imputation. We therefore also compared model performance in this context. A single imputed version of the validation was created in which missing continuous variables were imputed using the mean of the observed values for that variable, while missing binary variables were imputed using the mode of the observed values for that variable. For each of the variable selection approaches, the regression model selected and estimated in the derivation sample (using the 87 imputed versions of the derivation sample) was applied to this single imputed version of the validation sample. The performance of the fitted model in this single validation sample was assessed using the c-statistic, the generalized R^2 statistic, and the scaled Brier score.

RESULTS

Results are reported separately for the derivation and validation samples. A standard *t* test was used to compare the means of continuous variables between the derivation and validation samples while a χ^2 test was used to compare the distribution of binary variables between the derivation and validation samples. In the derivation sample, the percentage of subjects with missing data for a given variable ranged from 0% to 57.9%, with a median of 1.4% (25th and 75th percentiles: 0.9% and 4.0%). In the validation sample, the percentage of subjects with missing data for a given variable ranged from 0% to 33.9%, with a median of 0.9% (25th and 75th percentiles: 0.5% and 1.9%).

Variable Selection Using *P*=0.05 for Variable Retention

The variables selected using each of the variable selection approaches when a significance level of 0.05 was used for variable selection is reported in Table 2. Six variables were not selected using any of the variable selection approaches, while 9 variables were selected using all 9 variable selection approaches. The numbers of selected variables for the different variable selection methods were: 10 (complete case selection), 21 (S3), 27 (single sample selection and W3), 28 (W2 and Rubin's Rules), 29 (W1), 31 (S2), and 36 (S1) (see first row of Table 2). The estimated odds ratios and associated 95% CIs for the predictor variables in each model are reported in Table 3. The regression coefficients for these models were estimated in each of the imputed versions of the derivation sample and were then pooled using Rubin's Rules. Note that several of the estimated effects are not statistically significant. This is due to the final estimates and CIs for all models being estimated in all imputed datasets and then being pooled using Rubin's Rules. By definition, all of the variables in the model whose variables were selected using applications of Rubin's Rules were statistically significant ($P \le 0.05$). Similarly, the use of the S3 and W3 algorithms resulted in all included variables being statistically significant. However, the following variable selection methods resulted in the inclusion of variables that were not significant after estimation using Rubin's Rules: single (2 variables), S1 (9 variables), S2 (4 variables), W1 (2 variables), and W2 (2 variables).

The regression models reported in Table 3 (derived and estimated in the derivation sample) were then applied to each of the imputed versions of the validation sample. The mean of the model performance metrics (c-statistic, generalized R^2 statistic, scaled Brier score, and calibration slope) across the 69 imputed versions of the validation sample are reported in Table 4. The most notable observation is that the model whose variables were selected using the complete cases had noticeable worse performance than did the other models across all 3 measures of model performance.

The graphical assessment of calibration in the validation sample is described in Figure 1. Deviation of the smoothed calibration plot from the diagonal line with unit slope is indicative of lack of calibration. We have added to this plot a nonparametric estimate of the distribution of the predicted probability of the outcome in the first imputed

Table 2. Variables Selected Using Each Variable Selection Approach When Using a 0.05 Significance Level for Variable Retention

	Variable Selection Method								
Variable	Complete	Single	S1	S2	S3	W1	W2	W3	RR	Total Times Selected
Total number of variables selected	10	27	36	31	21	29	28	27	28	
Acute pulmonary edema	0	0	0	0	0	0	0	0	0	0
Asthma	0	0	0	0	0	0	0	0	0	0
Diastolic blood pressure	0	0	0	0	0	0	0	0	0	0
Hyperthyroidism	0	0	0	0	0	0	0	0	0	0
Previous revascularization	0	0	0	0	0	0	0	0	0	0
Diabetes mellitus	0	0	1	0	0	0	0	0	0	1
Family history of heart disease	0	0	1	0	0	0	0	0	0	1
Female	0	0	1	0	0	0	0	0	0	1
Triglycerides	0	0	1	0	0	0	0	0	0	1
Peptic ulcer disease	0	1	1	0	0	0	0	0	0	2
Total cholesterol	0	0	1	1	0	1	1	0	0	4
HDL cholesterol	0	0	1	1	0	1	1	0	0	4
Current smoker	1	1	1	1	0	0	0	0	0	4
ST-segment elevation MI	0	0	1	1	1	0	0	0	1	4
Cancer	0	1	1	1	0	1	0	1	1	6
Hypertension	0	0	1	1	0	1	1	1	1	6
LDL cholesterol	0	0	1	1	0	1	1	1	1	6
BMI	0	1	1	1	0	1	1	1	1	7
Dyslipidemia	0	1	1	1	0	1	1	1	1	7
Peripheral arterial disease	0	1	1	1	0	1	1	1	1	7
Sodium	0	1	1	1	0	1	1	1	1	7
Angina	0	1	1	1	1	1	1	1	1	8
Creatinine	0	1	1	1	1	1	1	1	1	8
Dementia	0	1	1	1	1	1	1	1	1	8
Depression	0	1	1	1	1	1	1	1	1	8
Elevated cardiac enzymes	0	1	1	1	1	1	1	1	1	8
Heart rate	0	1	1	1	1	1	1	1	1	8
International normalized ratio	0	1	1	1	1	1	1	1	1	8
Potassium	0	1	1	1	1	1	1	1	1	8
Previous AMI	0	1	1	1	1	1	1	1	1	8
Systolic blood pressure	0	1	1	1	1	1	1	1	1	8
White blood cell count	0	1	1	1	1	1	1	1	1	8
Age	1	1	1	1	1	1	1	1	1	9
Aortic stenosis	1	1	1	1	1	1	1	1	1	9
CHF	1	1	1	1	1	1	1	1	1	9
Cardiogenic shock	1	1	1	1	1	1	1	1	1	9
Stroke/TIA	1	1	1	1	1	1	1	1	1	9
Glucose	1	1	1	1	1	1	1	1	1	9
Hemoglobin	1	1	1	1	1	1	1	1	1	9
Respiratory rate	1	1	1	1	1	1	1	1	1	9
Urea	1	1	1	1	1	1	1	1	1	9

‘1’ in a cell indicates that the variable in that row was selected using the method for the given column. ‘0’ in a cell indicates that the variable in that row was not selected using the method for the given column. AMI indicates acute myocardial infarction; BMI, body mass index; CHF, congestive heart failure; HDL, high-density lipoprotein; LDL, low-density lipoprotein; and TIA, transient ischemic attack.

Table 3. Estimated Odds Ratios and 95% CIs in the Derivation Sample When Using a 0.05 Significance Level for Variable Retention

	Variable Selection Method								
Variable	Complete	Single	S1	S2	S3	W1	W2	W3	RR
Age (per 10 y increase)	2.02 (1.91–2.14)	1.96 (1.84–2.09)	1.93 (1.8–2.07)	1.91 (1.79–2.05)	2.01 (1.9–2.12)	1.88 (1.77–2.01)	1.88 (1.77–2.01)	1.9 (1.79–2.02)	1.91 (1.79–2.03)
Female			0.9 (0.78–1.04)						
Cardiogenic shock	9.18 (6.48–13)	5.01 (3.49–7.18)	5.08 (3.51–7.37)	5.04 (3.49–7.28)	4.96 (3.47–7.11)	5.04 (3.49–7.27)	5.05 (3.5–7.29)	5.19 (3.61–7.47)	5.11 (3.55–7.34)
Acute pulmonary edema									
Systolic blood pressure		0.99 (0.98–0.99)	0.99 (0.98–0.99)	0.99 (0.98–0.99)	0.99 (0.98–0.99)	0.99 (0.98–0.99)	0.99 (0.98–0.99)	0.99 (0.98–0.99)	0.99 (0.98–0.99)
Diastolic blood pressure									
Heart rate		1.01 (1.01–1.01)	1.01 (1.01–1.01)	1.01 (1.01–1.01)	1.01 (1.01–1.01)	1.01 (1.01–1.01)	1.01 (1.01–1.01)	1.01 (1–1.01)	1.01 (1.01–1.01)
Respiratory rate	1.04 (1.04–1.05)	1.03 (1.02–1.04)	1.03 (1.02–1.04)	1.03 (1.02–1.04)	1.03 (1.02–1.04)	1.03 (1.02–1.04)	1.03 (1.02–1.04)	1.03 (1.02–1.04)	1.03 (1.02–1.04)
BMI		0.98 (0.97–1)	0.98 (0.96–1)	0.98 (0.97–1)		0.98 (0.96–1)	0.98 (0.96–1)	0.98 (0.96–0.99)	0.98 (0.96–0.99)
Diabetes mellitus			1.08 (0.92–1.25)						
Hypertension			1.15 (1.02–1.3)	1.14 (1.01–1.29)		1.13 (1–1.28)	1.13 (1–1.28)	1.13 (1–1.27)	1.13 (1–1.28)
Current smoker	1.26 (1.09–1.45)	1.13 (0.98–1.31)	1.15 (0.99–1.34)	1.14 (0.98–1.32)					
Dyslipidemia		0.86 (0.75–0.99)	0.85 (0.74–0.98)	0.85 (0.73–0.97)		0.84 (0.73–0.97)	0.84 (0.73–0.96)	0.84 (0.73–0.96)	0.84 (0.73–0.96)
Family history of heart disease			0.91 (0.77–1.06)						
History of stroke/TIA	1.45 (1.25–1.68)	1.32 (1.13–1.55)	1.3 (1.11–1.53)	1.3 (1.11–1.53)	1.33 (1.14–1.56)	1.3 (1.11–1.53)	1.3 (1.11–1.53)	1.29 (1.1–1.51)	1.29 (1.1–1.51)
Angina		1.24 (1.1–1.4)	1.23 (1.09–1.4)	1.23 (1.09–1.39)	1.22 (1.08–1.37)	1.23 (1.08–1.39)	1.23 (1.09–1.39)	1.22 (1.08–1.38)	1.23 (1.09–1.39)
Cancer		1.31 (1–1.72)	1.31 (1–1.73)	1.32 (1–1.73)		1.32 (1–1.74)		1.34 (1.02–1.77)	1.32 (1.01–1.74)
Dementia		1.55 (1.23–1.97)	1.58 (1.24–2.01)	1.58 (1.24–2.01)	1.6 (1.27–2.02)	1.58 (1.25–2.01)	1.58 (1.25–2.01)	1.57 (1.24–1.99)	1.57 (1.23–1.99)
Peptic ulcer disease		0.79 (0.62–1)	0.79 (0.62–1.01)						
Previous AMI		1.21 (1.06–1.38)	1.2 (1.05–1.37)	1.21 (1.06–1.39)	1.2 (1.05–1.36)	1.21 (1.06–1.39)	1.21 (1.06–1.39)	1.2 (1.06–1.37)	1.22 (1.07–1.39)
Asthma									
Depression		1.32 (1.08–1.61)	1.34 (1.09–1.64)	1.31 (1.07–1.61)	1.32 (1.08–1.61)	1.32 (1.08–1.62)	1.32 (1.08–1.62)	1.32 (1.07–1.61)	1.32 (1.08–1.62)
Peripheral arterial disease		1.25 (1.04–1.5)	1.23 (1.02–1.48)	1.23 (1.02–1.48)		1.24 (1.03–1.5)	1.25 (1.03–1.5)	1.25 (1.04–1.51)	1.25 (1.04–1.5)
Previous revascularization									
Congestive heart failure	1.76 (1.44–2.14)	1.47 (1.19–1.81)	1.49 (1.2–1.84)	1.5 (1.22–1.86)	1.47 (1.2–1.81)	1.5 (1.21–1.85)	1.51 (1.22–1.87)	1.46 (1.19–1.81)	1.47 (1.19–1.82)
Hyperthyroidism									
Aortic stenosis	1.92 (1.35–2.73)	1.75 (1.21–2.52)	1.73 (1.2–2.51)	1.73 (1.2–2.51)	1.8 (1.25–2.58)	1.72 (1.19–2.49)	1.74 (1.2–2.52)	1.73 (1.2–2.49)	1.75 (1.21–2.52)
Hemoglobin	0.99 (0.98–0.99)	0.99 (0.99–1)	0.99 (0.99–1)	0.99 (0.99–1)	0.99 (0.99–0.99)	0.99 (0.99–1)	0.99 (0.99–1)	0.99 (0.99–1)	0.99 (0.99–1)
White blood cell count		1.03 (1.02–1.04)	1.03 (1.02–1.04)	1.03 (1.02–1.04)	1.04 (1.03–1.05)	1.03 (1.02–1.04)	1.03 (1.02–1.04)	1.03 (1.02–1.04)	1.03 (1.02–1.04)

(*Continued*)

Table 3. Continued

Variable	Variable Selection Method								
	Complete	Single	S1	S2	S3	W1	W2	W3	RR
Sodium		0.98 (0.97–1)	0.98 (0.97–1)	0.98 (0.97–1)		0.98 (0.97–1)	0.98 (0.97–1)	0.98 (0.97–1)	0.98 (0.97–1)
Potassium		1.15 (1.04–1.27)	1.16 (1.05–1.28)	1.17 (1.06–1.29)	1.16 (1.05–1.28)	1.17 (1.06–1.29)	1.17 (1.06–1.29)	1.15 (1.04–1.27)	1.16 (1.05–1.28)
Glucose	1.06 (1.05–1.07)	1.05 (1.04–1.06)	1.05 (1.03–1.06)	1.05 (1.04–1.06)	1.05 (1.04–1.06)	1.05 (1.04–1.06)	1.05 (1.04–1.06)	1.05 (1.04–1.06)	1.05 (1.04–1.06)
Urea	1.08 (1.07–1.09)	1.04 (1.03–1.06)	1.04 (1.03–1.06)	1.04 (1.03–1.06)	1.04 (1.03–1.05)	1.04 (1.03–1.06)	1.04 (1.03–1.06)	1.04 (1.03–1.05)	1.04 (1.03–1.05)
Creatinine		1 (1.00-1.00)	1 (1.00-1.00)	1 (1.00-1.00)	1 (1.00-1.00)	1 (1.00-1.00)	1 (1.00-1.00)	1 (1.00-1.00)	1 (1.00-1.00)
International normalized ratio		1.13 (1.04–1.23)	1.14 (1.05–1.24)	1.14 (1.05–1.24)	1.14 (1.05–1.24)	1.14 (1.05–1.24)	1.14 (1.05–1.24)	1.14 (1.05–1.24)	1.14 (1.05–1.24)
Total cholesterol			0.94 (0.85–1.04)	0.94 (0.86–1.02)		0.94 (0.86–1.02)	0.94 (0.86–1.02)		
HDL cholesterol			1.32 (0.98–1.78)	1.26 (0.95–1.67)		1.25 (0.94–1.66)	1.25 (0.94–1.65)		
LDL cholesterol			1.04 (1–1.07)	1.04 (1–1.07)		1.04 (1–1.07)	1.04 (1–1.07)	1.03 (1–1.07)	1.03 (1–1.07)
Triglycerides			1.02 (0.93–1.12)						
Elevated cardiac enzymes		0.65 (0.53–0.81)	0.65 (0.52–0.81)	0.65 (0.52–0.81)	0.65 (0.52–0.8)	0.65 (0.52–0.81)	0.65 (0.52–0.81)	0.65 (0.53–0.81)	0.65 (0.52–0.81)
ST-segment elevation MI			1.2 (1.07–1.36)	1.2 (1.06–1.35)	1.23 (1.09–1.39)				1.21 (1.07–1.36)

AMI indicates acute myocardial infarction; BMI, body mass index; HDL, high-density lipoprotein; LDL, low-density lipoprotein; and TIA, transient ischemic attack.

version of the validation sample using the model selected using Rubin's Rules (scale on the right vertical axis). All methods of variable selection resulted in models that displayed good calibration when the predicted probability of the outcome was <0.6; however, calibration deteriorated as the predicted probability exceeded 0.6. Calibration was poorest among those subjects with a high predicted probability of mortality. However, as illustrated by the overlaid density plot, there were relatively few subjects with high predicted probabilities of mortality. Differences in calibration between the different models were negligible.

The performance of the different selected models when applied to the validation sample when single mean imputation was used is reported in the top half of Table 5. The generalized R^2 statistic, the c-statistic, and the scaled Brier score were all lower for the model selected using the complete cases in the derivation sample compared with the models selected using the imputed versions of the derivation sample. In contrast to this, these 3 statistics did not differ meaningfully across the models obtained using different variable selection approaches in the imputed versions of the derivation sample.

Variable Selection Using *P*=0.157 for Variable Retention

When using a 0.157 significance level for variable retention, 3 variables were not selected using any of the variable selection approaches, while 14 variables were selected using all 9 variable selection approaches. The numbers of selected variables for the different variable selection methods were: 16 (complete case selection), 26 (S3), 30 (single sample selection), 32 (W2, W3, and Rubin's Rules), 33 (W1), 34 (S2), and 38 (S1). The estimated odds ratios and associated 95% CIs for the predictor variables in each model are reported in Table 6.

The performance of the selected models was evaluated in the imputed versions of the validation sample. The mean measures of model performance across the 69 imputed copies of the validation sample are reported in the bottom half of Table 4. Differences in model performance between the model selected using complete cases and the other models was attenuated compared with what was observed when a statistical significance level of 0.05 was used for variable selection.

The graphical assessment of calibration is described in Figure 2. Results for calibration were similar to those observed when a significance level of 0.05 was used for variable selection.

The performance of the different selected models when applied to the validation sample when single mean imputation was used is reported in the bottom half of Table 5. In contrast to the results obtained when using a significance level of 0.05

Table 4. Measures of Model Performance in the Validation Sample

Variable Selection Method	Number of Variables Selected	Generalized R^2	C-Statistic	Scaled Brier Score	Calibration Slope
P=0.05 criterion for variable selection					
Complete	10	0.359	0.841	0.272	0.993
Single	27	0.413	0.865	0.326	0.983
S1	36	0.413	0.865	0.330	0.973
S2	31	0.413	0.865	0.330	0.977
S3	21	0.407	0.862	0.332	0.971
W1	29	0.413	0.865	0.331	0.976
W2	28	0.411	0.864	0.331	0.971
W3	27	0.412	0.865	0.334	0.982
RR	28	0.413	0.864	0.326	0.982
P=0.157 criterion for variable selection					
Complete	16	0.397	0.859	0.335	0.986
Single	30	0.413	0.865	0.326	0.983
S1	38	0.413	0.865	0.330	0.972
S2	34	0.413	0.865	0.330	0.975
S3	26	0.409	0.863	0.331	0.976
W1	33	0.413	0.865	0.330	0.975
W2	32	0.412	0.864	0.329	0.973
W3	32	0.412	0.865	0.337	0.973
RR	32	0.412	0.864	0.324	0.975

Each cell contains the mean measure of model performance across the 69 imputed versions of the validation sample.

for variable selection, the difference between the performance of the model obtained using the complete cases in the derivation sample and that of the models obtained using the imputed versions of the derivation sample were attenuated when a significance level of 0.157 was used for variable selection.

DISCUSSION

We compared the predictive accuracy of prognostic models developed using different methods for variable selection with imputed data. Model accuracy was evaluated using data from a different temporal era compared with that in which variable selection and model estima-

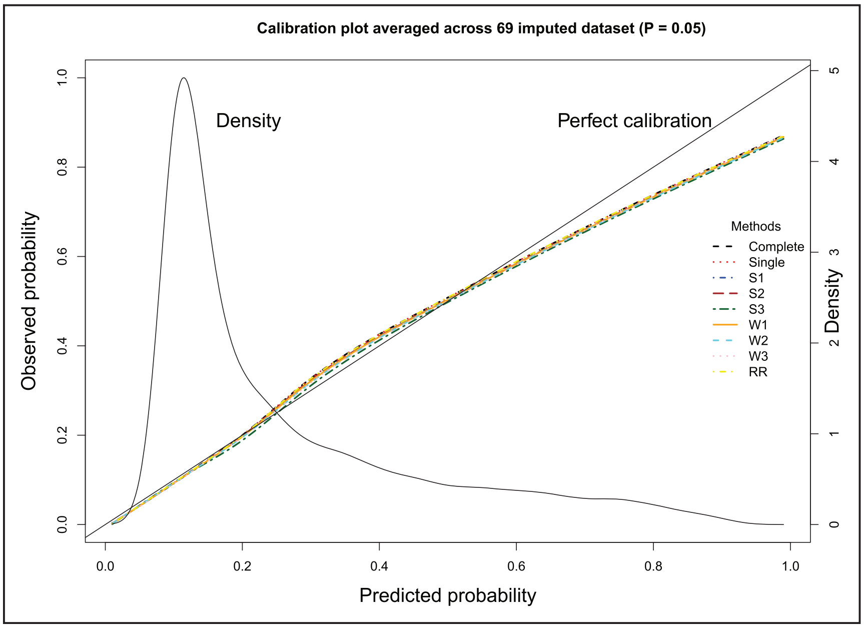

Figure 1. Calibration plot averaged across 69 imputed datasets (*P*=0.05).

Table 5. **Measures of Model Performance in the Validation Sample With Single Mean Imputation**

Variable Selection Method	Generalized R^2	C-Statistic	Scaled Brier Score	Calibration Slope
P=0.05 criterion for variable selection				
Complete	0.356	0.840	0.264	0.997
Single	0.409	0.863	0.314	0.995
S1	0.409	0.863	0.316	0.987
S2	0.409	0.863	0.314	0.992
S3	0.404	0.861	0.325	0.975
W1	0.409	0.863	0.316	0.991
W2	0.407	0.862	0.316	0.987
W3	0.410	0.863	0.324	0.993
RR	0.410	0.863	0.316	0.993
P=0.157 criterion for variable selection				
Complete	0.395	0.857	0.328	0.990
Single	0.410	0.863	0.316	0.993
S1	0.408	0.863	0.316	0.986
S2	0.409	0.863	0.316	0.989
S3	0.407	0.862	0.325	0.981
W1	0.409	0.863	0.316	0.989
W2	0.408	0.862	0.314	0.987
W3	0.408	0.863	0.322	0.988
RR	0.408	0.862	0.312	0.985

Number of variables selected when using the derivation sample is the same as in Table 4.

tion were conducted. The datasets used for model derivation and validation were large and contained a large number of candidate predictor variables. Furthermore, the proportion of subjects with missing data was high. Our primary observation was that the model whose variables were selected using only those subjects with complete data had substantially inferior prognostic ability in the validation sample compared with the models whose variables were selected using the imputed data. The variable selection methods that used all subjects (both those with complete data and those with imputed data) had comparable performance in the validation sample.

There are several reasons why the model selected using the complete cases differed to such a great extent from the models selected using other approaches. First, in the derivation sample, only 13% of subjects had complete data. These subjects may have differed systematically from the entire sample of hospitalized patients. It is plausible that the predictors of mortality differed in this subsample compared with the predictors of mortality in the overall sample. Second, the sample consisting of the complete cases was substantially smaller than the full sample (with a corresponding reduction in the number of observed events). This resulted in a substantially diminished statistical power to identify predictors of the outcome. It was noticeable that the complete case analysis selected substantially fewer predictors than did the other variable selection approaches. The omission of prognostically important variables would result in degraded prediction in the validation sample. These issues highlight the danger of conducting variable selection using only the complete cases.

Variable selection in multiply imputed data has received only modest attention. The most comprehensive study to date is that of Wood et al, who described 9 methods for variable selection.[8] As noted in the Introduction, they evaluated the variable selection methods in terms of their ability to correctly select variables from the true model and to exclude variables not in the true model. They recommended that variable selection be conducted using a stepwise application of Rubin's Rules, as this was the only approach that preserved the type I error rate (the probability that a method will incorrectly select a variable that is not part of the true model). We would note that type I error is of less concern when developing a prognostic model. For this reason, we examined the use of a significance level of 0.157 in addition to examining the performance of a significance level of 0.05 for variable retention.

Clark and Altman compared the performance of models for predicting mortality in patients with ovarian cancer.[16] They found that the model derived using complete cases had worse performance when applied to these subjects than did a model derived using multiple imputation and the full sample when applied to the full sample. We found when using the full validation sample that the models developed using imputed data had superior performance compared with the model developed using the complete cases in the derivation sample. Vergouwe et al compared different methods for developing and validating prognostic models, using prediction of deep venous thrombosis as a test case.[17] Using the terminology of the current paper, they compared the use of Rubin's Rules for variable selection with that of S2 and W3. They found that the 3 approaches resulted in similar results for variable selection.

In the current study, we focused on ideal model performance as opposed to pragmatic model performance.[10] The former refers to the performance of the model in future clinical settings in which all variables are measured and there are no missing data. The latter refers to the performance of the model in future clinical settings in which some subjects have missing data on some variables. Had we been interested in pragmatic model performance, the imputation models in the validation sample would have omitted the outcome variable. Furthermore, we could have developed a set of partial prediction models.[10] In doing so, we would develop a prediction model for the distinct missing data patterns. Given that our sample consisted of 41 predictor variables, with 643 distinct missing data patterns, such an approach was not feasible in our data.

There are certain limitations to the current study. First, our analyses were based upon empirical analyses in a single dataset. It is possible that different findings would

Table 6. Estimated Odds Ratios and 95% CIs in the Derivation Sample When Using a 0.157 Significance Level for Variable Retention

	Variable Selection Method								
Variable	**Complete**	**Single**	**S1**	**S2**	**S3**	**W1**	**W2**	**W3**	**RR**
Age (per 10 y increase)	2.1 (1.98–2.23)	1.93 (1.8–2.06)	1.92 (1.79–2.06)	1.92 (1.79–2.05)	1.97 (1.86–2.09)	1.92 (1.79–2.05)	1.93 (1.81–2.07)	1.92 (1.8–2.06)	1.95 (1.82–2.08)
Female	0.86 (0.76–0.96)		0.9 (0.78–1.04)	0.91 (0.79–1.04)		0.91 (0.79–1.04)	0.9 (0.79–1.04)	0.9 (0.79–1.04)	0.88 (0.77–1.01)
Cardiogenic shock	5.28 (3.7–7.54)	5.05 (3.51–7.27)	5.05 (3.49–7.33)	5.05 (3.49–7.3)	5.02 (3.5–7.19)	5.05 (3.49–7.3)	5.1 (3.53–7.37)	5.18 (3.58–7.49)	5.22 (3.62–7.54)
Acute pulmonary edema									
Systolic blood pressure	0.99 (0.98–0.99)	0.99 (0.98–0.99)	0.99 (0.98–0.99)	0.99 (0.98–0.99)	0.99 (0.98–0.99)	0.99 (0.98–0.99)	0.99 (0.98–0.99)	0.99 (0.98–0.99)	0.99 (0.98–0.99)
Diastolic blood pressure			1 (0.99–1)						
Heart rate	1.01 (1.01–1.01)	1.01 (1.01–1.01)	1.01 (1.01–1.01)	1.01 (1.01–1.01)	1.01 (1.01–1.01)	1.01 (1.01–1.01)	1.01 (1.01–1.01)	1.01 (1–1.01)	1.01 (1.01–1.01)
Respiratory rate	1.04 (1.03–1.04)	1.03 (1.02–1.04)	1.03 (1.02–1.04)	1.03 (1.02–1.04)	1.03 (1.02–1.04)	1.03 (1.02–1.04)	1.03 (1.02–1.04)	1.03 (1.02–1.04)	1.03 (1.02–1.04)
BMI		0.98 (0.96–1)	0.98 (0.96–1)	0.98 (0.97–1)		0.98 (0.97–1)	0.98 (0.97–1)	0.98 (0.97–1)	0.98 (0.97–1)
Diabetes mellitus			1.07 (0.92–1.25)						
Hypertension	1.12 (1–1.25)	1.14 (1.01–1.29)	1.15 (1.02–1.3)	1.15 (1.02–1.3)	1.1 (0.98–1.24)	1.15 (1.02–1.3)	1.15 (1.02–1.3)	1.15 (1.02–1.3)	1.15 (1.02–1.3)
Current smoker	1.18 (1.02–1.36)	1.13 (0.97–1.31)	1.15 (0.99–1.33)	1.15 (0.99–1.33)		1.15 (0.99–1.33)	1.14 (0.99–1.33)	1.15 (0.99–1.34)	1.14 (0.98–1.33)
Dyslipidemia		0.85 (0.74–0.98)	0.85 (0.73–0.98)	0.85 (0.74–0.98)	0.83 (0.73–0.96)	0.85 (0.74–0.98)	0.84 (0.73–0.97)	0.85 (0.74–0.97)	0.84 (0.73–0.97)
Family history of heart disease		0.9 (0.77–1.06)	0.91 (0.77–1.06)	0.9 (0.77–1.06)		0.9 (0.77–1.06)			
History of stroke/TIA	1.41 (1.21–1.65)	1.3 (1.11–1.52)	1.3 (1.11–1.53)	1.31 (1.11–1.53)	1.31 (1.12–1.53)	1.31 (1.11–1.53)	1.31 (1.11–1.54)	1.31 (1.11–1.54)	1.3 (1.11–1.53)
Angina		1.24 (1.1–1.4)	1.24 (1.09–1.4)	1.24 (1.09–1.4)	1.22 (1.08–1.38)	1.24 (1.09–1.4)	1.24 (1.09–1.4)	1.23 (1.09–1.39)	1.24 (1.1–1.4)
Cancer		1.32 (1.01–1.74)	1.32 (1–1.74)	1.31 (0.99–1.73)		1.31 (0.99–1.73)	1.31 (0.99–1.73)	1.33 (1.01–1.75)	1.32 (1–1.73)
Dementia		1.55 (1.22–1.97)	1.57 (1.24–2)	1.58 (1.24–2)	1.61 (1.27–2.03)	1.58 (1.24–2)	1.58 (1.25–2.01)	1.59 (1.25–2.02)	1.56 (1.23–1.98)
Peptic ulcer disease		0.79 (0.62–1.01)	0.79 (0.62–1.01)	0.79 (0.62–1.01)	0.78 (0.62–1)	0.79 (0.62–1.01)	0.79 (0.61–1)	0.79 (0.62–1.01)	0.78 (0.61–1)
Previous AMI		1.21 (1.06–1.39)	1.2 (1.05–1.37)	1.2 (1.05–1.37)	1.21 (1.06–1.37)	1.2 (1.05–1.37)	1.21 (1.06–1.38)	1.19 (1.04–1.36)	1.21 (1.06–1.39)
Asthma			0.85 (0.67–1.09)						
Depression	1.37 (1.12–1.67)	1.32 (1.08–1.62)	1.34 (1.09–1.64)	1.34 (1.09–1.65)	1.31 (1.07–1.6)	1.34 (1.09–1.65)	1.34 (1.09–1.64)	1.33 (1.09–1.63)	1.34 (1.1–1.65)
Peripheral arterial disease		1.25 (1.04–1.5)	1.22 (1.01–1.47)	1.23 (1.02–1.48)	1.28 (1.07–1.54)	1.23 (1.02–1.48)	1.23 (1.02–1.49)	1.24 (1.03–1.49)	1.24 (1.03–1.5)
Previous revascularization									
Congestive heart failure	1.63 (1.33–2)	1.47 (1.19–1.82)	1.49 (1.21–1.85)	1.5 (1.21–1.86)	1.47 (1.19–1.8)	1.5 (1.21–1.86)	1.51 (1.22–1.87)	1.5 (1.21–1.86)	1.5 (1.21–1.86)
Hyperthyroidism									
Aortic stenosis	1.83 (1.27–2.62)	1.75 (1.21–2.52)	1.75 (1.2–2.53)	1.72 (1.19–2.49)	1.77 (1.23–2.55)	1.72 (1.19–2.49)	1.74 (1.2–2.52)	1.72 (1.19–2.49)	1.77 (1.22–2.55)
Hemoglobin	0.99 (0.99–0.99)	0.99 (0.99–1)	0.99 (0.99–1)	0.99 (0.99–1)	0.99 (0.99–1)	0.99 (0.99–1)	0.99 (0.99–1)	0.99 (0.99–1)	0.99 (0.99–1)
White blood count	1.04 (1.02–1.05)	1.03 (1.02–1.04)	1.03 (1.02–1.04)	1.03 (1.02–1.04)	1.03 (1.02–1.04)	1.03 (1.02–1.04)	1.03 (1.02–1.04)	1.03 (1.02–1.04)	1.03 (1.02–1.04)

(*Continued*)

Table 6. Continued

Variable	Variable Selection Method								
	Complete	Single	S1	S2	S3	W1	W2	W3	RR
Sodium		0.98 (0.97–1)	0.99 (0.97–1)	0.98 (0.97–1)	0.98 (0.97–1)	0.98 (0.97–1)	0.98 (0.97–1)	0.98 (0.97–1)	0.98 (0.97–1)
Potassium		1.16 (1.05–1.27)	1.16 (1.05–1.28)	1.17 (1.06–1.29)	1.15 (1.05–1.27)	1.17 (1.06–1.29)	1.17 (1.06–1.29)	1.16 (1.05–1.28)	1.16 (1.05–1.28)
Glucose	1.05 (1.04–1.06)	1.05 (1.04–1.06)	1.05 (1.03–1.06)	1.05 (1.04–1.06)	1.04 (1.03–1.06)	1.05 (1.04–1.06)	1.05 (1.04–1.06)	1.05 (1.04–1.06)	1.05 (1.04–1.06)
Urea	1.07 (1.05–1.08)	1.04 (1.03–1.06)	1.04 (1.03–1.06)	1.04 (1.03–1.06)	1.04 (1.03–1.05)	1.04 (1.03–1.06)	1.04 (1.03–1.06)	1.04 (1.03–1.06)	1.04 (1.03–1.06)
Creatinine		1 (1–1)	1 (1–1)	1 (1–1)	1 (1–1)	1 (1–1)	1 (1–1)	1 (1–1)	1 (1–1)
International normalized ratio		1.14 (1.05–1.24)	1.14 (1.05–1.24)	1.14 (1.05–1.24)	1.13 (1.04–1.23)	1.14 (1.05–1.24)	1.14 (1.05–1.24)	1.14 (1.05–1.24)	1.14 (1.05–1.24)
Total cholesterol			0.94 (0.85–1.04)	0.94 (0.86–1.03)		0.94 (0.86–1.03)	0.94 (0.86–1.03)	0.94 (0.86–1.03)	
HDL cholesterol			1.32 (0.98–1.78)	1.28 (0.96–1.72)		1.28 (0.96–1.72)	1.28 (0.96–1.72)	1.29 (0.97–1.73)	1.24 (0.93–1.65)
LDL cholesterol		1.03 (1–1.07)	1.04 (1–1.07)	1.04 (1–1.07)		1.04 (1–1.07)	1.04 (1–1.07)	1.04 (1–1.07)	1.03 (1–1.07)
Triglycerides			1.02 (0.93–1.12)						
Elevated cardiac enzymes		0.65 (0.52–0.81)	0.65 (0.52–0.81)	0.65 (0.52–0.8)	0.65 (0.53–0.81)	0.65 (0.52–0.8)	0.65 (0.52–0.8)	0.65 (0.52–0.8)	0.64 (0.52–0.8)
ST-segment elevation MI			1.2 (1.07–1.36)	1.2 (1.07–1.36)	1.23 (1.09–1.39)				1.21 (1.07–1.36)

AMI indicates acute myocardial infarction; BMI, body mass index; HDL, high-density lipoprotein; LDL, low-density lipoprotein; and TIA, transient ischemic attack.

be observed in samples of subjects with different clinical conditions or for different outcomes. However, the datasets used for derivation and validation were large and from temporally distinct periods. This allowed us to assess the temporal transportability of the derived models.[7] A second limitation, pertaining to the generalizability of our findings, is that the derivation sample was large. We found minimal differences between the variable selection approaches (with the exception of the complete case approach) in terms of their prognostic ability. It is possible that differences between the variable selection approaches would be amplified in small samples. A third limitation pertains to the use of data from an earlier era (1999–2001 and 2004–2005). The distribution of patient

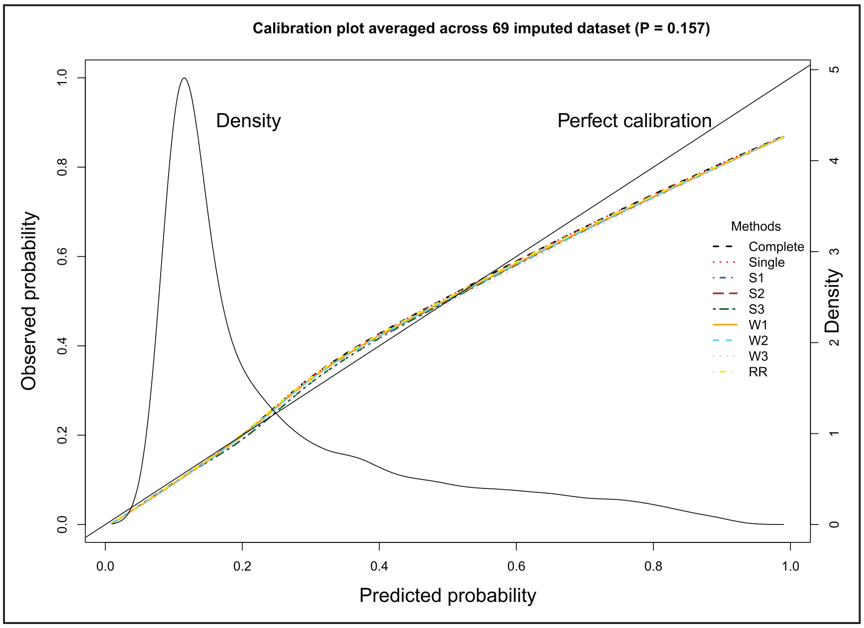

Figure 2. Calibration plot averaged across 69 imputed datasets (*P*=0.157).

characteristics and patterns of care for patients with acute myocardial infarction may differ between this era and the current era. It is conceivable that the selected predictor variables and the rate of missing data would differ in a more recent era. However, the use of these data was to illustrate statistical issues in variable selection when using multiple imputation. The objective of the current study was not to derive clinical prediction models for use in current clinical practice.

We illustrated that, apart from variable selection using the complete cases, the competing variable selection methods produced prognostic models that had comparable performance when evaluated in a temporally distinct validation sample. Despite the similar performance of the different models, we would argue for the use of the method based on the application of Rubin's Rules for variable selection. Such a selection process results in a final model that has the desirable property that the selected covariates all meet a predefined significance level. This is in contrast to several of the other variable selection methods that resulted in the inclusion of nonsignificant covariates once the final model was estimated using Rubin's Rules. A limitation to the use of Rubin's Rules (and to the S1, S2, S3, and W3 methods) is that they require user-written software, and they are typically not available in standard statistical software (though the mim stepwise procedure in the mim package for Stata can be used for stepwise variable selection in multiply imputed data). In contrast to this, the single method and the W1 and W2 methods can be easily implemented using conventional statistical software packages.

ARTICLE INFORMATION

Correspondence

Peter C. Austin, PhD, ICES, G106, 2075 Bayview Ave, Toronto, ON M4N 3M5, Canada. Email peter.austin@ices.on.ca

Affiliations

ICES, Toronto, ON, Canada (P.C.A., D.S.L., D.T.K.). Institute of Health Management, Policy and Evaluation, University of Toronto, ON, Canada (P.C.A., D.S.L., D.T.K.). Sunnybrook Research Institute, Toronto, ON, Canada (P.C.A., D.T.K.). Department of Medicine, University of Toronto, ON, Canada (D.S.L., D.T.K.). Peter Munk Cardiac Centre, University Health Network, Toronto, ON, Canada (D.S.L.). Medical Research Council Clinical Trials Unit, University College London, United Kingdom (I.R.W.).

Sources of Funding

This study was supported by ICES, which is funded by an annual grant from the Ontario Ministry of Health and Long-Term Care (MOHLTC). The opinions, results, and conclusions reported in this paper are those of the authors and are independent from the funding sources. No endorsement by ICES or the Ontario MOHLTC is intended or should be inferred. This research was supported by an operating grant from the Canadian Institutes of Health Research (CIHR; MOP 86508). Drs Austin, Ko, and Lee are supported in part by Mid-Career Investigator awards from the Heart and Stroke Foundation. Dr Lee is the Ted Rogers Chair in Heart Function Outcomes. Dr White was supported by the Medical Research Council Unit Programme MC_UU_12023/21. The Enhanced Feedback for Effective Cardiac Treatment data used in the study was funded by a CIHR Team Grant in Cardiovascular Outcomes Research (grant numbers CTP 79847 and CRT43823).

Disclosures

None.

REFERENCES

1. Steyerberg EW. *Clinical Prediction Models*. New York: Springer-Verlag; 2009.
2. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation*. 1998;97:1837–1847. doi: 10.1161/01.cir.97.18.1837
3. Eagle KA, Lim MJ, Dabbous OH, Pieper KS, Goldberg RJ, Van de Werf F, Goodman SG, Granger CB, Steg PG, Gore JM, Budaj A, Avezum A, Flather MD, Fox KA; GRACE Investigators. A validated prediction model for all forms of acute coronary syndrome: estimating the risk of 6-month postdischarge death in an international registry. *JAMA*. 2004;291:2727–2733. doi: 10.1001/jama.291.22.2727
4. Lee DS, Austin PC, Rouleau JL, Liu PP, Naimark D, Tu JV. Predicting mortality among patients hospitalized for heart failure: derivation and validation of a clinical model. *JAMA*. 2003;290:2581–2587. doi: 10.1001/jama.290.19.2581
5. Myers RH. *Classical and Modern Regression with Applications*. Second Ed. Belmont, California: Duxbury Press; 1990.
6. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons; 1987.
7. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med*. 1999;130:515–524. doi: 10.7326/0003-4819-130-6-199903160-00016
8. Wood AM, White IR, Royston P. How should variable selection be performed with multiply imputed data? *Stat Med*. 2008;27:3227–3246. doi: 10.1002/sim.3177
9. Tu JV, Donovan LR, Lee DS, Wang JT, Austin PC, Alter DA, Ko DT. Effectiveness of public report cards for improving the quality of cardiac care: the EFFECT study: a randomized trial. *JAMA*. 2009;302:2330–2337. doi: 10.1001/jama.2009.1731
10. Wood AM, Royston P, White IR. The estimation and use of predictions for the assessment of model performance using large samples with multiply imputed data. *Biom J*. 2015;57:614–632. doi: 10.1002/bimj.201400004
11. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med*. 2011;30:377–399. doi: 10.1002/sim.4067
12. Nagelkerke NJD. A note on a general definition of the coefficient of determination. *Biometrika*. 1991;78:691c692.
13. Cox DR. Two further applications of a model for binary regression. *Biometrika*. 1958;45:592–665.
14. Marshall A, Altman DG, Holder RL, Royston P. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. *BMC Med Res Methodol*. 2009;9:57. doi: 10.1186/1471-2288-9-57
15. Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. *Stat Med*. 2014;33:517–535. doi: 10.1002/sim.5941
16. 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. doi: 10.1016/s0895-4356(02)00539-5
17. Vergouwe Y, Royston P, Moons KG, Altman DG. Development and validation of a prediction model with missing predictor data: a practical approach. *J Clin Epidemiol*. 2010;63:205–214. doi: 10.1016/j.jclinepi.2009.03.017