Predicting prolonged dose titration in patients starting warfarin

Supplementary Material

Supplementary Table 1 2

Supplementary Table 2 3

Supplementary Table 3 4

Supplementary Table 4 5

Supplementary Figure 1 6

Supplementary Figure 2 7

Supplementary Figure 3 8

Supplementary Figure 4 9

Supplementary Methods 10

Supplementary References 13

Supplementary Table 1. Candidate baseline sociodemographic and clinical predictors

Candidate Predictor	Specification
Sociodemographic	
Gender	Binary (0 = male; 1 = female)
Self-reported race	Binary (0 = not African American; 1 = African American)
Marital status	Categorical (1 = married (ref); 2 = separated/divorced; 3 = widowed; 4 = never married)
Employment status	Categorical (1 = working; 2 = unemployed/disabled; 3 = retired (ref))
Education status	Binary (0 = more than high school; 1 = high school or less)
Insurance status	Categorical (1 = private (ref); 2 = any VA/Medicare only; 3 = Medicaid/no insurance)
Number of alcoholic drinks per week	Binary (0 = 0 drinks; 1 = 1 or more drinks)
Current smoking status	Binary (0 = not current smoker; 1 = current smoker)
Self-reported general health status	Categorical (1 = excellent/very good (ref); 2 = good; 3 = fair/poor)
No. hospitalizations in past 12 months	Categorical (1 = 0 visits (ref); 2 = 1–2 visits; 3 = 3 or more visits)
No. doctor's visits in past 12 months	Categorical (1 = 0–3 visits; 2 = 4–12 visits (ref); 3 = 13 or more visits)
Had difficulty receiving healthcare in the past 12 months	Binary (0 = no; 1 = yes)
Clinical	
Age (years) at baseline visit	Continuous (linear)
Body Mass Index	Continuous (linear)
Warfarin indication	Categorical (1 = atrial fibrillation/atrial flutter (ref); 2 = deep vein thrombosis/pulmonary embolism; 3 = other)
Previous use of warfarin	Binary (0 = no; 1 = yes)
Number of interacting medications being used at baseline	Binary (0 = 0–1 medications; 1 = 2 or more medications)
Amiodarone use at baseline	Binary (0 = no; 1 = yes)
Statin use at baseline	Binary (0 = no; 1 = yes)
$CHADS_2$ score	Categorical (1 = 0 (ref); 2 = 1; 3 = 2 or higher)
History of peptic ulcer disease or gastritis	Binary (0 = no; 1 = yes)
History of stroke	Binary (0 = no; 1 = yes)
History of cancer	Binary (0 = no; 1 = yes)
History of hypertension	Binary (0 = no; 1 = yes)
History of diabetes	Binary (0 = no; 1 = yes)
History of arrhythmia	Binary (0 = no; 1 = yes)
History of congestive heart failure	Binary (0 = no; 1 = yes)
History of myocardial infarction	Binary (0 = no; 1 = yes)

Supplementary Table 2. Comparison of variables selected for best subsets with different numbers of variables.

Number of variables in the model	Variables included in the best subset
1	Age
2	Age, Insurance status
3	Insurance status, Warfarin indication, # MD visits
4	Insurance status, Warfarin indication, # MD visits, Heart failure
5	Insurance status, Warfarin indication, # MD visits, Heart failure, Smoking status
6	Insurance status, Warfarin indication, # MD visits, Heart failure, Smoking status, Gender
7	Insurance status, Warfarin indication, # MD visits, Heart failure, Smoking status, General health status, Peptic ulcer disease
8	Insurance status, Warfarin indication, # MD visits, Heart failure, Smoking status, General health status, Peptic ulcer disease, Cancer
9	Insurance status, Warfarin indication, # MD visits, Heart failure, Smoking status, Gender, General health status, Peptic ulcer disease, Arrhythmia
10	Insurance status, Warfarin indication, # MD visits, Heart failure, Smoking status, Gender, General health status, Peptic ulcer disease, Cancer, Arrhythmia

Prediction models were compared by the time-dependent AUC at 12 weeks, as estimated by leave-one-out cross-validation (LOOCV).

Supplementary Table 3. Characteristics of the derivation and validation cohorts by site

Variable	HUP		CMCVAMC	
	Derivation (N = 184)	Validation (N = 263)	Derivation (N = 137)	Validation (N = 198)
Age				
< 45	51 (28)	75 (29)	12 (9)	7 (4)
45 – 55	39 (21)	61 (23)	23 (17)	28 (14)
55 – 65	35 (19)	62 (24)	57 (42)	96 (48)
65 – 75	35 (19)	41 (16)	28 (20)	45 (23)
75+	24 (13)	23 (9)	17 (12)	22 (11)
Female gender	89 (48)	136 (52)	5 (4)	9 (5)
African American race	103 (56)	188 (71)	70 (51)	147 (74)
Employment status:				
Working	79 (43)	78 (30)	32 (24)	18 (9)
Unemployed	17 (9)	22 (8)	17 (13)	12 (6)
Retired	49 (27)	62 (24)	49 (36)	91 (46)
Disabled	39 (21)	99 (38)	37 (27)	77 (39)
Annual income:				
< $15,000	48 (29)	90 (37)	41 (38)	95 (49)
$15,000 - $20,000	45 (27)	18 (7)	48 (45)	18 (9)
> $20,000	72 (44)	137 (56)	18 (17)	79 (41)
Insurance status				
Private	151 (84)	162 (63)	6 (4)	10 (5)
VA/Medicare/Other	7 (4)	43 (17)	107 (79)	154 (78)
Medicaid/None	21 (12)	54 (21)	23 (17)	34 (17)
Number MD visits in previous year				
< 4	48 (27)	76 (29)	35 (26)	35 (18)
4 – 12	87 (48)	116 (44)	56 (41)	90 (46)
> 12	46 (25)	71 (27)	46 (34)	72 (37)
Smoking status				
Never	90 (49)	130 (50)	19 (14)	37 (19)
Past	73 (40)	88 (34)	79 (58)	92 (46)
Current	21 (11)	42 (16)	39 (28)	69 (35)
Body Mass Index				
< 25	63 (34)	65 (25)	41 (30)	62 (31)
25 – 30	62 (34)	77 (30)	36 (26)	57 (29)
> 30	59 (32)	117 (45)	59 (43)	79 (40)
Warfarin indication				
AFib/AFlutter	68 (37)	81 (31)	70 (51)	87 (44)
DVT/PE	73 (40)	131 (50)	40 (29)	86 (44)
Other	43 (23)	51 (19)	27 (20)	24 (12)
Previously used warfarin	47 (26)	75 (29)	40 (29)	72 (36)
History of hypertension	87 (47)	170 (65)	66 (48)	151 (76)
History of diabetes	40 (22)	68 (26)	50 (36)	71 (36)
History of peptic ulcer disease	16 (9)	26 (10)	17 (12)	9 (5)
History of heart failure	31 (17)	58 (22)	34 (25)	40 (20)

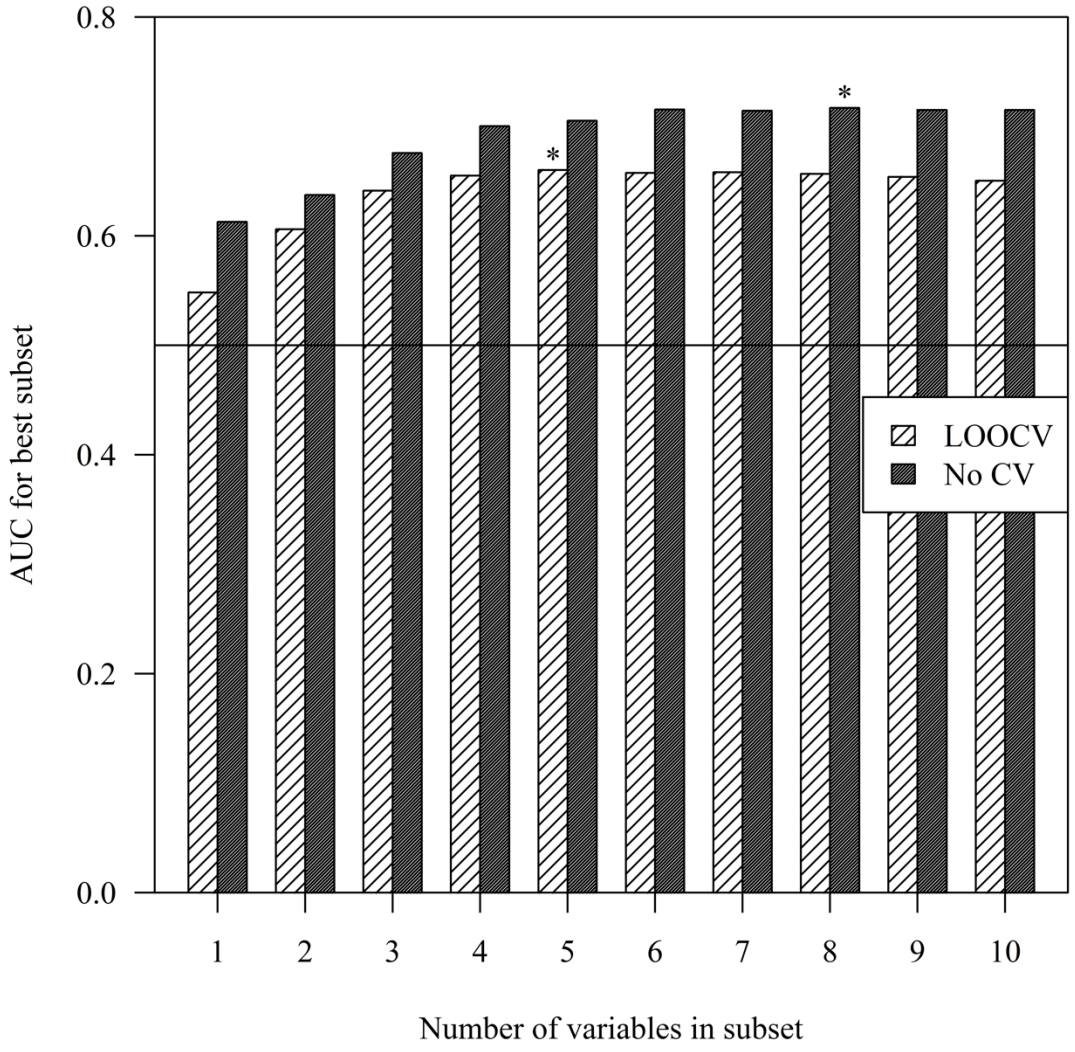
All values are reported as N (%), and sites were limited to those that were present in both derivation and validation cohorts.

Supplementary Table 4. Prediction model coefficients when developed in validation cohort in *post-hoc* analysis

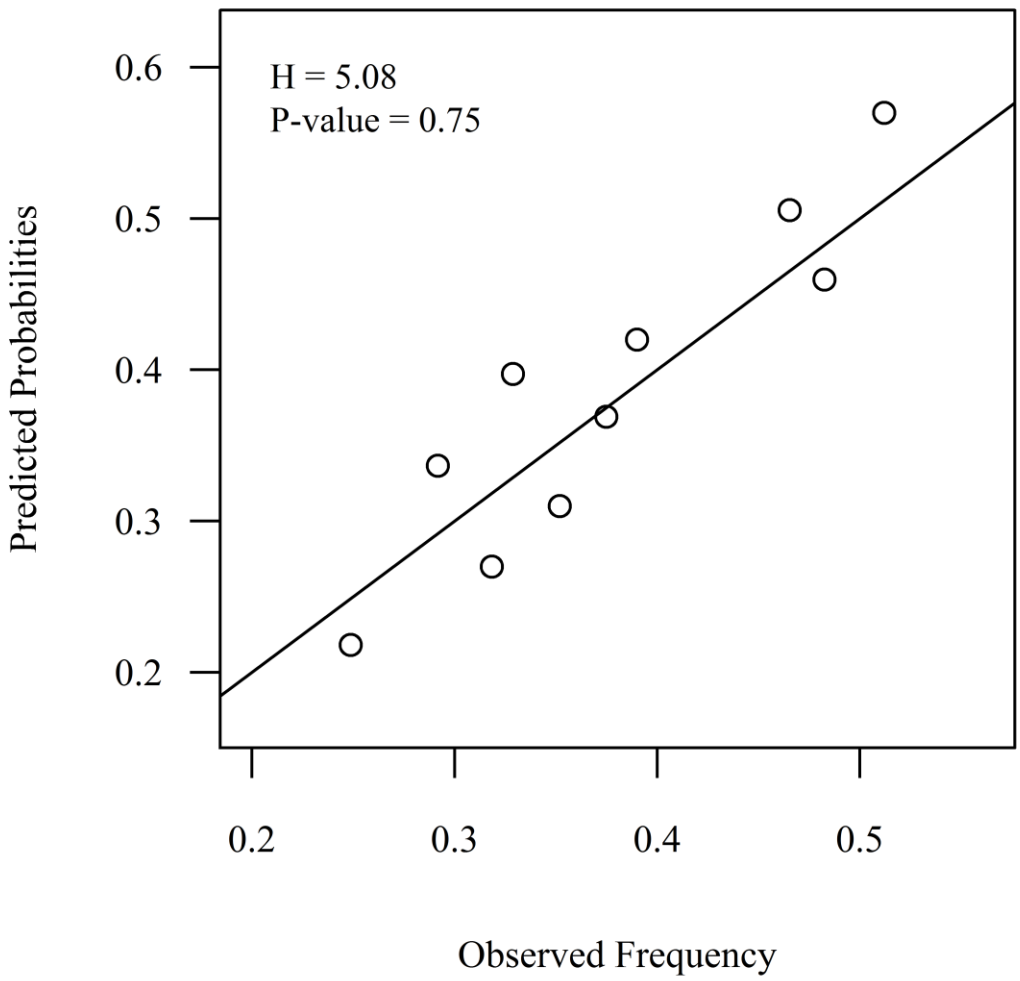
Predictor variable	Shrunk coefficient*
Age	0.01
Body Mass Index	0.02
Warfarin indication	
AFib/Aflutter	Ref
DVT/PE	-0.34
Other	-0.33
Insurance status	
Private insurance	Ref
VA/Medicare	-0.14
Medicaid/None	-0.36
Previous warfarin use	-0.22
History of heart failure	-0.12
History of arrhythmia	-0.16

*To improve expected model calibration, coefficients were shrunk using a linear shrinkage factor, equal to 0.84, which was estimated from 1,000 bootstrap replications. Negative coefficients indicate a higher probability of prolonged dose titration.

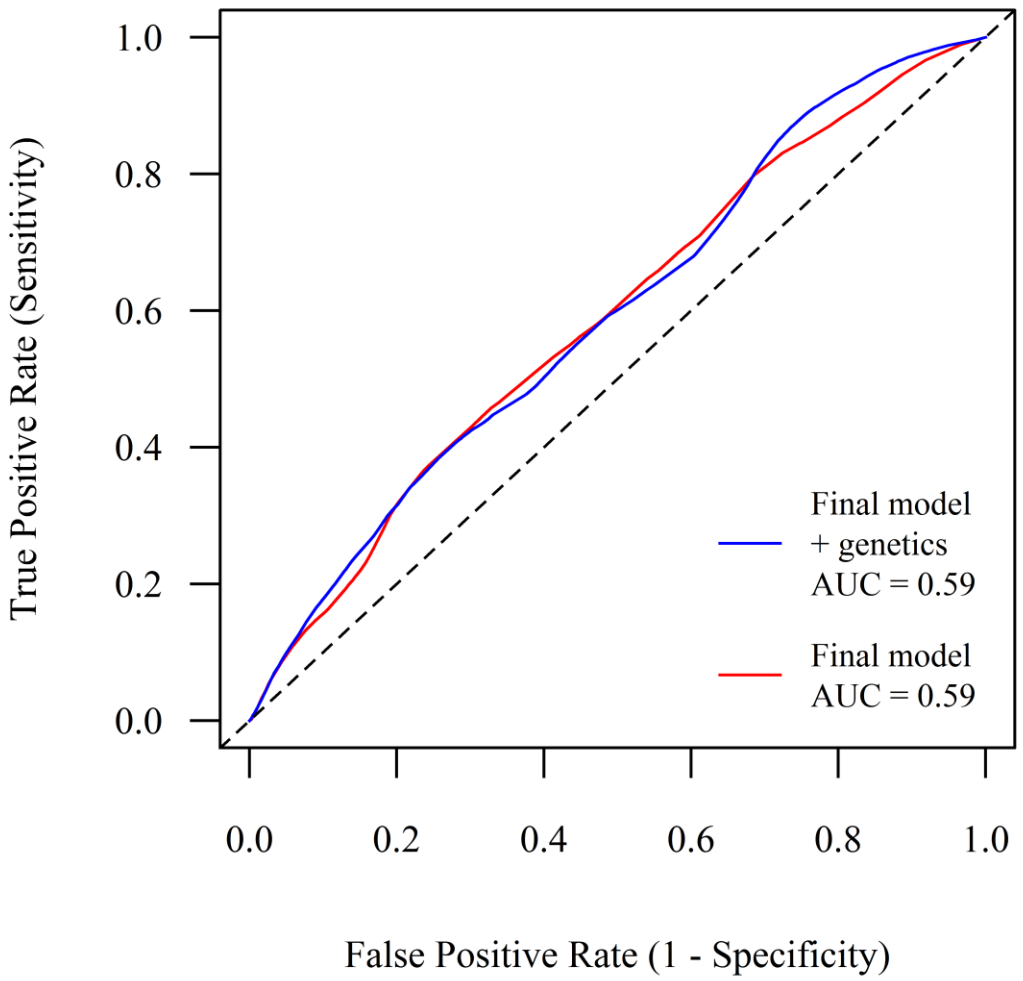
Supplementary Figure 1. Comparison of best prediction models by number of predictor variables in the model, as estimated with and without cross-validation. Prediction models were compared by the time-dependent AUC at 12 weeks, as estimated using leave-one-out cross-validation (LOOCV) or no cross-validation (CV). Asterisks indicate the model that would have been selected as the overall best model for each estimation method.



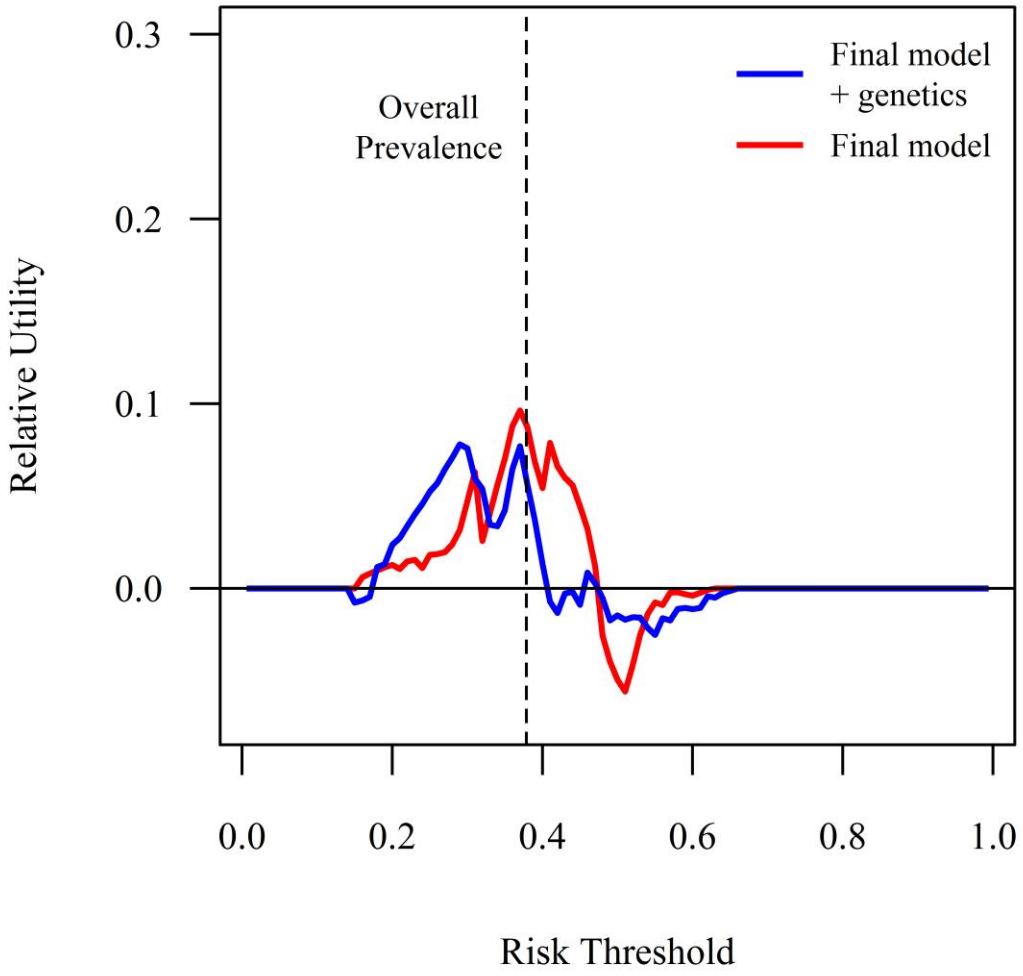
Supplementary Figure 2. Predicted probability vs. observed frequency of prolonged dose titration by risk decile.



Supplementary Figure 3. Comparison of ROC curves for the prediction models with and without the addition of genetic factors.



Supplementary Figure 4. Comparison of relative utility curves in prediction models with and without genetic factors.



Supplementary Methods

Univariable screen

To reduce overall computing time for our analyses to manageable levels, we chose to perform a univariable screen to reduce the number of candidate predictors from the initial 28 to 20, which was determined *a priori* to be an appropriate number of candidate variables, given computational constraints. For each candidate predictor, we constructed a univariable Cox regression model of the time from initiation of warfarin to the achievement of maintenance dose or censoring. We then estimated the time-dependent area under the ROC curve (AUC) at 12 weeks of follow-up using 10-fold cross-validation for each model. The 20 variables with the best time-dependent AUCs were selected for inclusion in the modified best subsets algorithm, as described below.

Time-dependent AUC

The time-dependent AUC—developed by Heagerty, et al.[1]—differs from the standard AUC because it accommodates censoring, and it differs from the commonly used C-index because it assesses model discrimination at a single point in time, rather than over the total duration of follow-up. The time-dependent AUC can thus be interpreted as the probability that a randomly selected individual who has experienced the failure event by time *t* will have a higher predicted probability of failure at time *t* than a randomly selected individual who has not experienced the failure event by time *t*. This statistic is estimated by integrating the time-dependent sensitivity and specificity across all possible cut-off values for the linear predictor derived from the model.[2] Because cross-validation was being used during the model development process, the linear predictor was calculated in the data subset that was withheld during estimation of the Cox model, repeated for all data subsets (e.g. 10 times for 10-fold cross-validation). When the model was assessed in the external validation cohort, the linear predictors in that cohort were used without cross-validation.

Because individuals may be censored prior to time *t*, the values for time-dependent sensitivity and specificity need to be estimated from the data. As recommended by Heagerty, et al., we used a nearest neighbor estimator—which is essentially a weighted Kaplan-Meier estimator based on a nearest neighbor kernel function, developed by Akritas[3]—which allows for monotonicity of sensitivity and specificity and for the censoring process to depend on the predictive marker of interest. This estimator is dependent on a smoothing parameter, λ, where 2λ represents the percentage of observations that are included in an individual observation's neighborhood; in our case, we chose the default value of $\lambda = 0.025$. The "survivalROC" package in R was used to facilitate these calculations.[4]

Modified best subsets selection algorithm

Variable selection was conducted using a modified best subsets algorithm.[5] This algorithm was designed to optimize model discrimination, or how well a model distinguishes between those who did and did not experience the outcome (in this case, those who had a prolonged vs non-

prolonged dose-titration phase, respectively). We calculated the time-dependent AUC at 12 weeks using 10-fold cross-validation for all possible combinations of the 20 remaining candidate predictors up to 10 predictor variables in length (616,665 combinations) to reduce our chances of selecting a combination based on overfitting. Because leave-one-out cross-validation (LOOCV)—in which one person at a time is removed from the dataset to build the model and then used for model testing, for all individuals in the dataset—can be a better estimate of external validation than 10-fold cross-validation,[6] we opted to estimate the time-dependent AUC using LOOCV in the 1,000 best models based on 10-fold cross-validation for each subset size (8,210 combinations). The combination of predictors that led to the highest time-dependent AUC using LOOCV was then selected as our final prediction model. This strategy had the advantage of choosing the best subset based on LOOCV, without the nearly 40-fold increase in computing time that would be required by calculating the time-dependent AUC using LOOCV in all possible combinations of predictors. A sensitivity analysis showed that this algorithm selected the exact same best combination of predictor variables as using LOOCV on all possible combinations up to 6 predictor variables in length.

Linear shrinkage factor

Because regression coefficients are often overestimated in small samples, prediction models will often show better calibration for out-of-sample predictions when coefficients are shrunk toward zero.[7] Thus, we sought to apply a linear shrinkage factor—which has been shown to perform well in small samples for improving model calibration, without sacrificing model discrimination[8]—to our final prediction model. To estimate the shrinkage factor, we fit the model in a bootstrap sample of the derivation cohort. We then calculated the linear predictors of the individuals in the actual derivation cohort using the model coefficients from the model fit in the bootstrap sample. The slope of the actual observed outcomes regressed on these bootstrapped linear predictors could then be used as an estimate of the shrinkage factor. To form a stable estimate of the shrinkage factor, we calculated the mean slope over 1,000 bootstrap replications. All of the original model coefficients were then multiplied by this shrinkage factor to produce the final shrunk coefficients, which were used for generating predictions in the external validation cohort. Because all of the coefficients are being multiplied by the same factor, the rank order of individual predictions is preserved and model discrimination is not affected by shrinkage.

In order to ensure that shrinkage was toward the overall mean and not toward the overall reference category, continuous variables needed to be centered at the mean and categorical variables had to be coded using simple contrasts. In this contrast method, reference groups were coded as $-1/k$, while non-reference categories were coded as $(k-1)/k$, where k is the number of categories. In this contrast method, the reference category of 0 is equivalent to the overall mean of the sample in which the model is being fit. Note that the difference between the reference and non-reference categories is still 1; thus, the interpretation of coefficients in this

contrast method is identical to the more common dummy coding for categorical variables (i.e. 0 for reference and 1 for non-reference categories).

Measures of clinical utility

The methods for determining clinical utility rely on the concept of the risk threshold, which is the probability of the outcome at which the clinician is indifferent about which treatment strategy to use; in other words, it is the probability at which the costs of false positive and false negative mistakes are equal.[9] Furthermore, the consequences of basing a clinical decision on the predicted probability from a risk prediction model can be estimated as a function of the risk threshold. While the exact threshold used in practice will vary depending on the value that physicians and patients place on certain outcomes, the metric can be used to determine the clinical usefulness of a given model under a range of possible thresholds. For our prediction model, given broadly similar safety and efficacy profiles for warfarin and the alternative anticoagulants (with the possible exception of apixaban),[10,11] the risk threshold for a given patient would likely depend primarily on his or her relative costs of INR monitoring on warfarin versus the out-of-pocket financial costs of the alternative anticoagulant agents. In this scheme, patients that are more burdened by financial costs would have a risk threshold above 0.5, while those that are more burdened by INR monitoring would have a risk threshold below 0.5.

The net benefit of a prediction model is equal to the true positive rate minus the false positive rate, weighted as a function of the risk threshold.[12] In this case, the net benefit is calculated relative to the strategy of using standard warfarin therapy in all patients. Relative utility is a related measure of the usefulness of a prediction model that is essentially a rescaling of net benefit, and it can be interpreted as the net benefit of the prediction model, compared to using the same treatment strategy in all patients, as a fraction of the net benefit of perfect prediction.[13] A relative utility of 1 indicates that the model performs as well as perfect prediction, while negative values indicate that the model leads to worse outcomes than using the same strategy in everyone.

Supplementary References

1. Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* 2000; **56**: 337–344.
2. French B, Saha-Chaudhuri P, Ky B, Cappola TP, Heagerty PJ. Development and evaluation of multi-marker risk scores for clinical prognosis. *Stat Methods Med Res* 2016; **25**: 255–271.
3. Akritas MG. Nearest Neighbor Estimation of a Bivariate Distribution Under Random Censoring. *Ann Stat* 1994; **22**: 1299–1327.
4. Heagerty PJ, Saha-Chaudhuri P. survivalROC: Time-dependent ROC curve estimation from censored survival data. 2013. Available at: http://cran.r-project.org/package=survivalROC. Accessed March 23, 2016.
5. Miller A. *Subset Selection in Regression*. 2nd ed. Chapman and Hall/CRC, 2002.
6. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer Series in Statistics, 2009.
7. Van Houwelingen JC, Le Cessie S. Predictive value of statistical models. *Stat Med* 1990; **9**: 1303–1325.
8. Steyerberg EW, Eijkemans MJC, Harrell FEJ, Habbema JDF. Prognostic modelling with logistic regression analysis : a comparison of selection and estimation methods in small data sets. *Stat Med* 2000; **19**: 1059–1079.
9. Pauker SG, Kassirer JP. Therapeutic Decision Making: A Cost-Benefit Analysis. *N Engl J Med* 1975; **293**: 229–234.
10. O'Dell KM, Igawa D, Hsin J. New oral anticoagulants for atrial fibrillation: a review of clinical trials. *Clin Ther* 2012; **34**: 894–901.
11. Rollins BM, Silva MA, Donovan JL, Kanaan AO. Evaluation of Oral Anticoagulants for the Extended Treatment of Venous Thromboembolism Using a Mixed-Treatment Comparison, Meta-Analytic Approach. *Clin Ther* 2014; **36**: 1454–1464.
12. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Mak* 2006; **26**: 565–574.
13. Baker SG, Cook NR, Vickers A, Kramer BS. Using relative utility curves to evaluate risk prediction. *J R Stat Soc Ser A* 2009; **172**: 729–48.