Journal of the American Medical Informatics Association, 24(6), 2017, 1052–1061
doi: 10.1093/jamia/ocx030
Advance Access Publication Date: 31 March 2017
Research and Applications

Research and Applications

Calibration drift in regression and machine learning models for acute kidney injury

Sharon E Davis,[1] Thomas A Lasko,[1] Guanhua Chen,[2] Edward D Siew,[3,4] Michael E Matheny[1,2,3,5]

[1]Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN, USA, [2]Department of Biostatistics, Vanderbilt University School of Medicine, [3]Geriatric Research Education and Clinical Care Service, VA Tennessee Valley Healthcare System, Nashville, TN, USA, [4]Division of Nephrology, Vanderbilt University School of Medicine, Vanderbilt Center for Kidney Disease and Integrated Program for AKI, Nashville, TN, USA and [5]Division of General Internal Medicine, Vanderbilt University School of Medicine

All correspondence and reprint requests should be addressed to: Michael E Matheny, GRECC, Room 4-B110, Veterans Administration TVHS, 1310 24th Ave. S., Nashville, TN 37212, USA. E-mail: michael@matheny.info. Phone: 615-873-8017. Fax: 615-873-7981

Received 21 November 2016; Revised 13 February 2017; Accepted 13 March 2017

ABSTRACT

Objective: Predictive analytics create opportunities to incorporate personalized risk estimates into clinical decision support. Models must be well calibrated to support decision-making, yet calibration deteriorates over time. This study explored the influence of modeling methods on performance drift and connected observed drift with data shifts in the patient population.

Materials and Methods: Using 2003 admissions to Department of Veterans Affairs hospitals nationwide, we developed 7 parallel models for hospital-acquired acute kidney injury using common regression and machine learning methods, validating each over 9 subsequent years.

Results: Discrimination was maintained for all models. Calibration declined as all models increasingly overpredicted risk. However, the random forest and neural network models maintained calibration across ranges of probability, capturing more admissions than did the regression models. The magnitude of overprediction increased over time for the regression models while remaining stable and small for the machine learning models. Changes in the rate of acute kidney injury were strongly linked to increasing overprediction, while changes in predictor-outcome associations corresponded with diverging patterns of calibration drift across methods.

Conclusions: Efficient and effective updating protocols will be essential for maintaining accuracy of, user confidence in, and safety of personalized risk predictions to support decision-making. Model updating protocols should be tailored to account for variations in calibration drift across methods and respond to periods of rapid performance drift rather than be limited to regularly scheduled annual or biannual intervals.

Key words: clinical prediction, machine learning, discrimination, calibration, acute kidney injury, clinical decision support

BACKGROUND AND SIGNIFICANCE

Risk prediction models are ubiquitous across clinical specialties and domains,[1–7] though there is limited evidence regarding whether and how modeling methods impact the tendency of model calibration to deteriorate over time.[7–14] Appropriately implemented, well-calibrated models support decision-making to improve patient outcomes, target therapeutic interventions, prioritize resource allocation, and reduce costs.[3,15,16] Historically, risk models relied on relatively few pieces of information, simple algorithms, and manual calculation.[15,17] Widespread adoption of electronic health records enables the delivery of automated,

real-time risk predictions synthesizing a broad set of demographic, clinical, and, most recently, genetic risk factors.[6,16] Integrating predictive analytics into electronic health record systems also enables the use of advanced regression and machine learning methods for model development.[1,16–19] Through the application of such models, clinical decision support is evolving from rule-based to data-driven, probability-based tools. In this paper, we investigate how well several different probabilistic algorithms maintain their clinical risk prediction accuracy over time, using as our example the specific problem of whether a patient will develop acute kidney injury (AKI) during a hospital admission.

AKI is an ideal target for probability-based clinical decision support, given its rapid growth and association with inpatient mortality, length of stay, and long-term sequelae, including chronic kidney and cardiovascular disease.[20–23] Key risk factors for AKI are collected during normal delivery of care and many are modifiable,[4,24] creating opportunities for electronic health record–integrated prediction models to support decision-making. Existing prediction models for AKI have generally been restricted to focused subpopulations and developed with logistic regression[25–34]; however, studies are beginning to pursue AKI models based on large national cohorts and apply advanced regression and machine learning methods.[24,35–37]

Historically, model evaluation has focused on discrimination, which is the ability to separate 2 populations in the data, such as patients with vs without a particular outcome. Discrimination accuracy is an important aspect of predictive models, but it does not assess the accuracy of individual risk predictions, which is crucial when using a predictive model to inform decisions about particular patients. For these patient-level use cases, the more important measure of accuracy is calibration, which estimates how well the predicted probability of an outcome (or *risk*) for a particular patient matches the observed frequency of that outcome among all similar patients.[1–3,8,11,38] Thus, a model might perform well based on discrimination measures while suffering substantial miscalibration. Errors in individual predicted probabilities can lead to overconfidence, inappropriate treatment, or poor allocation of resources.[1,8,39,40] Although the importance of calibration is emphasized in current recommendations,[39–42] many studies focus solely on discrimination,[43–45] and the few existing evaluations of calibration drift over time have been primarily restricted to logistic regression models for hospital mortality.[8,9,12–14]

Model performance can deteriorate over time, particularly in terms of calibration,[7–9,11,41] requiring guidance on recalibration protocols that effectively and efficiently maintain performance. While calibration drift over time is well documented among logistic regression models for hospital mortality,[8,9,12–14,46] the susceptibility of competing modeling methods to performance drift has not been well studied, and performance drift in other domains, such as AKI, has not been well reported.

Leveraging a large national cohort with 10 years of hospitalization data and modeling hospital-acquired AKI as an illustrative clinical outcome, we sought to understand whether modeling approaches influence performance drift, particularly in terms of calibration. Machine learning methods, which can more fully characterize relationships within clinical data than regression models by capturing flexible associations and complex interactions,[18,19,47,48] may be less susceptible to calibration drift than regression methods. We compared the performance over time of models for hospital-acquired AKI developed using 7 common regression and machine learning techniques. To inform modeling and recalibration best practices in this domain, we further studied how patient- and hospital-level characteristics shifted over time to influence model performance.

MATERIALS AND METHODS

We collected data on all admissions to Department of Veterans Affairs (VA) hospitals between January 1, 2003, and December 31, 2012. Laboratory, diagnosis, procedure, radiological study, and medication data were obtained through the national Corporate Data Warehouse.[49] A complete list of predictors is presented in Supplementary Table A1. Variables were selected based on prior publication.[24]

We limited our cohort to admissions with a length of stay between 48 h and 30 days. Admissions were required to have creatinine values measured prior to, within 48 h of, and >48 h after admission. Patients with dialysis or renal transplant prior to admission with community-acquired AKI (based on creatinine values collected between 24 h before and 48 h after admission) and receiving hospice care within 30 days of or 48 h after admission were excluded. Admissions to facilities with fewer than 100 admissions per year were excluded. The cohort was divided into a 1-year development period (2003) and subsequent quarterly validation cohorts (36 periods over 9 years).

We defined AKI stage 1+ as a 0.3 mg/dl or 50% increase from baseline to peak serum creatinine between 48 h and 9 days of admission and staged them according to Kidney Disease – Improving Global Outcomes guidelines.[50] Our predictor set, based on these guidelines and existing literature,[24] included demographics, medications, vital signs, body mass index, laboratory values, and diagnoses collected prior to admission (within 1 year) or within the first 48 h of admission. Missing values were imputed using predictive mean matching (Supplementary Table A1 gives rates of missing values across predictors).

Risk modeling

We implemented 7 common regression (logistic, L-1 penalized logistic,[51] L-2 penalized logistic,[52] and L-1/L-2 penalized logistic[53]) and machine learning (random forest [RF],[54] neural network,[55] and naïve Bayes[56]) methods to predict the probability of AKI for each admission. We fit parallel models using each method based on a common predictor set (Supplementary Table A1) and admissions in the development cohort. For those models with hyperparameters, values were selected using 5-fold cross-validation. Models were internally validated with bootstrap (200 iterations).

Evaluation over time

Model performance

We assessed performance at development and in each validation cohort. Discrimination was measured with the area under the receiver operating characteristics curve (AUC).[57] We characterized calibration with 3 increasingly stringent measures. The observed-to-expected-outcome ratio (O:E) measures agreement between the predicted and observed risk on average across all observations, the weakest form of calibration.[58] The Cox linear logistic recalibration curve assesses weak calibration[58] by characterizing systematic over-/underprediction with the intercept and overfitting, or whether predictions are too extreme for low- and high-risk observations with the slope.[42] We evaluated moderate calibration[58] with flexible calibration curves aimed at assessing alignment between predicted probabilities and observed outcome rates along the spectrum of predicted risk. We summarized these curves as the estimated calibration index (ECI), the mean squared difference between predicted probabilities and estimated observed probabilities based on the fitted flexible calibration curves.[58,59] We further explored how the performance of

each model might have shifted in and out of calibration across the full range of predicted probability by determining regions of predicted probability over which the flexible calibration curves indicated overprediction, underprediction, or successful calibration (illustrations in Supplementary Figure B1). Since observations are not uniformly distributed across the range of predicted probability, we also rescaled the regions of calibration by the volume of observations with predicted probabilities within each region (see illustrations in Supplementary Figure B2). For each metric, mean values and 95% confidence intervals were calculated with the percentile method with 1000 samples. Details regarding metric definitions, ideal values, and interpretations are provided in Supplementary Table B1.

Event rate and case-mix shift

To assess changes in the outcome rate and distribution of individual predictor variables, we calculated the mean or proportion for each continuous or categorical variable, respectively. We tested for linear and nonlinear changes over time in the distribution of each variable, adjusting for multiple comparisons with the Bonferroni method.

We also implemented membership models to explore whether the case mix and event rate were different enough to distinguish between the development and validation cohorts. Using all data from the development cohort and a single validation cohort, these models predict whether an observation is from the validation set based on covariates that include all predictors and the outcome of the original model.[60] We fit separate membership models comparing the development cohort to each consecutive validation cohort using both logistic regression and RF models. The AUCs of the logistic regression membership models were recorded to determine the presence of case mix and event rate shift.[60] AUCs were adjusted for optimism using bootstrap (B = 200). Odds ratios from the logistic regression membership models provided measures of the covariate-adjusted contribution of each predictor to case-mix shift.[60] We documented variable importance ranks from the RF membership models to explore how the relative importance of each predictor to case-mix shift changed over time. For each predictor, we tested for linear and nonlinear changes in variable importance rank over time, adjusting for multiple comparisons with the Bonferroni method.

Predictor-outcome association shift

We assessed predictor-outcome association shift by refitting regression and RF models in each quarterly validation cohort. We documented changes over time in odds ratios from logistic regression models and variable selection patterns from L-1 penalized logistic regression models. Since variable selection may be unstable, we implemented bootstrap (200 iterations) to calculate the proportion of bootstrapped samples in which each predictor was selected during each validation period.[53,61] In RF models, we measured changes over time of variable importance ranks. These refit RF models were developed with the hyperparameters of the original RF model. We adjusted for multiple comparisons with the Bonferroni method.

All analyses were conducted in R 3.2. This study was approved by the Institutional Review Board and the Research and Development Committee of the Tennessee Valley Healthcare System VA.

RESULTS

Our national VA cohort consisted of 1 841 951 admissions, 170 675 during the development period and 1 671 276 during the validation period. Each of the 36 consecutive temporal validation cohorts included a mean of 46 424 admissions (range 42 168–49 798). Summaries of the patient population at select points across the study period are presented in Table 1 and Supplementary Table SA1. Patients were primarily male, with a mean age of 66.1 years (standard deviation 13.0) and mean body mass index of 27.8 (standard deviation 7.5). The AKI rate was 6.8% ($n = 126\,010$) overall, ranging from 7.7% ($n = 13\,090$) in the development period to 6.3% ($n = 2737$) in the final 3-month validation period (Supplementary Figure SA1).

Initial model performance

Performance of the 7 parallel models is presented in Table 2. Discrimination was modest, with AUCs ranging from 0.69 to 0.76. The logistic, L-1 penalized logistic, and L-1/L-2 penalized logistic regression models were most discriminative; the naïve Bayes model was least discriminative. The 4 regression models and neural network model were well calibrated based on O:E ratios and ECIs; the random forest model was well calibrated based on ECI, while slightly underpredicting according to the O:E ratio (1.07, 95% CI, 1.06–1.07). Supplementary Table SB2 gives the results of the Cox linear logistic recalibration curve and accuracy metrics.

Performance over time

Discrimination was stable over the 9-year validation period for the logistic regression, neural network, and RF models, and it declined slightly for the penalized regression and naïve Bayes models (adjusted $P < .007$; Figure 1). While statistically significant, the declines were small. For example, the AUC of the L-1 penalized logistic regression model was 0.76 at development and 0.75 in the final validation cohort. The ranked performance of the different models did not change over time.

Calibration drift was observed in all models (Figure 2 and Supplementary Figure SB3). O:E ratios declined over the first 4 years of validation and were <1 for all models in the second half of the validation period, indicating overprediction. ECI increased over time for all models, indicating deteriorating calibration. In the second half of the validation period, ECIs deteriorated more substantially and became significantly higher for the regression models compared to the RF and neural network models. The naïve Bayes model consistently underperformed all other models.

The ranges of predicted probabilities and proportion of admissions over which each model was calibrated changed over time. These changes drive much of the calibration drift, as models well calibrated in regions of predicted probability capturing most of the data perform better overall than models well calibrated in sparsely populated regions of predicted probability. Figure 3 presents the regional calibration on the proportional volume scale with the cumulative fraction of admissions, and Supplementary Figure SB4 presents the regional calibration on the original predicted probability scale. For regression models, the majority of admissions fell within regions of overprediction, with the magnitude of overprediction increasing in later cohorts. For neural network and RF models, the majority of admissions fell within reasonably calibrated regions during the first 3 years, deteriorating somewhat in later cohorts while remaining less severely overpredicted than the regression models.

Population shifts over time

Event rate shift

Declines over time in the AKI rate corresponded with calibration drift in all models. O:E ratios and ECIs were strongly positively and

Table 1. Patient population at development (2003) and in 3 years of the validation period (2006, 2009, 2012)

Admission characteristics	2003	2006	2009	2012
N	170 675	176 341	193 917	184 827
% AKI	7.7	7.4	6.5	6.2
Age in years (mean and SD)	65.7 (12.9)	65.9 (12.9)	66.1 (13.0)	66.5 (13.0)
% Female	3.2	3.7	4.0	4.5
Race				
% White	75.0	75.9	75.4	74.9
% Black	20.1	19.1	19.0	19.1
% American Indian/Alaskan	0.8	0.9	0.9	0.9
% Asian/Pacific Islander	0.9	1.1	1.2	1.1
% Unreported	3.2	3.1	3.5	4.0
BMI at admission (mean and SD)	27.4 (7.8)	27.7 (7.7)	28.1 (7.9)	28.4 (7.5)
Mean outpatient GFR prior to admission (mean and SD)	69.5 (24.5)	70.5 (24.9)	72.3 (25.4)	74.5 (26.4)
Select medications (admission window)				
Vancomycin	5.3	10.4	14.5	16.3
ACEi	32.9	34.1	31.4	27.7
Antiemetics	3.3	5.1	9.3	13.2
Beta blockers	40.1	48.6	48.4	37.1
Opioids	50.8	59.2	63.3	64.3
Statins	27.9	38.9	43.8	44.0
Select diagnoses (preadmission)				
Anemia	14.5	23.3	28.6	31.1
Cancer	18.8	22.8	24.6	24.9
Chronic obstructive pulmonary disease	24.6	30.9	34.0	35.1
Congestive heart failure	15.2	18.6	19.7	20.0
Diabetes mellitus	29.7	34.6	39.1	42.6
Dyslipidemia	28.8	49.7	59.7	65.4
Alcoholism	12.1	18.9	23.5	26.4
Hypertension	55.0	69.4	74.5	76.7

Abbreviations: AKI = acute kidney injury; ACEi = angiotensin-converting enzyme inhibitor; GFR = glomerular filtration rate; SD = standard deviation.

Table 2. Model performance in development cohort

Performance metric	LR	Regression			RF	Machine Learning	
		L1	L2	L1-L2		NN	NB
Discrimination							
AUC	0.76 (0.76–0.76)	0.76 (0.76–0.76)	0.75 (0.75–0.75)	0.76 (0.76–0.76)	0.73 (0.73–0.73)	0.72 (0.72–0.72)	0.69 (0.69–0.69)
Calibration							
O:E	1.00 (1.00–1.00)	1.00 (1.00–1.00)	1.00 (1.00–1.00)	1.00 (1.00–1.00)	1.07 (1.06–1.07)	1.00 (1.00–1.01)	0.22 (0.22–0.23)
ECI	0.00 (0.00–0.00)	0.00 (0.00–0.00)	0.01 (0.01–0.01)	0.00 (0.00–0.00)	0.01 (0.01–0.01)	0.01 (0.01–0.01)	21.03 (20.19–21.87)

Regression methods included logistic regression (LR), L-1 penalized logistic regression (L1), L-2 penalized logistic regression (L2), and L-1/L-2 penalized logistic regression (L1–L2). Machine learning methods included random forest (RF), neural networks (NN), and naïve Bayes (NB).

Abbreviations: AUC = area under the receiver operating characteristics curve; ECI = estimated calibration index; O:E = observed-to-expected-outcome ratio.

negatively correlated with event rate, respectively (adjusted $P < .001$). AKI was also a significant predictor in membership models, predicting whether admissions belonged to the validation or development cohort. Additional results are summarized in Supplementary Figure SA1.

Case-mix shift

Shifts were observed in individual predictors from one cohort to the next. Linear, monotone nonlinear, and nonmonotone changes occurred in 95% of predictors (adjusted $P < .0002$). The largest changes were in the prevalence of certain medications and chronic diseases (Table 1 and Supplementary Table SA1). The smallest changes were in the use of antifungals and monoamine oxidase inhibitors before admission and use of nonsteroidal anti-inflammatory drugs, monoamine oxidase inhibitors, anhydrase diuretics, and lithium during the admission window.

Membership models increasingly discriminated between admissions from the development and validation cohorts as time progressed (AUC increased from 0.60 to 0.92; Supplementary Figure SA2). Fluctuating odds ratios and variable importance ranks for most predictors were observed without clear patterns over time and prevented direct linking of changes in individual predictors with performance drift. Detailed results for 6 exemplar predictors (use of antiemetics and vancomycin during the admission window, age, mean outpatient glomerular filtration rate [GFR] prior to admission,

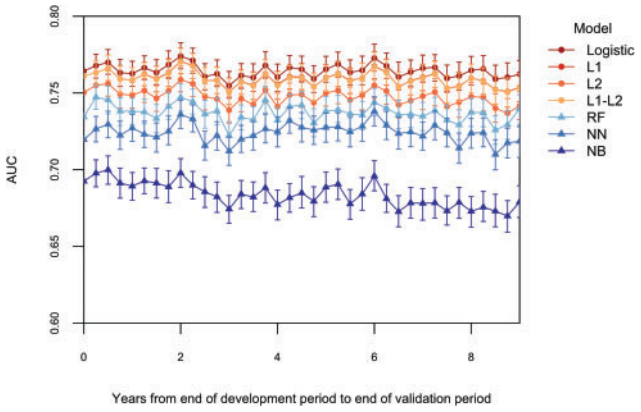

Figure 1. Discrimination over time by modeling method. Regression models – logistic regression (LR), L-1 penalized logistic regression (L1), L-2 penalized logistic regression (L2), and L-1/L-2 penalized logistic regression (L1-L2) – are displayed with circular markers and red-orange colors. Machine learning models – random forest (RF), neural network (NN), and naïve Bayes (NB) – are displayed with triangular markers and blue-purple colors.

history of cancer, and history of diabetes) are presented in Supplementary Figures SA3–SA5.

Predictor-outcome association shift

Changes in the strength of associations between predictors and AKI were measured by changes in the structure of models refit in each validation cohort (Table 3). No changes were observed over time for the majority of predictors. In logistic regression models, we observed few significant changes over time in the strength of predictor-outcome. However, while we observed tendencies toward strengthening or weakening of associations over time for a limited number of predictors, these changes generally did not reach statistical significance. This pattern was most pronounced for use of vancomycin during the admission window, age at admission, and history of diabetes, each of which experienced shifts in odds ratios during the second half of the validation period. In L-1 penalized regression models, 6 predictors (use of loop and thiazide diuretics during the admission window, mean and change in GFR in the admission window, mean outpatient GFR before admission, and race) were consistently selected for inclusion, indicating no meaningful association shift. Two predictors, age and use of fluoroquinolones during the admission window, changed over time in their selection frequency (adjusted $P < .0002$). In RF models, 11 predictors (9.3%) experienced significant changes in variable importance (adjusted $P < .0002$). Shifts in associations were concentrated in the second half of the validation period. Full results for the 6 exemplar predicators are presented in Supplementary Figures SA6–SA8.

DISCUSSION

In this rigorous comparison of regression and machine learning models for development of AKI during a hospital admission, discrimination remained quite stable over time and calibration deteriorated substantially, with all methods drifting toward overprediction within 1 year of development. While discrimination statistically significantly declined over time for the penalized regression and naïve Bayes models, the magnitude of these changes was minimal and this did not result in practically meaningful changes in AUCs. For the most stringent calibration measures, machine learning models

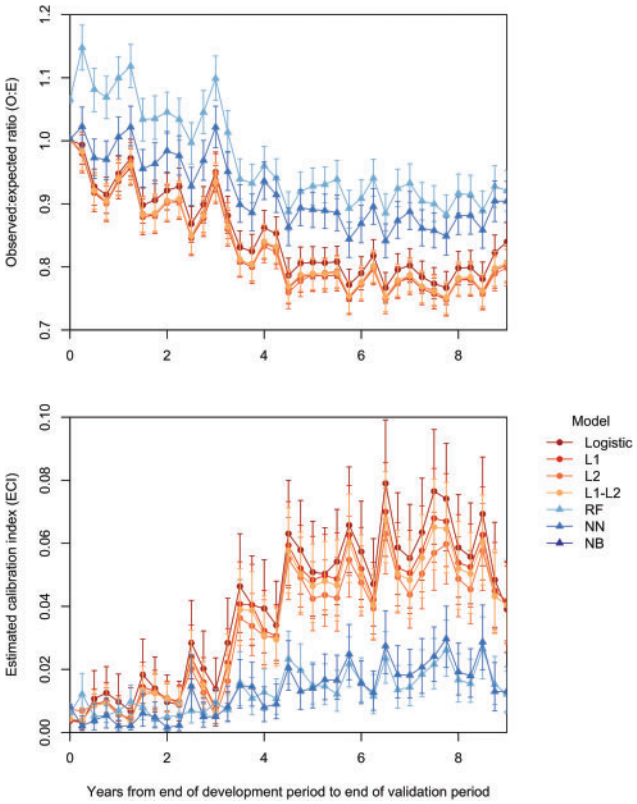

Figure 2. Calibration over time by modeling method. Regression models – logistic regression (LR), L-1 penalized logistic regression (L1), L-2 penalized logistic regression (L2), and L-1/L-2 penalized logistic regression (L1-L2) – are displayed with circular markers and red-orange colors. Machine learning models – random forest (RF), neural network (NN), and naïve Bayes (NB) – are displayed with triangular markers and blue-purple colors. Due to the large discrepancy between NB performance and performance of the other models, the vertical axes are scaled such that NB values are excluded from the plots.

exhibited superior stability compared to regression models. Decreases in the rate of AKI over time coincided with increasing overprediction in all models, while changes in predictor-outcome associations temporally corresponded with diverging calibration between machine learning and regression models.

The stability of discrimination and increasing overprediction parallel previous findings among logistic regression hospital mortality models.[8,9,12–14] Although few clinical studies compare calibration drift across methods, 1 study found that O:E ratios were stable for a tree-based model but deteriorated within 4 years for a corresponding logistic regression model.[9,10] While we did not implement a single tree-based model, our RF model experienced O:E ratio drift that paralleled our logistic regression model. However, O:E ratios indicated less overprediction by our RF model than our logistic regression model. We have expanded on this limited literature by comparing a wider set of methods and exploring calibration in more detail.

Deterioration in ECIs over time was substantively greater for regression models compared to machine learning models. One important reason is that the neural network and RF models maintained calibration over probability ranges that covered more of the data. In addition, beginning 3 years after development, the regression models exhibited an increasing magnitude of overprediction, while the RF and neural network models had more stable and lower levels of overprediction. These results suggest that for patient-level

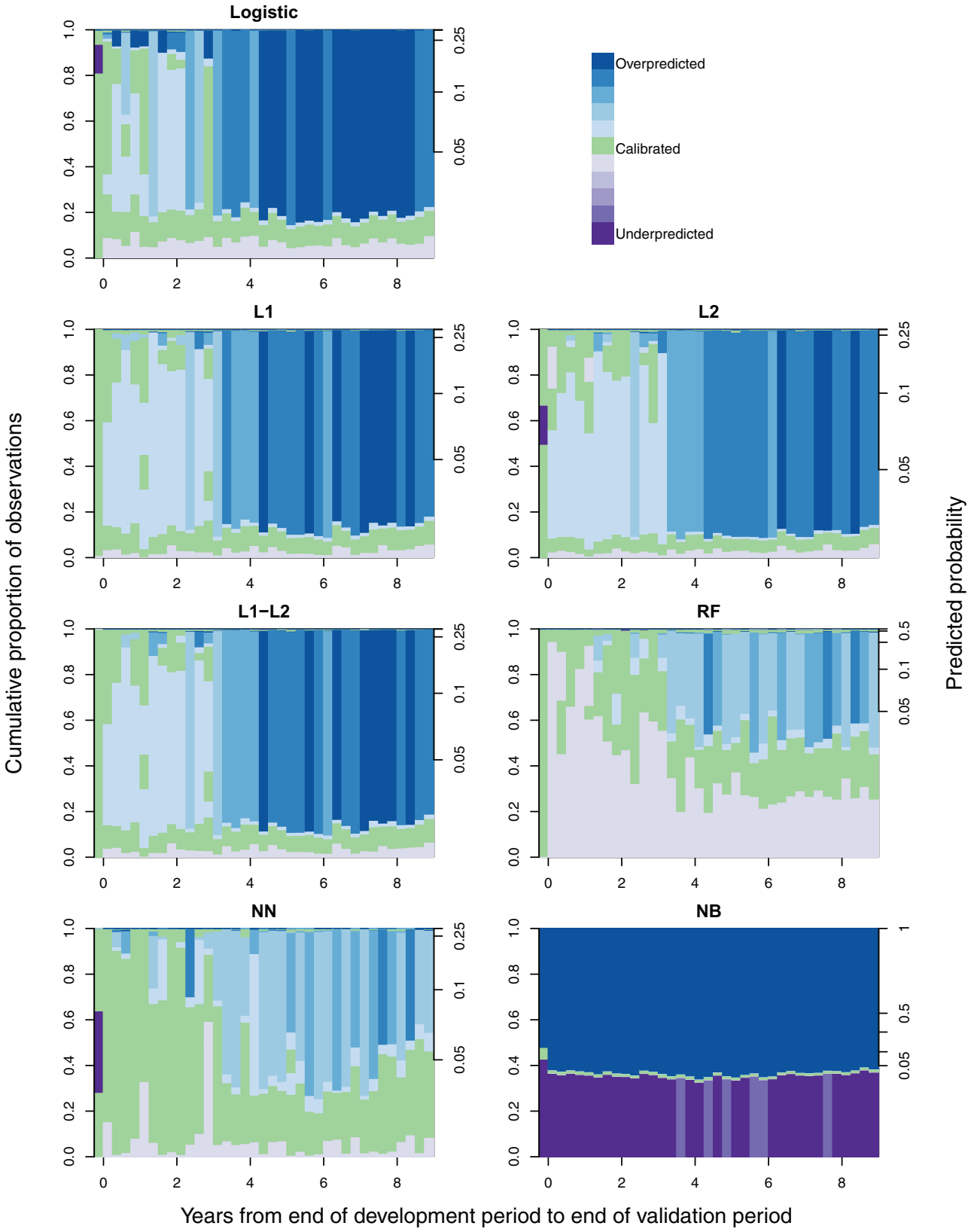

Figure 3. Regions of calibration across the range of predicted probabilities, scaled by proportion of observations in each region and shaded by the magnitude of the within-region estimated calibration index. Regression models include logistic regression (LR), L-1 penalized logistic regression (L1), L-2 penalized logistic regression (L2), and L-1/L-2 penalized logistic regression (L1-L2). Machine learning models include random forest (RF), neural network (NN), and naïve Bayes (NB).

Table 3. Odds ratios from logistic regression models (LR OR), variable importance ranks from random forest models (RF rank), and proportion of times predictors were selected in 200 L-1 penalized logistic regression (L1) modeling iterations based on models fit at development and within select temporal validation cohorts

Predictor	Development (2003)			2006 – Q4			2009 – Q4			2012 – Q4		
	LR OR[a] (95% CI)	RF rank	L1 % selected	LR OR (95% CI)	RF rank	L1 % selected	LR OR (95% CI)	RF rank	L1 % selected	LR OR (95% CI)	RF rank	L1 % selected
Highest ranking predictors at development												
Mean GFR in admission window	0.16 (0.15–0.18)	1	100	0.18 (0.15–0.22)	1	100	0.14 (0.12–0.17)	1	100	0.13 (0.11–0.16)	1	100
Change in GFR during admission window	0.61 (0.59–0.64)	2	100	0.62 (0.57–0.67)	2	100	0.61 (0.56–0.66)	2	100	0.59 (0.55–0.64)	2	100
Mean outpatient GFR before admission	4.14 (3.81–4.51)	3	100	4.39 (3.72–5.18)	3	100	4.60 (3.87–5.46)	3	100	3.75 (3.15–4.47)	3	100
Blood urea nitrogen	1.26 (1.2–1.33)	4	100	1.36 (1.23–1.51)	4	100	1.27 (1.14–1.42)	4	100	1.06 (0.94–1.20)	7	88
BMI at admission	1.00 (0.90–1.12)	5	77	1.01 (0.81–1.25)	5	50.5	1.16 (0.93–1.45)	5	86	1.05 (0.81–1.36)	4	93
White blood cell count	1.20 (1.15–1.26)	6	100	1.18 (1.08–1.29)	11	98.5	1.22 (1.11–1.34)	9	100	1.24 (1.12–1.37)	6	100
Platelets	0.89 (0.85–0.94)	7	100	0.89 (0.81–0.97)	9	75	1.01 (0.91–1.13)	11	75	1.08 (0.96–1.22)	5	90.5
Alkaline phosphatase	1.06 (1.02–1.09)	8	100	1.02 (0.94–1.1)	8	76	1.01 (0.93–1.09)	8	89.5	1.02 (0.93–1.12)	8	82
Glucose	1.06 (1.01–1.11)	9	97.5	0.99 (0.91–1.08)	6	50	1.09 (1.00–1.19)	6	73	1.04 (0.95–1.15)	11	90
Standard deviation of preadmission GFR	0.99 (0.95–1.04)	10	81.5	0.99 (0.90–1.09)	12	42.5	0.94 (0.86–1.04)	12	45	1.05 (0.95–1.16)	13	88.5
Variables with shifts in association												
Age	1.22 (1.15–1.29)	21	100	1.27 (1.13–1.43)	19	100	1.07 (0.94–1.22)	18	88.5	1.12 (0.97–1.3)	25	94
GFR count during admission window	1.02 (0.97–1.07)	33	100	0.98 (0.90–1.07)	32	99	1.03 (0.95–1.12)	33	64	1.00 (0.95–1.06)	29	89.5
History of hypertension	1.25 (1.31–1.19)	39	100	1.39 (1.57–1.24)	39	100	1.36 (1.56–1.19)	54	100	1.24 (1.43–1.07)	60	100
History of diabetes mellitus	1.14 (1.08–1.21)	40	100	1.14 (1.03–1.27)	41	100	1.03 (0.92–1.15)	47	80.5	1.11 (0.99–1.24)	49	97.5
ACEi within 90 days prior to admission	1.15 (1.09–1.21)	41	100	1.05 (0.95–1.17)	53	84	1.15 (1.03–1.28)	60	98.5	1.01 (0.9–1.13)	65	47
CCB within 90 days prior to admission	1.16 (1.10–1.23)	50	100	1.09 (0.98–1.23)	54	87.5	1.03 (0.91–1.16)	80	49	1.06 (0.93–1.2)	69	77
History of cancer	1.22 (1.16–1.28)	61	100	1.09 (0.99–1.20)	74	86	0.95 (0.87–1.05)	101	48.5	1.05 (0.95–1.16)	93	71.5
History of dyslipidemia	1.03 (0.99–1.09)	65	94	1.02 (0.93–1.13)	85	57	1.17 (1.05–1.30)	93	97	1.01 (0.90–1.13)	104	39.5
Fluoroquinolones during admission window	1.05 (0.97–1.15)	66	86.5	0.88 (0.75–1.02)	100	64	0.89 (0.75–1.04)	81	58.5	0.88 (0.72–1.06)	95	73
Lactated Ringer's IVF	2.43 (0.33–17.72)	67	98	2.34 (0.21–25.62)	42	62.5	0.41 (0.01–16.52)	44	71.5	8.43 (0.77–91.72)	40	69
Vancomycin during admission window	1.13 (1.03–1.23)	81	100	1.18 (1.04–1.34)	62	98.5	1.44 (1.28–1.63)	43	100	1.58 (1.39–1.79)	39	100
History of COPD	1.07 (1.02–1.12)	84	98.5	1.02 (0.93–1.11)	92	36.5	0.95 (0.87–1.05)	97	38.5	1.05 (0.95–1.15)	97	77

(continued)

Table 3. continued

Predictor	Development (2003)			2006 – Q4			2009 – Q4			2012 – Q4		
	LR OR[a] (95% CI)	RF rank	L1 % selected	LR OR (95% CI)	RF rank	L1 % selected	LR OR (95% CI)	RF rank	L1 % selected	LR OR (95% CI)	RF rank	L1 % selected
Penicillins during admission window	0.97 (0.92–1.04)	86	49	1.05 (0.94–1.18)	89	67	0.94 (0.83–1.06)	85	32	0.99 (0.88–1.13)	59	63.5
Antiemetics during admission window	1.20 (1.07–1.35)	90	99	1.14 (0.96–1.36)	80	72.5	1.31 (1.14–1.51)	50	98	1.22 (1.07–1.39)	68	96.5
Antiemetics within 90 days prior to admission	1.09 (0.97–1.23)	101	83	1.13 (0.90–1.41)	82	50	1.01 (0.81–1.26)	89	29	1.18 (0.98–1.42)	71	89

[a]Odds ratios for continuous predictors are for an interquartile range increase in value.

Abbreviations: GFR = glomerular filtration rate; ACEi = angiotensin-converting enzyme inhibitor; CCB = calcium channel blocker; COPD = chronic obstructive pulmonary disease; OR = odds ratio; CI = confidence interval; LR = logistic regression; RF = random forest; L1 = L-1 penalized logistic regression; BMI = body mass index; IVF = intravenous fluid.

prediction, RF and neural network models may have different updating requirements compared to regression models.

Performance drift results from data shifts in the patient population, including changes in outcome rate, patient case mix, clinical practice, and documentation practices.[3,6–8,11,60] We observed complex, multidimensional data shifts in our cohort. A declining proportion of admissions complicated by AKI was strongly correlated with overprediction, resulting in drift of both O:E ratios and Cox intercepts. While membership models clearly indicated the presence of case-mix shift, the complexity of co-occurring changes across the majority of predictors prevented us from directly linking changes in individual predictors with calibration. Although we observed limited evidence of shifts in the strength of associations between predictors and AKI, these shifts tended to occur over the second half of the validation period, corresponding with the years during which the regression models experienced more substantial calibration drift than the machine learning models. This may indicate greater susceptibility of regression methods to predictor-outcome association shift compared to RF and neural network models. The ability of the RF and neural network models to capture complex interactions and nonlinear associations[18,19,47,48] may make these methods more robust to subtle association shifts. Additional studies, especially simulations isolating association shifts, are necessary to explore this hypothesis.

These findings have important implications for long-term use of predictive modeling in clinical decision support for AKI and other clinical domains. For use cases focused on classifying patients by risk level where discrimination is more important than calibration, our finding of stable discrimination across models may alleviate concerns regarding model updating. However, when individual patient risk predictions are of interest, calibration becomes the most important aspect of model accuracy, and the substantial calibration drift that we observed in all models suggests that periodic recalibration or retraining is necessary. In our cohort, we found that regression models had a greater need for updating compared to RF and neural network models.

This study also provides evidence that modeling methods are variably susceptible to calibration drift in ways that may not be apparent based on traditional calibration assessments and may be particularly relevant to clinical utility. We observed similar calibration drift patterns across methods when considering the commonly reported O:E ratio and Cox logistic recalibration curve. These weak measures of calibration, however, may not be sufficient for the use of personalized patient-level risk predictions.[58] More stringent calibration evaluations based on flexible calibration curves ensure that models have a net benefit greater than or equal to treat-all or treat-none strategies, thus ensuring that predictions are not harmful to clinical decision-making.[58] The variable robustness of the different models under these more sensitive measures suggests that model updating protocols should be tailored to the particular model used, and that calibration accuracy should be evaluated frequently.

Our findings can inform recommendations regarding the timing of and approach to clinical prediction model updating. We recommend repeated validation for all models, with careful consideration of the timing of repeated assessment and sample sizes supporting each assessment,[62] and whether discrimination or stringent calibration metrics should be measured. In addition, monitoring population data shifts could provide early warning of the need for model updating and insight into the updating approach required to correct performance. Recalibration would be indicated in the case of calibration drift with event rate shift and stable discrimination; however, when calibration drift occurs in the presence of predictor-outcome association shifts, significant case-mix shift, or changes in discrimination, full model revision may be required.[3,62–64] Tracking of model performance and population shifts could be managed through the implementation of surveillance systems.[7] As our study suggests, models based on different methods have variable updating needs; such surveillance systems should be tailored to the distinct susceptibilities of modeling methods to performance drift. Of course, we also emphasize the importance of local knowledge, as changes in local clinical practice or data definitions should always trigger model updates.[63] Finally, we recommend flexibility in updating protocols, in order to address miscalibration as it occurs rather than at scheduled intervals.

Our study is not without limitations. Although we utilized a large national cohort, we examined performance drift in a single population data shift scenario. Replicating this study in settings influenced by different combinations and intensities of event rate, case mix, and association shift may allow generalizable conclusions regarding the susceptibility of modeling methods to performance drift. The complexity of linking multiple co-occurring forms of data shift with model performance limits our ability to understand why certain methods may be more robust to particular forms of data

shift. Changes in the frequency of omitted variables or the association between such variables and AKI may also impact model performance in unexpected ways. Simulation studies with isolated, prespecified shifts could address this limitation. In addition, calibration does not directly characterize clinical utility. Miscalibration in some ranges of predicted probability may not impact clinical decisions, particularly among patients with the highest and lowest risk, for whom appropriate decisions may be clear to clinicians. Understanding whether, when, and how calibration drift affects the clinical utility of predictions for decision-making is an important consideration in informing recalibration guidelines. A recent study recommended that assessment of flexible calibration curves be required to ensure nonharmful predictions,[58] making our findings regarding the divergent patterns of ECI and regions of calibration across models an initial assessment of clinical utility and an important consideration in model implementation.

CONCLUSION

Growing opportunities to leverage predictive analytics to integrate personalized risk prediction into clinical decision support requires well-calibrated models consistently providing accurate predictions. This study extends our understanding of model performance over time across modeling methods. Our finding of stable discrimination over time may alleviate model updating concerns for predictions used to assign risk levels rather than individualized risk estimates. However, our calibration drift findings strongly support the need for routine recalibration of models incorporated into clinical decision support tools presenting personalize predicted probabilities. Recalibration protocols should be tailored to account for variations in calibration drift across modeling methods. Given the short time frames over which we documented significant calibration drift, recalibration protocols should be flexible enough to respond to periods of rapid performance drift rather than limited to regularly scheduled intervals. Efficient and effective updating protocols will be essential to maintain accuracy of and user confidence in personalized risk predictions integrated into clinical decision support for AKI and other clinical outcomes. While the full suite of best practice guidelines remains to be developed, modeling methods will be an important component in determining when and how clinical prediction models must be revised.

COMPETING INTERESTS

The authors have no conflicts of interest to disclose.

FUNDING

This work was supported by National Library of Medicine grants 5T15LM007450-15 and 1R21LM011664-01; Veterans Health Administration grants VA HSR&D CDA-08-020, VA HSR&D IIR 11-292, VA HSR&D IIR 13-052, and VA HSR&D IIR 13-073; the Edward Mallinckrodt Jr Foundation; and the Vanderbilt Center for Kidney Disease.

CONTRIBUTORS

SED and MEM designed the study and acquired the data. TAL and GC contributed to the design and selection of data analysis methods. EDS provided critical clinical domain knowledge. SED conducted all data analysis and drafted the initial manuscript. All authors contributed to interpretation of the results and critical revision of the manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGMENTS

Initial results of this study were presented at the American Medical Informatics Association 2016 Summit on Clinical Research Informatics, with an associated abstract published in the conference proceedings, as well as at the National Library of Medicine 2016 Informatics Training Conference.

REFERENCES

1. Amarasingham R, Patzer RE, Huesch M, Nguyen NQ, Xie B. Implementing electronic health care predictive analytics: considerations and challenges. *Health Affairs*. 2014;33(7):1148–54.
2. Ohno-Machado L, Resnic FS, Matheny ME. Prognosis in critical care. *Ann Rev Biomed Engineering*. 2006;8:567–99.
3. Moons KG, Kengne AP, Grobbee DE, *et al*. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart*. 2012;98(9):691–98.
4. Matheny ME, Miller RA, Ikizler TA, *et al*. Development of inpatient risk stratification models of acute kidney injury for use in electronic health records. *Med Decision Making*. 2010;30(6):639–50.
5. Kansagara D, Englander H, Salanitro A, *et al*. Risk prediction models for hospital readmission: a systematic review. *JAMA*. 2011;306(15):1688–98.
6. Steyerberg EW, Moons KG, van der Windt DA, *et al*. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med*. 2013;10(2):e1001381.
7. Toll DB, Janssen KJ, Vergouwe Y, Moons KG. Validation, updating and impact of clinical prediction rules: a review. *J Clin Epidemiol*. 2008;61(11):1085–94.
8. Hickey GL, Grant SW, Murphy GJ, *et al*. Dynamic trends in cardiac surgery: why the logistic euroscore is no longer suitable for contemporary cardiac surgery and implications for future risk models. *Eur J Cardiothoracic Surg*. 2013;43(6):1146–52.
9. Minne L, Eslami S, De Keizer N, De Jonge E, De Rooij SE, Abu-Hanna A. Effect of changes over time in the performance of a customized SAPS-II model on the quality of care assessment. *Intensive Care Med*. 2012;38(1):40–46.
10. Minne L, Eslami S, de Keizer N, de Jonge E, de Rooij SE, Abu-Hanna A. Statistical process control for monitoring standardized mortality ratios of a classification tree model. Methods Inf Med. 2012;51(4):353–58.
11. Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *BMJ*. 2009;338:b606.
12. Harrison DA, Lone NI, Haddow C, *et al*. External validation of the Intensive Care National Audit & Research Centre (ICNARC) risk prediction model in critical care units in Scotland. *BMC Anesthesiol*. 2014;14:116.
13. Paul E, Bailey M, Van Lint A, Pilcher V. Performance of APACHE III over time in Australia and New Zealand: a retrospective cohort study. *Anaesthesia Intensive Care*. 2012;40(6):980–94.
14. Madan P, Elayda MA, Lee VV, Wilson JM. Risk-prediction models for mortality after coronary artery bypass surgery: application to individual patients. *Int J Cardiol*. 2011;149(2):227–31.
15. Amarasingham R, Audet AJ, Bates DW, *et al*. Consensus statement on electronic health predictive analytics: a guiding framework to address challenges. *eGEMs*. 2016;4(1):1–11.
16. Parikh RB, Kakad M, Bates DW. Integrating predictive analytics into high-value care: the dawn of precision delivery. *JAMA*. 2016;315(7):651–52.
17. Pencina MJ, Peterson ED. Moving from clinical trials to precision medicine: the role for predictive modeling. *JAMA*. 2016;315(16):1713–14.

18. Sajda P. Machine learning for detection and diagnosis of disease. *Ann Rev Biomed Engineering*. 2006;8:537–65.
19. Steyerberg EW, van der Ploeg T, Van Calster B. Risk prediction with machine learning and regression methods. *Biometrical J. Biometrische Zeitschrift*. 2014;56(4):601–06.
20. Uchino S, Kellum JA, Bellomo R, *et al*. Acute renal failure in critically ill patients: a multinational, multicenter study. *JAMA*. 2005;294(7):813–18.
21. Brivet FG, Kleinknecht DJ, Loirat P, Landais PJ. Acute renal failure in intensive care units – causes, outcome, and prognostic factors of hospital mortality: a prospective, multicenter study. French Study Group on Acute Renal Failure. *Crit Care Med*. 1996;24(2):192–98
22. Coca SG, Yusuf B, Shlipak MG, Garg AX, Parikh CR. Long-term risk of mortality and other adverse outcomes after acute kidney injury: a systematic review and meta-analysis. *Am J Kidney Dis*. 2009;53(6):961–73.
23. Liano F, Junco E, Pascual J, Madero R, Verde E. The spectrum of acute renal failure in the intensive care unit compared with that seen in other settings. *The Madrid Acute Renal Failure Study Group*. Kidney Int Suppl. 1998;66:S16–24.
24. Cronin RM, VanHouten JP, Siew ED, *et al*. National Veterans Health Administration Inpatient Risk Stratification Models for Hospital-Acquired Acute Kidney Injury. *J Am Med Inform Assoc*. 2015;22(5):1054–71.
25. Breidthardt T, Christ-Crain M, Stolz D, *et al*. A combined cardiorenal assessment for the prediction of acute kidney injury in lower respiratory tract infections. *Am J Med*. 2012;125(2):168–75.
26. Kim WH, Lee SM, Choi JW, *et al*. Simplified clinical risk score to predict acute kidney injury after aortic surgery. *J Cardiothorac Vasc Anesth*. 2013;27(6):1158–66.
27. Kristovic D, Horvatic I, Husedzinovic I, *et al*. Cardiac surgery–associated acute kidney injury: risk factors analysis and comparison of prediction models. *Interact Cardiovasc Thorac Surg*. 2015;21(3):366–73.
28. McMahon GM, Zeng X, Waikar SS. A risk prediction score for kidney failure or mortality in rhabdomyolysis. *JAMA Int Med*. 2013;173(19):1821–28.
29. Ng SY, Sanagou M, Wolfe R, Cochrane A, Smith JA, Reid CM. Prediction of acute kidney injury within 30 days of cardiac surgery. *J Thoracic Cardiovasc Surgery*. 2014;147(6):1875–83, 83 e1.
30. Park MH, Shim HS, Kim WH, *et al*. Clinical risk scoring models for prediction of acute kidney injury after living donor liver transplantation: a retrospective observational study. *PloS One*. 2015;10(8):e0136230.
31. Slankamenac K, Beck-Schimmer B, Breitenstein S, Puhan MA, Clavien PA. Novel prediction score including pre- and intraoperative parameters best predicts acute kidney injury after liver surgery. *World J Surgery*. 2013;37(11):2618–28.
32. Wang YN, Cheng H, Yue T, Chen YP. Derivation and validation of a prediction score for acute kidney injury in patients hospitalized with acute heart failure in a Chinese cohort. *Nephrology*. 2013;18(7):489–96.
33. Rodriguez E, Soler MJ, Rap O, Barrios C, Orfila MA, Pascual J. Risk factors for acute kidney injury in severe rhabdomyolysis. *PloS One*. 2013;8(12):e82992.
34. Schneider DF, Dobrowolsky A, Shakir IA, Sinacore JM, Mosier MJ, Gamelli RL. Predicting acute kidney injury among burn patients in the 21st century: a classification and regression tree analysis. *J Burn Care Res*. 2012;33(2):242–51.
35. Legrand M, Pirracchio R, Rosa A, *et al*. Incidence, risk factors and prediction of post-operative acute kidney injury following cardiac surgery for active infective endocarditis: an observational study. *Crit Care*. 2013;17(5):R220.
36. Brown JR, MacKenzie TA, Maddox TM, *et al*. Acute kidney injury risk prediction in patients undergoing coronary angiography in a national Veterans Health Administration cohort with external validation. *J Am Heart Assoc*. 2015;4(12):e002136.
37. Gurm HS, Seth M, Kooiman J, Share D. A novel tool for reliable and accurate prediction of renal complications in patients undergoing percutaneous coronary intervention. *J Am Coll Cardiol*. 2013;61(22):2242–48.
38. Matheny ME, Ohno-Machado L, Resnic FS. Discrimination and calibration of mortality risk prediction models in interventional cardiology. *J Biomed Inform*. 2005;38(5):367–75.
39. Jiang X, Osl M, Kim J, Ohno-Machado L. Calibrating predictive model estimates to support personalized medicine. *J Am Med Inform Assoc*, 2012;19(2):263–74.
40. Van Calster B, Vickers AJ. Calibration of risk prediction models: impact on decision-analytic performance. *Med Decis Mak*. 2015;35(2):162–69.
41. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. New York, NY: Springer; 2009.
42. Steyerberg EW, Vickers AJ, Cook NR, *et al*. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21(1):128–38.
43. Bouwmeester W, Zuithoff NP, Mallett S, *et al*. Reporting and methods in clinical prediction research: a systematic review. *PLoS Med*. 2012;9(5):1–12.
44. Collins GS, de Groot JA, Dutton S, *et al*. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol*. 2014;14:40.
45. Mallett S, Royston P, Waters R, Dutton S, Altman DG. Reporting performance of prognostic models in cancer: a review. *BMC Med*. 2010;8:21.
46. Cook DA, Joyce CJ, Barnett RJ, *et al*. Prospective independent validation of APACHE III models in an Australian tertiary adult intensive care unit. *Anaesth Intensive Care*. 2002;30(3):308–15.
47. Malley JD, Kruppa J, Dasgupta A, Malley KG, Ziegler A. Probability machines: consistent probability estimation using nonparametric learning machines. *Methods Inform Med*. 2012;51(1):74–81.
48. Breiman L. Statistical modeling: the two cultures. *Statistical Science*. 2001;16(3):199–231
49. Perlin JB, Kolodner RM, Roswell RH. The Veterans Health Administration: quality, value, accountability, and information as transforming strategies for patient-centered care. *Am J Managed Care*. 2004;10(11 Pt 2):828–36.
50. Khwaja A. KDIGO clinical practice guidelines for acute kidney injury. *Nephron Clin Pract*. 2012;120(4):c179–84.
51. Tibshirani R. Regression shrinkage and selection via the lasso. *J Royal Stat Soc Series B*. 1996;58(1):267–88.
52. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*. 1970;12(1):55–67.
53. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J Royal Stat Soc Series B*. 2005;67(2):301–20.
54. Breiman L. Random forests. *Machine Learning*. 2001;45(1):5–32.
55. Bishop CM. *Neural Networks for Pattern Recognition*. New York: Oxford University Press; 1995.
56. Hand DJ. Naive Bayes. In: Wu X, Kumar V, eds. *The Top Ten Algorithms in Data Mining*. Chapman & Hall/CRC; 2009:163–78.
57. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29–36.
58. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol*. 2016;74:167–76.
59. Van Hoorde K, Van Huffel S, Timmerman D, Bourne T, Van Calster B. A spline-based tool to assess and visualize the calibration of multiclass risk predictions. *J Biomed Inform*. 2015;54:283–93.
60. Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol*. 2015;68(3):279–89.
61. Meinshausen N, Bühlmann P. Stability selection. *J Royal Stat Soc Series B (Statistical Methodology)*. 2010;72(4):417–73.
62. Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJ, Habbema JD. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Stat Med*. 2004;23(16):2567–86.
63. Kappen TH, Vergouwe Y, van Klei WA, van Wolfswinkel L, Kalkman CJ, Moons KG. Adaptation of clinical prediction models for application in local settings. *Med Decis Mak*. 2012;32(3):E1–10.
64. Janssen KJ, Moons KG, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. *J Clin Epidemiol*. 2008;61(1):76–86.