Performance Drift of Clinical Prediction Models: Impact of modeling methods on prospective model performance

By

Sharon E. Davis

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Biomedical Informatics

May, 2017

Nashville, Tennessee

Approved By:

Michael E. Matheny, M.D., M.S., M.P.H.

Thomas A. Lasko, M.D., Ph.D.

Guanhua Chen, Ph.D.

ACKNOWLEDGEMENTS

This work was made possible by the generous support of multiple funding agencies. My graduate studies and research were supported by a training grant from the National Library of Medicine (grant number: 5T15LM007450-15). Time and effort of my committee members and key collaborators were supported by the National Library of Medicine (grant number: 1R21LM011664-01), the U.S. Department of Veterans Affairs Health Services Research & Development (grant numbers: CDA-08-020, IIR 11-292, IIR 13-052, and IIR 13-073), the Edward Mallinckrodt, Jr. Foundation, and the Vanderbilt Center for Kidney Disease.

I am grateful for the consistent and enthusiastic support of my three committee members, Dr. Michael Matheny, Dr. Tom Lasko, and Dr. Guanhua Chen. As my primary advisor, Dr. Matheny challenged, encouraged, and mentored me throughout the past three years with generosity and patience. Drs. Lasko and Chen were active and engaged collaborators, providing insightful guidance and critical perspective throughout the project. Together, my committee members created an environment that encouraged my academic curiosity, facilitated my successful Masters' research, and prepared me to confidently undertake the next phase of my graduate studies.

I also had the pleasure to work with numerous collaborators at both the U.S. Department of Veterans Affairs and Vanderbilt's Department of Biomedical Informatics. Alex Cheng, Dara Eckerle Mize, and Jacob Eichenberger provided crucial support as reviewers for the systematic literature review. Dr. Edward Siew generously offered critical clinical domain knowledge for interpreting the acute kidney injury analyses. Aize Cao offered insight into the intricacies of the VA clinical data warehouse and constructed essential datasets. Jesse Brannen, Dax Westerman, Josh Gieringer, Jason Denton, Vinnie Messina, and Rob Cronin each assisted with access to computational resources to continuously improve the efficiency of the analysis as the project progressed. I greatly appreciate of the assistance and support from each.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

Page

LIST OF ABBREVIATIONS

ACEi – angiotensin converting enzyme inhibitors

AKI – acute kidney injury

AUC – area under the receiver operating characteristics curve

AVR – aortic valve replacement

CABG – coronary artery bypass graft

CCB – calcium channel blockers

CDS – clinical decision support

CI – confidence interval

CMS -- Centers for Medicare and Medicaid Services

COPD – chronic obstructive pulmonary disease

CPB – cardiopulmonary bypass

CPRS – Computerized Patient Record System

CPT – current procedural terminology

e-HPA – electronic health predictive analytics

ECI – estimated calibration index

EHR – electronic health record

GFR – glomerular filtration rate

ICD – international classification of diseases

ICU – intensive care unit

IVF – intravenous fluids

L1 – L-1 penalized logistic regression

L2 – L-2 penalized logistic regression

L1-L2 – L-1/L-2 penalized logistic regression

LR – logistic regression

MAOI – monoamine oxidase inhibitor

MCHC – mean corpuscular hemoglobin concentration

MI – myocardial infarction

NB – naïve Bayes

NN – neural network

NSAID – non-steroidal anti-inflammatory drugs

O:E – observed to expected outcome ratio

OR – odds ratio

RF – random forest

SD – standard deviation

VA – US Department of Veterans Affairs

VI – variable importance

VINCI – VA Informatics and Computing Infrastructure

VistA – Veterans Health Information Systems and Technology Architecture

VISN – Veterans Integrated Service Network

CHAPTER 1

INTRODUCTION AND SPECIFIC AIMS

Risk prediction models are ubiquitous across clinical specialties and domains[1-7] These models estimate patient risk for diagnostic or prognostic outcomes based on clinical, demographic, social, and genomic risk factors.[5, 8-10] These models may support patient and provider decision-making,[11-13] assist in resource allocation,[14] and adjust quality metrics for acuity.[1, 5, 15] Use of predictive analytics for these tasks requires models that consistently deliver high quality predictions. With the increasingly widespread adoption of electronic health records and use of advanced modeling methods,[1, 16-20] the role of clinical prediction models and our understanding of the challenges presented by the incorporation of predictive analytics into clinical care are rapidly evolving. One such challenge is deteriorating model performance, particularly in terms of calibration, as patient populations shift over time.[6, 7, 21-24] In this study, we focus on understanding whether and how modeling methods affect the tendency of model calibration to deteriorate over time, with the goal of informing methodological recommendations, developing automated modeling surveillance frameworks, and future research.

Motivation

Predictive Analytics: An Evolving Role in Clinical Care

Clinical prediction models have been developed for a broad set of clinical outcomes, including models for the prognosis and diagnosis of acute and chronic diseases,[3, 25-28] hospital mortality,[2] response to medical interventions,[29-31] survival after cancer diagnosis,[29, 32] hospital length of stay,[33] and hospital readmission.[4, 34] Applied prospectively, risk prediction models may support medical decision-making, quality benchmarking, and alignment of health systems with the Triple Aim framework by improving patient outcomes, reducing costs, and increasing patient satisfaction.[1, 5, 7, 19, 35] Impact assessments using comparative study designs are necessary to determine whether the implementation of accurate prediction models result in such improvements for patients and healthcare organizations.[5-7, 36, 37] Although such impact assessments are rare,[5-7, 36, 37] the use of prediction models has been linked with reductions in antibiotic use;[11] reductions in readmissions in patients with heart failure;[14] increased prescription of antihypertensive and cholesterol-lowering medications among patients at high risk of cardiovascular disease;[12] increased uptake of chemotherapy among patients with higher predicted benefit;[13] and reduced disability, improved quality of life, and reduced costs among patients with low back pain.[38, 39]

Historically, risk models relied on relatively few pieces of information, simple algorithms, and manual calculation.[18, 35] However, as clinical data warehouses continue to expand with the adoption of electronic health records (EHRs), models are increasingly synthesizing a broad set of demographic, clinical, and, most recently, genetic risk factors.[5, 9, 10, 19] When sufficient sample

sizes are available for model development,[40-43] incorporating a wider array of risk factors that are in fact associated with the outcome of interest can enable modern prediction models to better characterize associations in the data by capturing more information, explaining more variability in the data, and avoiding some degree of omitted variable and misspecification bias.[40] Although additional predictors may not always offer additional explanatory power beyond existing predictors,[9, 40] when such predictors do improve model fit, model users benefit more accurate predictions.

In addition to making available a broader set risk factors, the data warehouses underlying EHRs are increasing the volume of observations (e.g., patients and admissions) available for use in model development and validation. Larger samples sizes support the inclusion of more predictors, allow for more complex modeling approaches, and provide for more stable estimation.[40, 41, 43-45] Some modeling techniques may require over 200 events per predictor for model development, while others fewer than 20.[44] Substantial sample sizes may also be required to ensure sufficient information content for stable model validation, with simulations suggesting at least 10 events per variable be required.[46] For complex models with many predictors and for models of rare outcomes, these recommendations can large sample sizes and necessitate the use of EHR data warehouses with their substantial data volumes.

Thus a new generation of clinical prediction models are being developed that both incorporate a broader set of risk factors and leverage large EHR-based datasets to improve predictive accuracy. These models can provide personalized estimates of risk for individual patients that can be delivered in real-time within the EHR.[1, 5, 18, 19, 21] Through the application of such models, clinical decision support (CDS) is evolving from rule-based to personalized, probability-based tools. These CDS tools support medical decision-making by synthesizing multi-dimensional data into risk estimates that providers can incorporate with their clinical experience.[1, 5] Yet despite these growing opportunities for the development and deployment of advanced clinical prediction models, few clinical prediction models have been deployed to provide real-time risk prediction at the point-of-care.[1, 5, 7, 36] Recognizing the disparity between model development and implementation efforts, Amarasingham et al[1] proposed electronic health predictive analytics (e-HPA) systems. These information systems would access EHR data to develop, validate, and apply prediction models to enhance clinical care through real-time risk prediction. Ideally, such systems would automate much of the necessary data and model management in order to minimize requirements of analytical staffing resources. While technical, policy, and methodological challenges remain, e-HPA systems hold great promise to promote quality care by integrating risk prediction into clinical practice.

Modeling Methods: Increasing Utilization of Machine Learning Models

Risk modeling research has developed in two parallel fields, with biostatisticians tending to focus on classical regression methods and computer scientists and biomedical informaticists tending to advance machine learning methods.[47, 48] Classical statistical regression techniques, such as logistic and Cox regression, have been widely used for clinical prediction.[47-49] These parametric or semiparametric data modeling methods leverage subject matter knowledge to *a priori* select predictors and identify interactions, as well as determine the form of the relationship

between predictors and the outcome.[17, 47, 49] While parametric methods are familiar, provide human-interpretable models, and may be applied to relatively small cohorts, misspecification of predictor effects may bias risk estimates and limit predictive performance.[17, 47] Machine learning methods, on the other hand, use model-free algorithm-based approaches to learn relationships between variables and develop predictions without the need to pre-specify the form of predictor-outcome associations or interactions between predictors. These computer science motivated modeling approaches may provide improved predictive accuracy by leveraging information within complex associations that are difficult to include in parametric models.[16, 47, 49] However, while avoiding the problem of pre-specification, machine learning methods require careful decisions regarding model structure, potentially complicated tuning of multiple, are computationally complex, require large sample sizes (potentially 10x more events per variable than some regression models[44]), and are not directly interpretable (i.e., "black boxes").[17, 47, 50] The computational and sample size limitations of machine learning methods can be readily overcome with modern computing resources and the availability of modeling cohorts from large EHR-based data warehouses. Similarly, while health care providers may raise concerns over predictions provided without explanatory information (such as that afforded by effect estimates from regression models), new methods such as such as locally interpretable model-agnostic explanations (LIME)[51] are enabling the interpretation of machine learning models.[51-55]

Overall, both regression and machine learning techniques have benefits and limitations for their use in clinical prediction modeling (see Table 1). Although machine learning approaches often report improved performance compared to model-based regression methods,[56-58] both approaches may produce valid predictions[47, 48, 59] and machine learning

Table 1. General advantages and limitations of regression and machine learning models

Issue	Regression	Machine Learning
Model tuning	Not required (generally)[a]	Required, but complexity varies by method
Pre-specification of effects	Required	Not required
Complex interactions	Only included if pre-specified	Automatically incorporated, including high order interactions
Sample size requirements	Relatively small sample sizes required	Data hungry, require larger sample sizes
Computational complexity	Simple, fast	Complex, possibly time-consuming
Familiarity	Familiar	Less familiar, although increasingly common
Interpretability	Easily interpreted coefficients characterizing associations	Not innately interpretable, some additional steps may allow interpretation

[a] Penalized regressions require some degree of tuning

methods do not provide superior performance in all cases.[60-64] Thus, no single approach will provide superior performance and utility in all circumstances.[47, 52] EHR-integrated predictive analytics enables the use of both regression and machine learning approaches by both providing high-dimensional datasets for model development and automating calculation of risk predictions from complex models.[1, 16-19] As both regression and machine learning methods may provide valid, clinically useful risk estimates, model developers must weigh the advantages and limitations of available approaches when determining which technique is most appropriate for a given risk prediction problem.[47] Thus, recommendations for the development and implementation of e-HPA systems must address models constructed through either approach and provide guidance on how modeling methods may impact system design.

Model Performance: The Importance of Calibration and Concern of Performance Drift

There are two aspects of model performance – discrimination (i.e., the ability to separate populations with and without the outcome or to correctly rank-order observations by risk) and calibration (i.e., the agreement between individual predicted and true probabilities).[65] While discrimination focuses on how commonly models assign higher probabilities to observations with the outcome than observations without the outcome, it does not consider whether those probabilities are well-aligned with observed outcome rates (i.e., calibrated). Thus, a model may perform well based on discrimination measures, while suffering substantial miscalibration. For example, predictions of 10% and 50% for a patient who does not go on to experience an outcome and one that does, respectively, are discriminative but not well-calibrated if the observed outcome rates among similar patients are 1% and 5%, respectively. Although both discrimination and calibration are important facets of model performance, they may not be equally important in all contexts.[23, 66] Prediction models with inferior calibration but high discriminatory ability may be sufficient for applications aiming to dichotomize patients into high and low riskgroups. However, model calibration is at least as important as discrimination when individual risk predictions are needed, as is frequently the case in risk-adjusted quality profiling and some types of clinical decision support tools. At the bedside, using predictions for patient-level decision-making requires well-calibrated models that provide individualized predicted probabilities of an outcome that are well-aligned with the true probability the patient will experience the outcome.[1, 2, 7, 21, 23, 66]

Use cases presenting personalized predicted probabilities in support of decision-making depend critically upon model exhibiting and maintaining high levels of calibration.[1, 2, 7, 21, 23, 66] Misleading patient-level risk estimates produced by miscalibrated models may lead to over-confidence, inappropriately alter treatment choices, or misappropriate resources.[1, 23, 67, 68] Although current recommendations emphasize the importance of calibration,[22, 65, 67, 68] validation studies often focus on discrimination and neglect to report calibration.[32, 69, 70] Recent advances in methods for characterizing model calibration further emphasize the clinical importance of aligning predicted probabilities with true risk across the range of patient risk.[71, 72] Decision-analysis of models adhering to stringent measures of calibration have shown such models to have a net benefit greater than or equal to default treat-all or treat-none approaches. Less stringently calibrated models did not provide the same assurance. Thus, strict calibration

assessments can ensure a model will not be harmful to decision-making, although such well-calibrated models are not necessarily helpful either.[71]

In addition to requiring special emphasis on calibration, the implementation of probability-based CDS tools is complicated by the tendency of model performance, particularly in terms to calibration, to drift over time.[5, 6, 21-23, 66] Performance drift results from differences that arise over time between the population on which the model was developed and the population on which the model is applied. Referred to as data shift, changes in the population may take the form of shifts in the underlying outcome rate, patient case mix, or associations between predictors and outcomes.[6, 7, 23, 73] These shifts can be the result of changes in a health care system's patient population, treatment and diagnosis patterns, or variable measurement methods.[5, 6, 21] As clinical prediction models become incorporated into CDS tools and e-HPA systems, understanding how model accuracy changes over time is essential for interpreting risk estimates and developing and maintaining user confidence.

Model Updating: Limited Guidance for Efficient Recalibration

Inadequate performance of clinical prediction models commonly prompts researchers to develop entirely new models.[7, 21] As a result, many prediction models are published for the same outcome.[6] There are over 80 and 100 models of prognosis after stroke and neurological trauma, respectively.[7] Numerous competing models complicate broad implementation and impact assessment of prediction models. Additionally, this approach neglects information from previous model development efforts and commonly utilizes smaller datasets than the original model.[6, 7, 21] Model updating, on the other hand, preserves and extends knowledge by incorporating new data into an existing model.[7, 21] Updating methods vary from intercept correction to adjustment of coefficients and inclusion of new predictors.[6, 7, 21] These methods vary in complexity, data requirements, and analytical staffing resource demand. The frequency of new model development rather than model updating may be partially due to a lack of clear recommendations regarding model updating, including selection between competing updating approaches under varying circumstances and modeling frameworks.[1, 21, 22]

In order for personalized predicted probabilities to be useful in clinical care, e-HPA systems must provide accurate, reliable risk estimates. e-HPA systems must thus enable model updating in response to data shifts that result performance drift beyond acceptable levels. Current model updating protocols often call for regularly scheduled model revision on an annual or biannual basis, with little or no attention to model performance between scheduled maintenance periods. Despite increasing use of machine learning methods, there may not be sufficient evidence regarding whether models developed using regression and machine learning methods are differentially susceptible or robust to various forms of data shift.[74] A deeper understanding the impact of data shifts on model performance and whether modeling methods affect this relationship would support the development of efficient and effective model maintenance components within e-HPA systems.

Central Objective and Specific Aims

Our central objective is to support the adoption of clinical risk prediction and e-HPA systems by advancing research into how performance drift, particularly in terms of calibration, is influenced by modeling methods. The success and broad adoption of probability-based CDS and e-HPA systems will require delivery of consistently high quality risk predictions. Acknowledging the continued importance of both regression and machine learning approaches in clinical prediction, we seek to understand whether modeling methods impact the long-term accuracy and consistency of clinical prediction models implemented in evolving clinical environments and changing patient populations. Our findings may advise e-HPA developers and managers as they select modeling approaches and plan model updating strategies tailored to each model's strengths and limitations. This understanding will lay the ground work for the design of automated EHR-embedded model performance surveillance tools that support sustained model performance and impact on patient and health system outcomes. We will pursue the following specific aims.

Aim 1: Characterize existing knowledge of performance drift in clinical settings

We will conduct a literature review to characterize the state of understanding surrounding prediction model performance over time for models based on regression and machine learning techniques. With particular emphasis on temporal calibration drift, this review will synthesize the available evidence and highlight gaps in the knowledge needed to inform model updating recommendations.

Aim 2: Compare *temporal performance of prediction models for clinical outcomes using common regression and machine learning methods*

We will model two binary clinical outcomes using ordinary logistic regression, penalized logistic regression (i.e., L-1, L-2, and L-1/L-2 penalized regression), and common machine learning methods (i.e., random forests, neural networks, and naïve Bayes). We will assess each model's accuracy, discrimination, and calibration over the seven to nine years after model development, comparing temporal performance trajectories across methods.

Aim 3: Link temporal shifts in patient populations with performance drift to identify drivers of performance drift across models

To inform modeling and recalibration best practices in this domain, we will study how patient and hospital-level characteristics shift over time to influence model performance. We will characterize the forms and extents of data shifts occurring in the same patient populations on which we observed temporal model performance under Specific Aim 2. Using multiple approaches, we will explore event rate, case mix, and predictor-outcome association shifts,

linking each with patterns in model accuracy, discrimination, and calibration to compare the susceptibility of modeling methods to each form of data shift.

While we expect all models to experience performance drift in the presence of substantial data shifts in the patient population, in cases of less dramatic data shifts, modeling methods can be expected to react to variable degrees. As machine learning methods may more fully characterize the complex relationships within clinical data than regression methods by capturing flexible associations and complex interactions,[16, 17, 47, 49] these methods may be less susceptible to performance drift and certain forms of data shift. We anticipate that models with intercepts will immediately systematically over or underpredict as prevalence of an outcome changes in the population (i.e., event rate shift occurs). However, since random forest models do not rely on an initial intercept but instead generate a prediction based on the data in leaf nodes particular to the observation of interest, prevalence changes at the population-level may not impact the accuracy of all observations to the same degree and thereby retain model performance under some degree of event rate shift. As regression models require pre-specified associations, these models likely begin with some level of misspecified associations, so we would may expect shifting predictor-outcome associations to exacerbate these errors and regression models to be particularly susceptible to this form of data shift. Similarly, oversimplified predictor effects or omitted key interactions in regression models may result in clusters of patients for whom these models do not perform especially well, and case mix shift may have a larger impact on their performance compared to machine learning models which are able to learn complex associations specific to many patient clusters. Therefore, we hypothesize that models based on machine learning techniques will retain performance or experience a smaller magnitude of deterioration in performance compared to models based on regression technique. Further, we hypothesize that differences in performance drift by modeling method will result from each form of data shift, with event rate having a smaller influence on random forest models, and machine learning methods in general being less susceptible than regression methods to case mix and predictor-outcome association shifts.

CHAPTER 2

LITERATURE REVIEW

In order to characterize the current evidence, we conducted a systematic review of the literature documenting the performance of clinical prediction models over time, with an emphasis on model calibration. Our aim was to characterize observed patterns in temporal calibration of clinical prediction models; evaluate the types and extents of data shift documented in the literature and its impact on model performance; and synthesize evidence regarding whether different modeling methods are equally robust or susceptible to data shifts.

Search Strategy and Evidence Assessment Methods

Data Sources and Searches

Using both MeSH terms and keywords, we identified potentially relevant articles in the Embase 1974 through May 20, 2015 and MEDLINE In-Process & Other Non-Indexed Citations through May 22, 2015 databases. Our search strategy captured all citations with key concepts in the title or abstract, including terms describing prediction modeling, modeling methods, model performance, and temporal validation. We required all citations to mention calibration or recalibration in the title or abstract, as model calibration is particularly important for clinical risk estimation. We subsequently excluded animal studies and non-English language publications, as well as undesirable publication formats (e.g., case reports and commentaries) and studies in domains outside the scope of this review that were frequently captured in our inclusion searches (e.g., environmental monitoring and analytical chemistry). A detailed description our search protocol is provided in Table 2. We also reviewed the reference lists of eligible studies for relevant studies.

Study Selection

We included original research studies that conducted repeated temporal validations of clinical risk prediction models within a single study cohort. We required studies to include at least 500 patients or cases presenting over at least 2 years. For studies developing a new prediction model, we required the training and validation datasets to be temporally split and applied the sample size and timeframe restrictions to the validation dataset only. As we are interested in model performance over time, we included articles assessing model validation in at least two non-overlapping time periods (e.g., annually or quarterly). Studies reporting only discrimination in each temporal validation cohort were excluded, and thus all included studies reported model calibration or recalibrated model coefficients for sequential validation cohorts.

We did not limit our search by clinical domain, modeling methods, or particular discrimination and calibration metrics.

Two reviewers independently assessed the abstracts of all publications identified from Embase and MEDLINE for eligibility. We retrieved full-text articles for publications meeting the inclusion/exclusion criteria based on either reviewer's determination. Two independent reviewers evaluated each full text article against the inclusion/exclusion criteria and discussed any disagreements until consensus was reached.

Table 2. Detailed literature review search strings

Query number	Search string
1	exp Multivariate Analysis/ or exp Models, Statistical/ or exp Data Mining/ or exp Survival Analysis/ or exp Regression Analysis/ or exp Artificial Intelligence/ or exp "Neural Networks (Computer)"/ or exp Support Vector Machines/ or multivariate.ti,ab. or univariate.ti,ab. or logistic.ti,ab. or regression.ti,ab. or "machine learning".ti,ab. or "statistical learning".ti,ab. or "supervised learning".ti,ab. or "random forest*".ti,ab. or "naive bayes".ti,ab. or "support vector machine*".ti,ab. or "artificial neural network*".ti,ab. or "decision tree*".ti,ab. or "prognostic model*".ti,ab or "predictive analy*".ti,ab. or prediction.ti,ab. or "decision curve*".ti,ab. or "nomogram*".ti,ab.
2	exp Prognosis/ or "prognostic".ti,ab. or "prediction rule*".ti,ab. or "clinical prediction*".ti,ab. or "predicti* factor*".ti,ab. or "clinical predictor*".ti,ab. or predictive.ti,ab. or "risk prediction*".ti,ab. or discrimination.ti,ab. or calibrat*.ti,ab. or recalibrat*.ti,ab. or "model updating".ti,ab. or validation.ti,ab. or "case mix".ti,ab. or "case-mix".ti,ab.
3	"calibrat*".ti,ab. or "recalibrat*".ti,ab.
4	prospective.ti,ab or "external* validat*".ti,ab or temporal.ti,ab or "repeated validation".ti,ab or "repeated recalibration".ti,ab or updating.ti,ab or "calibration drift".ti,ab or "local calibration".ti,ab or "over time".ti,ab.
5	1 and 2 and 3 and 4
6	exp animal/
7	exp human/
8	6 not 7
9	"Biophysics, Bioengineering and Medical Instrumentation".ec. or chemistry.sh. or chemometric.ti,ab. or "instrumentation".sh. or "environmental monitoring".sh. or "Clinical and Experimental Biochemistry".ec. or "land use".sh.
10	(case reports or letter or comment or editorial or practice guideline or historical article or news or newspaper article or legal cases).pt.
11	8 or 9 or 10
12	5 not 11
13	12 and English.la

Data Extraction and Quality Assessment

We extracted key information on the clinical population, modeling and validation approaches, and temporal validation results. Information was extracted from studies included in the review and any associated publications describing original model development. For both the derivation and temporal validation cohorts, we collected information on sample size, clinical setting, geographic setting, timeframe of enrollment, variable definitions, patient population characteristics, and inclusion/exclusion criteria. In terms of methodology, we extracted information on variable selection, modeling approach, internal validation strategy, and model updating techniques. All repeatedly measured discrimination metrics, calibration metrics, and recalibrated model coefficients were recorded. In addition, we recorded original performance metrics for each model. When key information was presented graphically, we requested more detailed information from study authors and estimated values from figures if the authors were unable to provide supplemental information.

Comparing the performance of a prediction model in different populations can be complicated by differences in data definitions. Since disparate definitions may impact model performance systematically or interact with data shifts, any discrepancies between the development cohorts and the temporal validation studies included in our review may limit our ability to interpret linkages between data and performance drift over time. We therefore focus our quality assessment of each temporal validation study based on how well variable definitions and inclusions/exclusion criteria coordinated with those used to define the derivation cohort.

Data Synthesis and Analysis

We synthesized findings across studies graphically and narratively. For each study, we calculated the time elapsed from the end of the enrollment period in the models' development or local-update cohort to the end of the enrollment period in each validation time step. This provided a common basis for assessing temporal performance across studies.

We focused our graphical analysis on performance metrics reported in the majority of studies. Discrimination assesses the ability of a model to distinguish events from non-events or correctly rank-order observations by risk. We focused our assessment of temporal model discrimination on the area under the receiver operating characteristics curve (AUC). AUC provides an assessment of the probability that an observation with the outcome is assigned a higher risk estimate than an observation without the outcome. AUC ranges from 0 to 1, with 0.5 indicating an uninformative model and 1.0 indicating perfect discrimination.[75] Calibration, the agreement between observed and predicted probabilities, is summarized by the observed to expected ratio (O:E), a generalization of the standardized mortality ratio constructed by comparing the overall outcome rate with the mean predicted probability. O:E ratio values of 1 indicate good calibration, with perfect agreement between observed probability and mean predicted probability. Values less than 1 indicate overprediction of risk and values greater than 1 indicate underprediction.[24] Where available, we plotted O:E ratios for both original and recalibrated models to assess the effect of repeated model updating.

We narratively evaluated the connection between data shift and temporal model performance as most studies did not provide detailed information characterizing the clinical population at each validation time step.

Review Results

Selected Literature

Our search returned 1,671 references, with 35 additional references extracted from references of eligible studies. Figure 1 details each references' disposition. Excluding 173 duplicates, 1,534 references were available for abstract review, during which 879 were excluded. We reviewed full texts for the remaining 655 references, of which 16 met all inclusion criteria. An overview of the 16 studies eligible for review is presented in Table 3.

Figure 1. Flow chart of article disposition

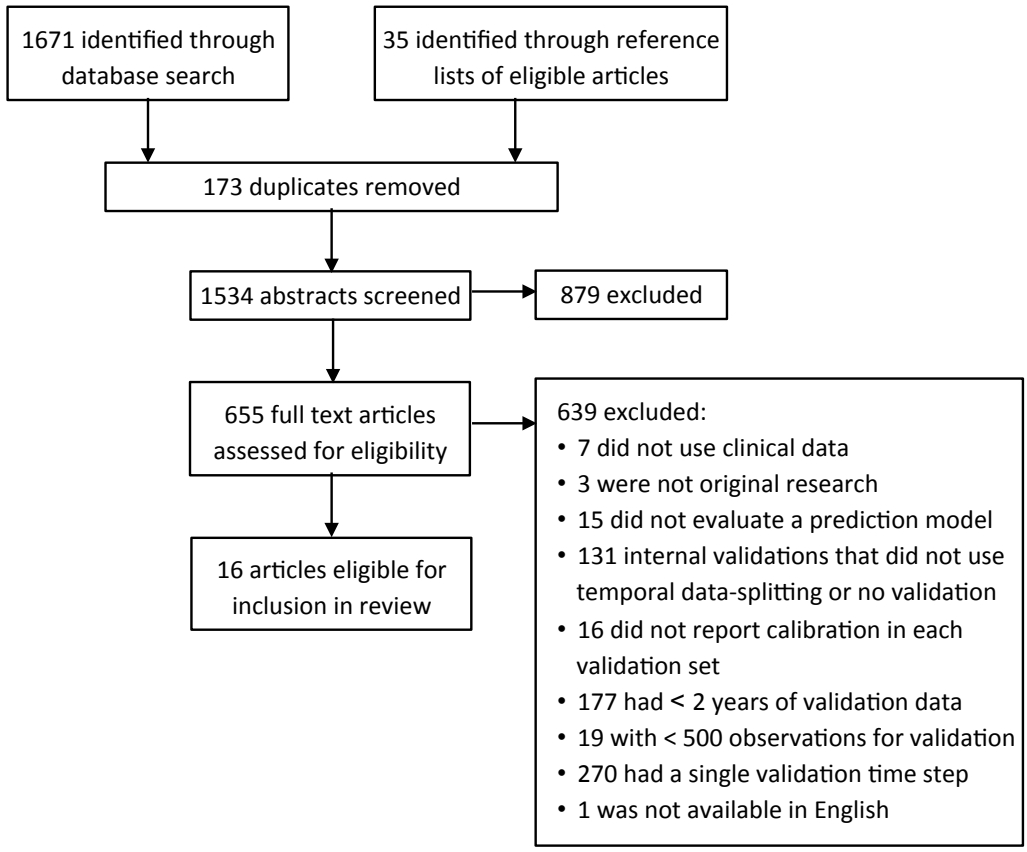


Table 3. Characteristics of studies selected for literature review

			Temporal validation sample					
Study	**Model**	**Local updating**	**Outcome**	**Setting**	**Timeframe**	**N**	**Time since derivation/ update**	**Temporal splitting**
Cook (2002) [76]	APACHE-III [77]	Proprietary adjustments	Hospital mortality among ICU admissions	Brisbane, Queensland, Australia; 1 hospital	Jan. 1, 1995 to Jan. 1, 2000	5,278	5.2 - 10.2 yrs *	1-yr periods
Harrison (2006) [78]	APACHE-III; APACHE-II [79]; SAPS-II [80]; MPM-II [81]	Refit on data from Dec. 1995 – Dec. 1999 (n=65,427)	Hospital mortality among ICU admissions	England, Wales, and Northern Ireland; 163 ICUs	Jan. 2000 to Aug. 2003	75,679	0 – 3.6 yrs	Two 1-yr periods and one 1.6-yr period
Harrison (2014) [82]	APACHE-II [79]	n/a	Hospital mortality among ICU admissions	Scotland; 24 ICUs	Jan. 1, 2007 to Dec. 31, 2009	22,700	25 – 27 yrs	1-yr periods
Hekmat (2005) [83]	APACHE-II; MODS [84]; CASUS [83]	n/a	30-day mortality among ICU admissions of cardiac surgery patients undergoing CPB	Cologne, Germany; 1 ICU	APACHE-II and MODS: Apr. 1999 to May 2001; CASUS: May 2000 to May 2001 and Feb. 2002 to Feb. 2003	APACHE-II and MODS: 1,441; CASUS: 2,161	APACHE-II: 16.3 - 18.3 yrs; MODS: 10.1 - 12.2 yrs; CASUS: 0 - 2.7 yrs	1-yr periods
Hekmat (2010) [85]	APACHE-II; MODS; † CASUS	n/a	30-day mortality among ICU admissions of cardiac surgery patients undergoing CPB	Cologne, Germany; 1 ICU	CASUS: May 2000 to May 2001 and Feb. 2002 to Oct. 2005	CASUS: 4,858	CASUS: 0 - 5.4 yrs	CASUS: 1-yr period and 3.5-yr period
Hickey (2013a) [86]	Logistic EuroSCORE [87]	Multiple repeated updating methods	Hospital mortality after cardiac surgery	England and Wales; 37 hospitals	Apr.1, 2001 to Mar. 31, 2011	316,632	5.3 - 15.3 yrs	Varying

Table 3. (continued) Characteristics of studies selected for literature review

			Temporal validation sample					
Study	**Model**	**Local updating**	**Outcome**	**Setting**	**Timeframe**	**N**	**Time since derivation/ update**	**Temporal splitting**
Hickey (2013b) [23]	Logistic EuroSCORE	n/a	Hospital mortality after cardiac surgery	England and Wales; 37 hospitals	Apr. 1, 2001 – Mar. 31, 2011	317,292	5.3 - 15.3 yrs	3-mth periods
Madan (2011) [88]	study-specific model	n/a	Mortality within 30 days of surgery or within the same hospital admission after cardiac surgery with CPB	Houston, Texas, United States (single center)	Jan. 1, 2000 to Aug. 27, 2007	7,160	0 - 7.6 yrs	One 5-yr period and one 2.6-yr period
McCormick (2012) [89]	study-specific model	Dynamic logistic regression and dynamic model averaging	Laparoscopic appendectomy among pediatric patients	United States; 2,449 hospitals	1996-2002	72,189	No elapsed time	1-mth periods
Mikkelsen (2012) [90]	Logistic EuroSCORE	n/a	30-day mortality after cardiac surgery	Denmark; 4 cardiac centers	Jan. 1, 1999 to Mar. 31, 2010	21,664	3.1 - 14.3 yrs	2-yr periods
Minne (2012a) [24]	rSAPS-II (logistic) [91]	Original and updated each time step ‡	Hospital mortality among elderly ICU admissions	The Netherlands; 21 ICUs	Jan. 2004 to July 2009	12,143	0 - 5.6 yrs	30 time steps of equal sample size
Minne (2012b) [92]	rSAPS-II (tree) [91]	Original and updated each time step ‡	Hospital mortality among elderly ICU admissions	The Netherlands	Jan. 2004 to July 2009	12,143	0 - 5.6 yrs	30 time steps of equal sample size

Table 3. (continued) Characteristics of studies selected for literature review

			Temporal validation sample					
Study	**Model**	**Local updating**	**Outcome**	**Setting**	**Timeframe**	**N**	**Time since derivation/ update**	**Temporal splitting**
Osswald (2009) [93]	Logistic and additive EuroSCORE [87, 94]	n/a	30-day mortality after aortic valve replacement	Heidelberg, Germany; 1 hospital	Jan. 1, 1994 to Mar. 31, 2006	1,545	1.9 - 10.3 yrs	4-yr periods
Paul (2012) [95]	APACHE-III (version j)	n/a	Hospital mortality among ICU admissions	Austrialia and New Zealand	Jan. 1, 2000 to Dec. 31, 2009	558,585	10.2 - 20.2 yrs *	2-yr periods
Rogers (2012) [96]	TRISS [97]	n/a	Hospital mortality after admission with traumatic injury	Pennsylvania, United States	1990-2010	408,489	2 - 23 yrs	1-yr periods
Siregar (2014) [§ 46]	Logistic EuroSCORE	Multiple repeated updating methods	Hospital mortality after cardiac surgery	The Netherlands	2007-2012	95,240	11 - 17 yrs	Varying time windows

***** Time since end of APACHE-III development dataset since timing of updating dataset unknown.

[†] APACHE-II and MODS validations are identical to those reported in Hekmat (2005)

[‡] First-level recalibration of logistic models involves fitting a new logistic model with original model's log odds as the only predictor. First-level recalibration for the tree-based model utilized the same tree structure, but adjusted the mortality prediction at each leaf based on outcomes of all patients in each leaf in the updating cohort.

[§] Abstract only

Abbreviations: CPB – cardiopulmonary bypass; ICU – intensive care unit

Mortality after cardiac surgery or ICU admission were the most common clinical outcomes predicted, considered in eight and six studies, respectively. Mortality after traumatic injury and type of appendectomy were the outcome in one study each. Two studies developed and temporally validated new models; 13 studies temporally validated an existing model or locally-updated version of an existing model (one study took both approaches). Derivation details for the original models are included in Table 4, and local model updating, if applicable, is noted in Table 3. Four studies reported the performance of multiple models within the same validation dataset.

We identified studies in European, North American, and Australian hospitals. Five studies (two of which were related) collected data at a single institution, whereas the remaining studies leveraged regional, national, or international datasets. The smallest dataset included 1,441 observations and the largest 558,585.

We identified several related studies. Two articles by Minne et al[24, 92] utilized the same validation cohort to explore temporal performance of parallel models based on different development methods. Two articles from Hickey et al[23, 86] relied on the same dataset to support different approaches to understanding temporal performance. Overlapping validations of the APACHE-II and MODS models were presented in two articles by Hekmat et al.[83, 85] We thus considered the results presented in their 2005 article for these models, and extracted results for the CASUS model from both studies.

Quality Assessment

In addition to the temporal validation details presented in Table 3, model development details are presented in Table 4. Four studies evaluated temporal performance in cohorts directly linked to the model development data. In these cases, data definitions (both outcome and covariate), missing value processing, and inclusion/exclusion criteria were identical in derivation and validation cohorts. We assessed the remaining eleven studies for quality based on coordination of variable definitions and inclusion/exclusion criteria with those used to define the derivation cohort.

Although few studies provided detailed definitions of model covariates, data collection methods suggested definitions corresponded with those used during model development. Eight studies utilized registry data on which the relevant published risk scores are routinely calculated. The four studies using study-specific data to validate existing models noted adherence to variable definitions reported in the original publication. Consistent with methods used for model development, missing covariate data was generally assumed to be normal. Rogers et al [96] utilized complete case analysis; we were unable to determine whether this strategy was used for TRISS development.

The outcome definitions in our reviewed studies were not always consistent with those used for model development. Two studies applied original outcome definitions. Hekmat et al [83, 85] assessed the ability of the APACHE-II and MODS models to predict 30-day mortality among ICU admissions after cardiac surgery, whereas the models were originally developed for hospital and ICU mortality, respectively, without the 30-day limit.[79, 84] The EuroSCORE predicts mortality within 30-days post-cardiac surgery or within the same hospital stay.[94] In their

Table 4. Modeling methods and derivation/internal validation cohort characteristics for risk prediction models assessed in studies included in literature review

						Development methods		
Model	**Original article**	**Outcome**	**Setting**	**Timeframe**	**N**	**Model**	**Variable selection**	**Internal validation**
APACHE-II	Knaus et al. 1985 [79]	Hospital mortality among ICU admissions	United States; 19 ICUs from 13 hospitals	1982 (one hospital 1979-1981)	5030	Logistic regression	Domain knowledge	No formal validation
APACHE-III	Knaus et al. 1991 [77]	Hospital mortality among ICU admissions	United States; 42 ICUs from 40 hospitals	May 1988 to November 1989	15680 (7840/ 7840)	Logistic regression	Domain knowledge	Random 50:50 data splitting
CASUS	Hekmat et al. 2005 [83]	30-day mortality after cardiac surgery with CPB	Cologne, Germany; 1 ICU	April 1999 to April 2000	384	Logistic regression	Filtering based on univariate tests	Temporal data-splitting[a]
EuroSCORE (additive and logistic versions)	Nashef et al. 1999 [94] and Roques et al. 2003 [87]	Mortality within 30 days of surgery or within the same hospital stay after cardiac surgery with CPB	8 European countries; 128 medical centers	September-November 1995	14799 (13302 / 1497)	Logistic regression	Filtering based on univariate tests followed by backward selection	Random 90:10 data-splitting
MODS	Marshall et al 1995 [84]	ICU mortality	Halifax, NS, Canada; 1 ICU	May 1988 to February 1990	692 (336 / 356)	Logistic regression	Literature review	Temporal data-splitting (internal validation: March 1989-February 1990)

Table 4. (continued) Modeling methods and derivation/internal validation cohort characteristics for risk prediction models assessed in studies included in literature review

						Development methods		
Model	**Original article**	**Outcome**	**Setting**	**Timeframe**	**N**	**Model**	**Variable selection**	**Internal validation**
MPM-II	Lemeshow et al. 1993 [81]	Hospital mortality among ICU admissions	Europe, North America; 143 ICUs	Dataset I: April 17, 1989 to May 10, 1991; Dataset II: September 30, 1991 to December 27, 1991	19124 (12610 / 6514)	Logistic regression	Filtering based on univariate tests followed by backward selection	Temporal data-splitting for dataset I (September 1990-May 1991 reserved for validation) and random 65:35 data-splitting for dataset II
rSAPS-II - logistic	de Rooji et al. 2007 [91]	Hospital mortality among eldery ICU admissions	The Netherlands; 21 hosptials	January 1997 to December 2003	6867 (4578 / 2289)	Logistic regression (refit SAPS-II model in elderly subpopulation)	All SAPS-II variables included	Random 66:33 data-splitting
rSAPS-II - tree	de Rooji et al. 2007 [91]	Hospital mortality among eldery ICU admissions	The Netherlands; 21 hosptials	January 1997 to December 2003	6867 (4578 / 2289)	Recursive partitioning analysis	All SAPS-II variables included; Tree pruned with 10-fold cross-validation	Random 66:33 data-splitting
SAPS-II	Le Gall et al. 1993 [80]	Hospital mortality among ICU admissions	Europe, North America; 137 ICUs	September 30, 1991 to December 27, 1991	12997 (8369 / 4628)	Logistic regression	Filtering based on univariate tests	Random 65:35 data-splitting

Table 4. (continued) Modeling methods and derivation/internal validation cohort characteristics for risk prediction models assessed in studies included in literature review

						Development methods		
Model	**Original article**	**Outcome**	**Setting**	**Timeframe**	**N**	**Model**	**Variable selection**	**Internal validation**
TRISS	Champion et al. 1995 [97]	Hospital mortality after traumatic injury	United States and Canada; 51 institutions	October 1982-1987	23177	Logistic regression	unknown	unknown
Madan 2011 - study-specific model	Madan et al. 2011 [88]	Mortality within 30 days of surgery or within the same hospital stay after cardiac surgery with cardiopulmonary bypass	Houston, Texas, United States (single center)	January 1, 1993 to December 31, 1999	8959	Logistic regression	Forward selection	Temporal splitting *
Model for pediatric appendectomy type	Hagendorf et al. 2007 [98]	Laproscopic rather than open appendectomy among pediatric patients	United States	1996 – 2002	72,189	Logistic regression	Backward selection	N/A

* Temporal splitting for initial model validation included data used for the repeated temporal validation cohorts considered in our review. In these cases, the time frame and sample size of the derivation cohort only is reported here.

Abbreviations: CPB – cardiopulmonary bypass; ICU – intensive care unit

validations of the EuroSCORE, Mikkelsen et al [90] and Osswald et al [93] defined the outcome as mortality within 30 days (eliminating deaths occurring during the hospital stay but more than 30-days after surgery), whilst Hickey et al [23, 86] and Siregar et al [46] defined the outcome as hospital mortality (eliminating deaths occurring within 30-days but after discharge).

Detailed inclusion/exclusion criteria for each study and the development cohorts of existing models are provided in Table 5. Data restrictions generally followed similar patterns to those applied during model development. Osswald et al restricted their analysis to patients receiving primary isolated aortic valve replacement,[93] in contrast to the EuroSCORE which was developed for all cardiac surgeries with cardiopulmonary bypass.[94] Hekmat et al [83, 85] restricted to ICU admissions after cardiac surgery rather to the general ICU admissions on which the APACHE-II and MODS models were developed.[79, 84]

Table 5. Comparison of population exclusion criteria between model development and temporal validation studies in literature review

	Exclusions															
Model/Study	**Cardiac surgery**	**CABG**	**Non-cardiac surgery**	**Non-surgical**	**Coronary care**	**MI observation**	**Burns**	**Transplants**	**Other than AVR**	**Min. age (yrs)**	**Max. age (yrs)**	**Min. ICU stay (hrs)**	**Hospital stay <48hrs**	**Readm. In hospital stay**	**Transfers to other ICUs**	**Other**
APACHE-II [79]		X					X			16		8				
Hekmat (2005, 2010) [83, 85]			X							18						
Harrison (2006) [78]	X	X					X	X		16		4		X	X	Missing outcome, APACHE score, admission type, or ventilation information for PaO2/fraction of inspired oxygen ratio
Harrison (2014) [82]							X			16		4		X		Flagged as "Excluded from severity of illness scoring"; Missing outcome, age, admission reason or location.
APACHE-III [77]		X				X	X	X		16		4				
Cook (2002) [76]	X					X	X			16		4		X		
Harrison (2006) [78]	X	X					X	X		16		4		X	X	Missing outcome, APACHE score, admission type, or ventilation information for PaO2/fraction of inspired oxygen ratio
Paul (2012) [95]										16		4		X	X	Missing APACHE-III score or outcome; CABG cases prior to 2007

Table 5. (continued) Comparison of population exclusion criteria between model development and temporal validation studies in literature review

	Exclusions															
Model/Study	**Cardiac surgery**	**CABG**	**Non-cardiac surgery**	**Non-surgical**	**Coronary care**	**MI observation**	**Burns**	**Transplants**	**Other than AVR**	**Min. age (yrs)**	**Max. age (yrs)**	**Min. ICU stay (hrs)**	**Hospital stay <48hrs**	**Readm. in hospital stay**	**Transfers to other ICUs**	**Other**
CASUS [83]			X							18		24				
Hekmat (2005, 2010) [83, 85]			X							18						
EuroSCORE [94]			X													Excluded at center-level if >99% missing (4 centers excluded)
Hickey (2013a) [86]			X													Missing procedure date, missing outcome, and within admission re-do procedures
Mikkelsen (2012) [90]			X													Missing EuroSCORE
Siregar (2014) [46]			X													
Hickey (2013b) [23]			X				X									Missing procedure date, missing outcome, and within admission re-do procedures. Traumas and ventilator assisted device procedures.
Osswald (2009) [93]			X						X							
MODS [84]												24				
Hekmat (2005, 2010) [83, 85]			X							18						
MPM-II [81]	X				X	X	X			18				X		
Harrison (2006) [78]	X	X					X	X		16		4		X	X	Missing outcome, APACHE score, admission type, or ventilation information for PaO2/fraction of inspired oxygen ratio
rSAPS-II [91]										80						Missing admission type or SAPS-II score
Minne (2012a, 2012b) [24, 92]										80						Missing admission type or SAPS-II score
SAPS-II [80]	X				X	X	X			18						Missing admission type or ventilation information for PaO2/fraction of inspired oxygen ratio
Harrison (2006) [78]	X	X					X	X		16		4		X	X	Missing outcome, APACHE score, admission type, or ventilation information for PaO2/fraction of inspired oxygen ratio

Table 5. (continued) Comparison of population exclusion criteria between model development and temporal validation studies in literature review

	Exclusions															
Model/Study	**Cardiac surgery**	**CABG**	**Non-cardiac surgery**	**Non-surgical**	**Coronary care**	**MI observation**	**Burns**	**Transplants**	**Other than AVR**	**Min. age (yrs)**	**Max. age (yrs)**	**Min. ICU stay (hrs)**	**Hospital stay <48hrs**	**Readm. In hospital stay**	**Transfers to other ICUs**	**Other**
TRISS [97]										15						Other than penetrating or blunt trauma
Rogers (2012) [96]										16			X			Other than penetrating or blunt trauma; Patients who received paralytic agents and were recorded as having a respiratory rate of zero
Madan (2011) [88]	X															
McCormick (2012) [89]											15					Appendectomy patients only.

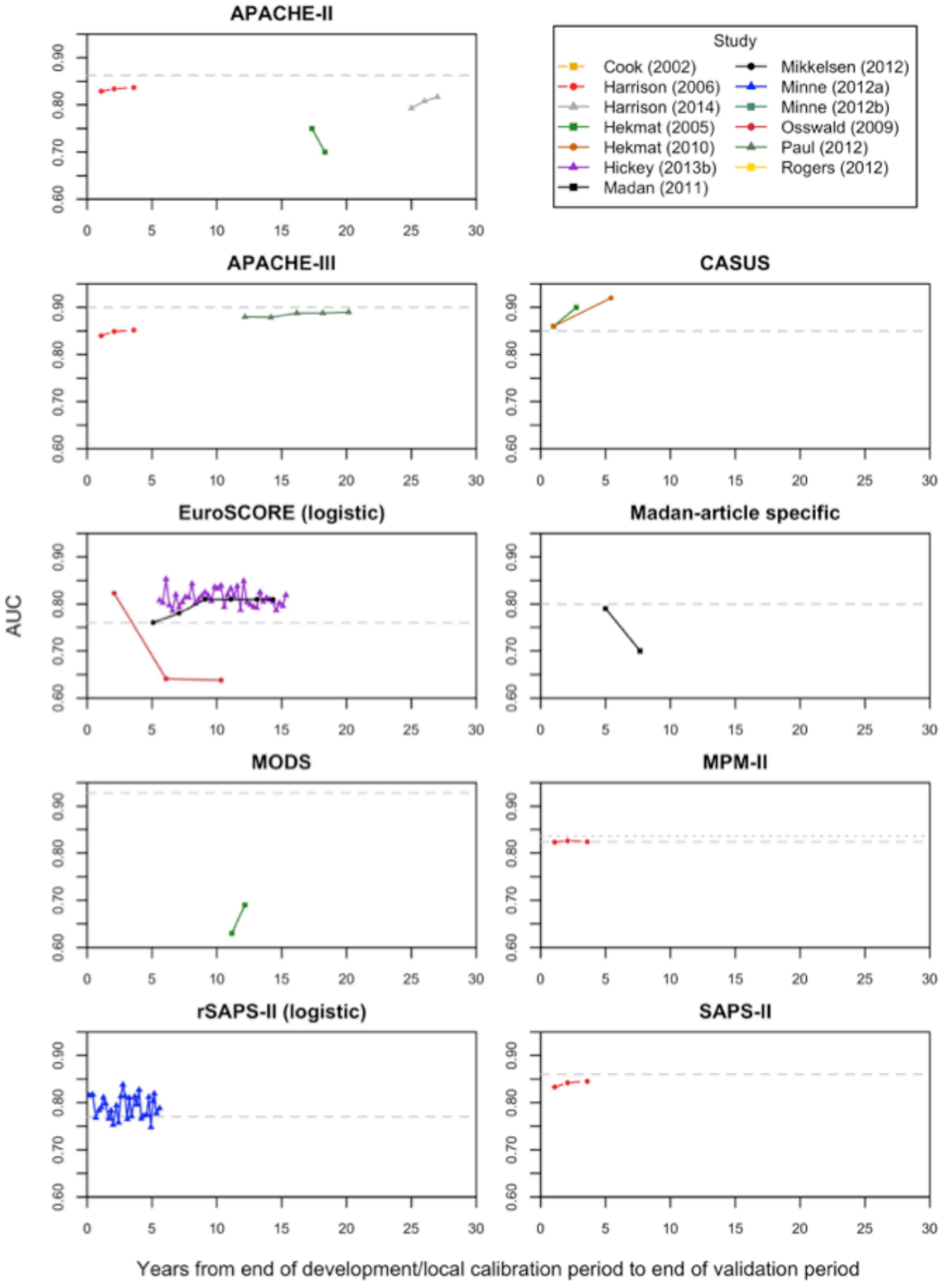
Note: Grey rows for Madan (2011)[88] and McCormick (2012)[89] represent both development and validation studies due to evaluation of study-specific models.

Abbreviations: AVR=aortic valve replacement; CABG=coronary artery bypass graft

Temporal Model Discrimination

Although not the focus of this review, we assessed discrimination when it appeared in our selected studies. Eleven studies reported discrimination, excluding the TRISS and tree-based rSAPS-II models from assessment of temporal discrimination. Models exhibited good to excellent discrimination at development, with AUCs between 0.75 and 0.90. Figure 2 illustrates AUC over time by model and study. Discrimination was maintained over time, with AUCs generally between 0.70 – 0.92. Several temporal validation studies observed AUC values above those reported during model development.[23, 24, 83, 85, 90, 93] The logistic EuroSCORE and rSAPS-II models exhibited improved discrimination from baseline but no temporal pattern across the validation time period in studies by Minne et al [24] or Hickey et al.[23] In a Danish cohort, Mikkelsen et al observed an increase in AUC from 0.76 to 0.81 over the first six years of validation, followed by stability over the next seven years.[90] Small temporal improvements in AUC were reported for APACHE-II by Paul et al,[95] as well as locally-updated versions of APACHE-II, APACHE-III, MPM-II, and SAPS-II by Harrison et al.[78] We observed large declines in discrimination for APACHE-II in Hekmat et al,[83] the logistic EuroSCORE in Osswald et al,[93] and the study-specific model developed by Madan et al.[88] Over two time steps, increases in AUC were observed for CASUS and MODS.[83, 85]

Figure 2. Discrimination over time by reviewed prediction model and study. Models updated as a component of the study prior to temporal validation are distinguished with dashed lines connecting validation time steps. Performance at the time of model development is indicated by horizontal gray reference line.



Temporal Model Calibration

Fourteen studies reported model calibration in multiple, sequential validation cohorts. These cohorts were collected from immediately after the end of the development or update period up to 27 years later, with most studies falling within 10-15 years after model development/update. We were not able to establish the timeframe of the update cohorts for the versions of APACHE-III validated by Cook et al[76] and Paul et al,[95] thus we measured elapsed time from development and acknowledge these times are overestimates. Studies reported model calibration in 2 to 40 time steps ranging in length from 3 months to 5 years.

Twelve studies contained sufficient detail for graphical analysis. The study by Siregar et al[46] has only been published in abstract form and detailed results were not available.

Figure 3 illustrates temporal patterns in calibration for the ten studies reporting O:E ratios. In general, calibration deteriorated over time, with O:E ratios declining, an indication of increasing overprediction. Substantial changes in calibration were reported for the APACHE-III,[76, 95] logistic EuroSCORE,[23, 90] TRISS,[96] and logistic rSAPS-II[24] models. Within 3.5 years after local model updating, Harrison et al[78] observed small declines in calibration for all four models assessed. Minne et al[24] observed degraded calibration of the logistic rSAPS-II model within the same timeframe. Although O:E ratios were not reported, deteriorated calibration was observed for the model developed and validated by Madan et al[88] and for APACHE-II and MODS validated in Hekmat et al's cohort.[83] We encountered some exceptions to this general pattern: the EuroSCORE in the cohort evaluated by Osswald et al[93] and the tree-based rSAPS-II model assessed by Minne et al[24, 92] showed no deterioration.

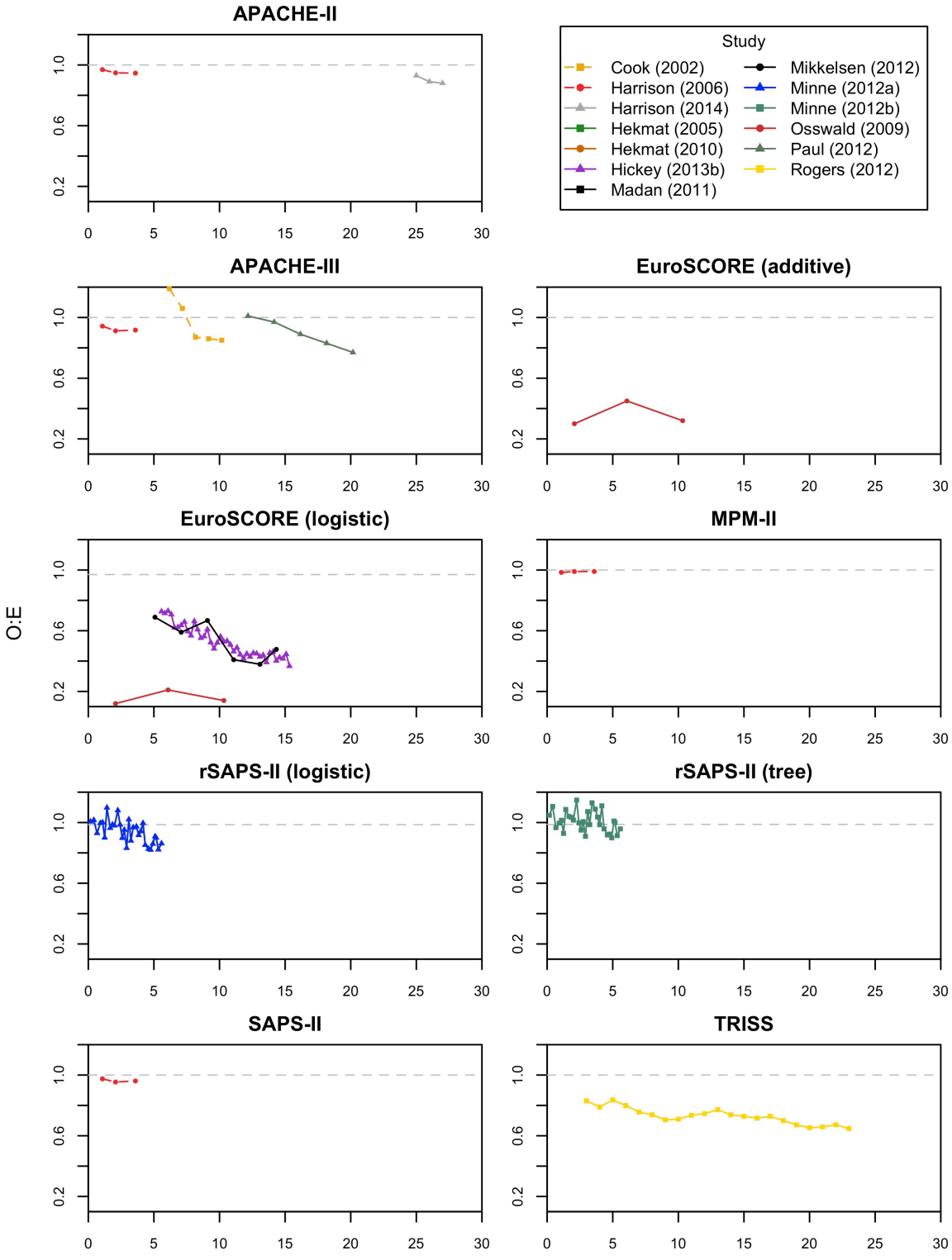
Observed Data Shifts

We evaluated three forms of data shift: outcome rate shift (a change in the baseline rate of the outcome within the population), case mix shift (a change in the distribution of risk factors within the population), and predictor-outcome association shift (a change in relationships between predictors and risk of the outcome). Outcome rate shift was the most commonly assessed form. Four studies documented stable outcome rates, and four documented declining rates. In the cohort evaluated by Hekmat et al, there was no clear pattern in the outcome rate across three temporal cohorts.[83, 85]

Changes in case mix were generally reported narratively without detailed information for each time step. Hickey et al,[23] Mikkelsen et al,[90] and Madan et al[88] observed increasing patient severity, while Osswald et al[93] observed no trends in patient severity. Hickey et al[23] provided the most detailed exploration of case mix shift, observing linear and nonlinear trends in numerous predictor distributions. The authors, however, noted the observed changes in most risk factors were small.[23]

Through repeated model updating and a focus on model coefficients, two studies explored shifts in predictor-outcome associations. McCormick et al[89] and Hickey et al[86] observed complex, co-occurring forms of predictor-outcome association shift. Some associations were stable throughout the study, some increased or decreased in strength, and some exhibited both periods of stability and shifting strength. For still other predictors, association shifts in one

Figure 3. Calibration over time by reviewed prediction model and study. Models updated as a component of the study prior to temporal validation are distinguished with dashed lines connecting validation time steps. O:E ratio at development or the ideal value (1.0) is indicated by horizontal gray reference lines.

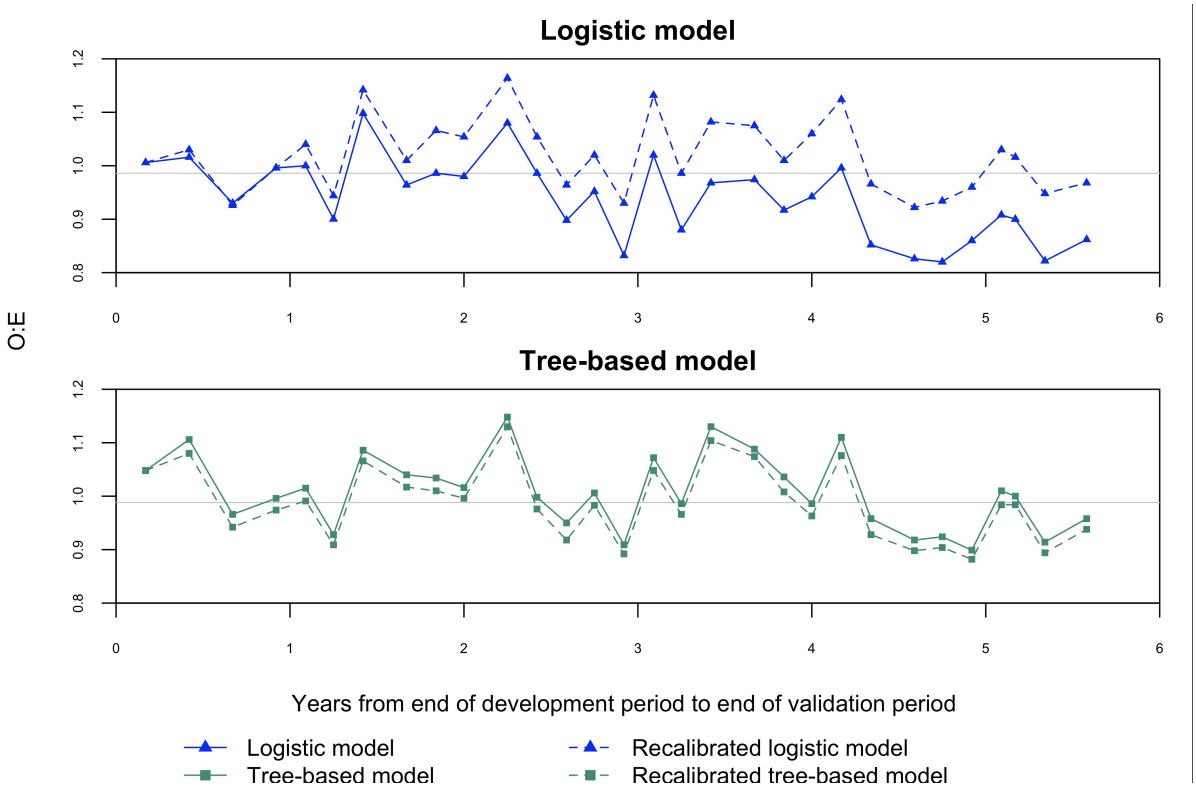


direction were reversed in subsequent years. Associations also drifted in and out of the range of statistical significance.

Susceptibility of Different Modeling Methods to Performance Drift

Two coordinated studies by Minne et al support direct comparison of the temporal performance of a logistic regression model[24] and a recursive partitioning, classification tree model.[92] These models were derived on the same cohort using the same covariates and were subsequently temporally validated on the same dataset. These studies employed statistical process control to determine whether model performance was acceptable at each validation time step. Discrimination was maintained over the 6-year validation period by both models. The tree-based model maintained calibration, but the logistic model began significantly overpredicting risk within four years (See Figure 4). With repeated first-level recalibration after each time step, the O:E ratio for the logistic model was restored to acceptable levels and similar to the performance of the original tree-based model.[24, 92] Repeated recalibration had a small effect on the tree-based model and was not considered necessary.[92]

Figure 4. Temporal calibration of corresponding logistic and tree-based rSAPS-II models (extracted from studies by Minne et al [24, 92]).



Discussion

Incorporating high-dimensional clinical, genomic, and demographic factors, predictive analytics leverage rich EHR data to provide clinical decision support at the point of care. Such applications require well-calibrated models consistently delivering accurate individualized risk estimates. Our review of the literature documenting temporal performance of clinical prediction models provided consistent evidence of deteriorating calibration over time. On the other hand, discrimination was stable for extended periods in our selected studies. Despite flexible screening criteria, we found existing evidence focused almost exclusively on models predicting mortality and developed with logistic regression.

Discrimination was stable over time within study populations, even more than 20 years after model development. Since we limited our search to studies considering calibration, a larger body of literature on temporal patterns of discrimination was excluded from our review. However, two studies evaluating discrimination but not calibration in sequential validation cohorts were captured by our search and reported similar findings.[99, 100] Studies exhibiting temporal instability in AUC[83, 85, 88, 93] relied on substantially smaller datasets than the other studies reviewed. These AUC changes may be related to less stable estimates rather than true differences or to variation in data restrictions between time steps. In several cases, validation AUCs were consistently higher than those reported at model development.[23, 24, 83, 85, 90, 93] Such improvements in discrimination may result from more heterogeneous case mixes in the validation versus development cohorts[22] or an increasing proportion of patients in subgroups with higher discrimination.[78]

Calibration deteriorated over time, usually in the direction of overprediction, and in most cases within five years of model development. Clearly, poor calibration can lead to poor decision-making. Patients may be dissuaded from pursing potentially effective treatments when presented with elevated estimates of complication risk or may elect to undergo difficult treatments when presented with inflated estimates of negative disease prognosis.[23, 67] Increasing miscalibration over time may also adversely influence quality assessments utilizing prediction models to risk-adjust quality metrics for differences in patient mix and severity between hospitals or care units. Minne et al found calibration drift of the rSAPS-II logistic model resulted in overly optimistic quality assessments. Fifteen percent of hospitals were identified as underperforming by the model affected by calibration drift, while 35% of the hospitals were identified as underperforming when the model was recalibrated to correct for overprediction.[24]

Few studies provided detailed assessments of observed data shifts and the impact of data shifts on model performance; however, temporal trends in calibration were attributed to both outcome rate shift and case mix shift. In Cook et al,[76] three years of declining outcome rates followed by a period of stability paralleled trends observed in the O:E ratio which indicated increasing and then stable overprediction. Osswald et al[93] observed no change in patient severity and no trend in O:E ratios, while three studies documented increasing patient severity with corresponding declines in O:E ratios.[23, 88, 90] Combining the two studies by Hickey et al[23, 86] provided the most comprehensive evidence linking temporal model performance and data shifts. Within a cohort spanning 10 years, these studies documented co-occurring outcome rate, case mix, and predictor-outcome association shift. Despite multifaceted data shifts, the EuroSCORE's discrimination was stable. Calibration, on the other hand, decayed across the

entire study period as the EuroSCORE increasingly overpredicted risk.[23] A steadily declining intercept during repeated model updating underscored a declining mortality rate as a key component of data shift affecting calibration.[23, 86] Analysis exploring the contribution of case mix shift to deterioration of calibration indicated numerous and complex trends in predictor distributions over time. The authors noted, however, that observed changes in risk factors were small, and thus declines in O:E ratios were likely either unrelated to case mix or associated with complex interactive effects among risk factors.[23] With repeated model updating, trends in coefficients identified complex fluctuations in predictor-outcome associations, including periods of stability and both monotonic and non-monotonic trends in predictor strength.[86] These studies highlight the robustness of model discrimination; vulnerability of model calibration; and challenges of attributing patterns in model performance to specific components of data shift.

We designed our search strategy and screening criteria to capture broad evidence of calibration drift and data shifts in the clinical literature; however, we found the available evidence in the literature to be limited. Among 16 studies reviewed, 14 assessed mortality prediction models. While these models focused on mortality within different populations (i.e., cardiac surgery patients, ICU admissions, and traumatic injury patients), the limitation of the available literature to mortality prediction restricts generalizability of the evidence. In addition, although we did not design our search to exclude models for continuous, time-to-event, or multinomial outcomes, all 16 studies modeled binary outcomes. Research exploring the performance of prediction models across a variety of clinical domains is needed to allow for generalizable conclusions regarding the impact of temporal model performance and data shifts on clinical prediction tool development and implementation.

There was also little variation in modeling methods among the prediction models evaluated (See Table 4). With the exception of one study, models were fit using logistic regression. Minne et al[24, 92] provided direct comparisons between the temporal performance of a logistic regression and a tree-based version of the rSAPS-II model. In both cases, discrimination was maintained. Within four years, the logistic regression model began overestimating mortality risk at a level deemed significant and unacceptable, whereas the tree-based model maintained calibration across the study period. Minne et al suggest these findings may relate to the dichotomization of predictors in the tree structure, which may require data shifts to cross branching thresholds before affecting predictions.[92] Beyond the rSAPS-II tree-based model, our review found no studies assessing temporal performance of models based on modern techniques that minimize overfitting, handle collinearity, incorporate complex interactions, or automate variable selection, such as penalized regression and flexible machine learning algorithms.

Our systematic review is strengthened by inclusion of studies leveraging large national and international datasets covering extended timeframes, as well as several studies applying identical data definitions and restrictions in development and validation cohorts. The findings are limited by the unexpected restriction to evidence primarily from a single clinical domain (i.e., mortality) and the narrative nature of most evidence regarding case mix shift. In addition, in order to synthesize evidence with metrics common across studies, we limited our analysis of calibration to overall O:E ratios, which may obscure trends in performance within subgroups that may have allowed for more direct links with data shifts.

Conclusions

Risk prediction is ubiquitous in health care, supporting medical decision-making and quality benchmarking. As EHR utilization increases, the next frontier is integrating these data resources to support advanced predictive analytics delivering individualized risk estimates at the point of care. Real-time application of prediction models within EHRs will allow clinical decision support to move from Boolean logic rules to flexible model-based tools incorporating high-dimensional information. This transition will require well-calibrated models that consistently provide estimates of individual risk that are well-aligned with true risk. A key challenge to broad implementation, however, is the deterioration of model calibration over time. Additional research is required to determine whether and how modeling methods exacerbate or alleviate calibration drift under varying data shift scenarios. Such understanding of how model performance may shift over time is essential for interpreting risk estimates, developing and maintaining user confidence, developing guidelines for model updating, and designing efficient systems for routine model maintenance.

CHAPTER 3

STUDY DESIGN

In response to the limited evidence regarding variations in the robustness of modeling methods to performance drift, we explored the temporal performance of seven common regression and machine learning methods for two illustrative clinical outcomes — hospital-acquired acute kidney injury and 30-day all-cause mortality after hospital admission. We assessed calibration at varying levels of stringency to both compare our findings with the existing literature and to extend our assessments to measures particularly relevant for e-HPA clinical decision support systems. Further, we applied multiple methods to assess each form of data shift in order to link changes in patient and hospital-level characteristics to fluctuations in model performance.

Illustrative Clinical Outcomes

This study is aimed at understanding the connection between data shift, modeling methods, and model performance in general, and for this reason outcomes of interest in any clinical domain could potentially serve as exemplar endpoints for this work. We elected to study two clinical use cases with different clinical patient patterns, exposures, and distinct outcomes. We explored performance drift and data shift in models for hospital-acquired acute kidney injury and 30-day all-cause mortality after hospital admission. These outcomes may be optimal for prediction modeling for a number of reasons. They are well-defined and easily measurable and have a short time span from predictor observation to outcome occurrence, which reduces the risk of measurement error and follow-up bias. In addition, both domains are associated with an existing body of risk prediction literature, which provides a basis for predictor selection. Finally, both outcomes of interest have risk factors that are routinely collected during normal delivery of care, which makes these outcomes reasonable targets for automated EHR-integrated probability-based decision support and quality benchmarking.

Hospital-Acquired Acute Kidney Injury

Hospital-acquired acute kidney injury (AKI) affects 5-7% of hospitalized patients[101-104] and is associated with myocardial infarction, chronic kidney disease and end stage renal disease.[105-107] Among the general inpatient population, the AKI mortality rate is 10-15%, and among critically ill patients and those with AKI requiring dialysis, the mortality rate may be more than 50%.[103, 106, 108, 109] Existing clinical prediction models for AKI have generally been restricted to focused subpopulations and developed with logistic regression;[110-119] however, with the increasing availability of large EHR cohorts, a growing number of studies have begun pursuing AKI models based on large national patient populations and advanced modeling methods.[62, 120-

[122] Key risk factors in these models are collected during normal delivery of care and many, such as use of nephrotoxic medications, are modifiable,[3, 62] creating opportunities for EHR-integrated prediction models to support decision-making. However, while clinical use of these models for patient-level decision-making depends critically upon well-calibrated models, calibration has not been consistently reported and calibration drift has not been studies among models for AKI.[110, 111, 118-120, 123]

30-Day All-Cause Mortality After Hospital Admission

While the rate of inpatient death has been declining, there were over 700,000 deaths among hospitalized patients in 2010.[124] Mortality rates are particularly high for older patients, with 73% of deaths occurring among patients over 65 years, and those admitted with respiratory failure, pneumonia, or septicemia.[124] Mortality within 30-days of hospital admission, a key metric of hospital quality and patient safety, is assessed and tracked by the Centers for Medicare and Medicaid Services (CMS) to both inform the public and, since 2013, to adjust reimbursements.[125] CMS quality metrics rely on prediction models to standardize mortality rates by adjusting for the case mix of each hospital's patient population.[125] In addition to quality benchmarking, prediction models for hospital mortality are used to support decision-making and reduce mortality rates, particularly in the critical care setting.[2] CDS tools presenting a patient's predicted probability of 30-day mortality at or near the time of admission/transfer could create and promote opportunities to allocate limited resources, provide additional education/assistance to patients and caregivers, or prompt additional interventions. Our literature review highlighted the long history of prediction models for hospital mortality and concern over calibration drift in this domain. While advanced regression and machine learning methods have been implemented for mortality prediction,[126-128] as noted in our review, studies of calibration drift have primarily focused on logistic regression models. In the only exception, Minne et al[24, 92] reported more stable calibration for a tree-based model compared to a corresponding logistic model. By including hospital mortality in our study, we extend this existing literature through consideration of additional modeling methods, more stringent assessments of calibration, and systematic quantitative evaluations of data shifts.

Data Sources and Definitions

Veterans Affairs Inpatient Data

We collected data on all admissions to all US Department of Veterans Affairs (VA) hospitals nationwide with date of admission between January 1, 2003 and December 31, 2013. These data were available through VA Informatics and Computing Infrastructure (VINCI), a data and analysis resource containing national retrospective data for patients hospitalized at any of 116 VA hospitals and bringing together data from the VA's Computerized Patient Record System (CPRS) and Veterans Health Information Systems and Technology Architecture (VistA).[129, 130] For each admission, we accessed data on laboratory results; diagnosis and

procedure codes (International Classification of Diseases version 9 [ICD-9], ICD-9 Procedure, and Current Procedural Terminology [CPT]); preadmission and administered medications; radiology reports; orders; and vital status.[129] Parallel information and health care utilization data for the year prior to each admission (including events in 2002) was also collected. Further, we linked admissions involving the same patient in order to assess admission history. All associated data collection and analyses in this study was approved by the Institutional Review Board and the Research and Development committee of the Tennessee Valley Healthcare System VA.

Acute Kidney Injury Study Population

Our AKI cohort included admissions beginning between January 1, 2003 and December 31, 2013. A complete list of predictors, details of variable definitions, and details of cohort construction have been published previously.[62] We modeled the probability of patients developing Stage 1+ AKI as defined by the KDIGO classification guidelines.[131] AKI status was defined using the baseline creatinine level (mean outpatient value) and the maximum creatinine value and dialysis procedure codes recorded between 48 hours and 9 days after admission.

Predictor variables and exclusion criteria were based on data collected prior to admission (within 1 year) or within 48 hours of admission. We included predictors selected for prior AKI modeling efforts, which were based on KDIGO guidelines and existing literature.[62] Predictors included demographics, medications, vital signs, body mass index, laboratory values, and diagnoses. Comorbidity information was based on International Classification of Diseases version 9 (ICD-9) Procedure and Current Procedural Terminology codes recorded in the year prior to admission. Medications, vital signs, and body mass index risk factors were summarized separately for the pre-admission and admission windows (24 hours before admission to 48 hours after admission). Baseline creatinine and values collected between 24 hours before and 48 hours after admission were used to determine community-acquired AKI status. Admission window laboratory values were collected between 24 hours before and 48 hours after admission. Other preadmission/index laboratory values were defined by the most recent value collected between 5 days before admission and 48 hours after admission. Preadmission medication covered medications taken 90 days to 24 hours prior to admission. A complete list of predictors is included in Appendix A.

We limited our cohort to admissions with a length of stay between 48 hours and 30 days. Admissions were required to have creatinine values measured prior to admission, within 48 hours of admission, and more than 48 hours after admission. Patients under 18 years of age, patients with dialysis or renal transplant prior to admission, patients with community-acquired acute kidney injury, and patients receiving hospice care within 30 days of admission or within 48 hours after admission were excluded. All admissions to VA facilities with fewer than 100 admissions per year or to facilities not reporting key laboratory data to the central data warehouse were also excluded.

30-Day All-Cause Mortality Study Population

Our 30-day all-cause mortality after hospital admission cohort included admissions beginning between January 1, 2006 and December 31, 2013. Administrative and logistical considerations lead us to leverage a slightly different admissions cohort for our mortality than our AKI analyses. Key mortality predictors were not readily available in our existing AKI cohort, and the cohort underlying our mortality analyses was in-production at the time of study development, allowing us to incorporate our data definitions into the initial build. Since our analyses of the two outcomes are analogous but fully independent and we characterize temporal performance relative to the time of model development (see *Temporal Data-Splitting* section), the chronological misalignment between the AKI and mortality training cohorts and different lengths of validation periods do not complicate interpretation of our findings.

We modeled the probability of death within 30-days of admission to any VA facility, regardless of the cause of death and whether the death occurred during admission or after discharge. We defined a predictor set based on a review of previously published risk models for hospital mortality.[2, 132-134] Predictors included demographics, admission characteristics, body mass index, laboratory values, diagnoses, and health care utilization. We modeled the probability of mortality 48 hours after admission, allowing the use of data collected prior to inpatient stay and during the admission window (24 hours before through 48 hours after admission). For each admission, we applied the algorithm used by CMS to determine whether the admission was planned or unplanned.[135] Comorbidity information was based on ICD-9 codes recorded prior to admission. We recorded dialysis status, a history of dyslipidemia, and conditions included in the Elixhauser comorbidity classification system.[136] The Elixhauser obesity category was excluded in favor of body mass index at admission. We recorded the most recent value for select laboratory tests during the admission window. Health care utilization was characterized as the number of inpatient stays at any VA facility, the number of outpatient visits to VA providers, and whether the patient had any unplanned readmissions over the year prior to the index admission. A complete list of predictors is included in Appendix B.

In order to avoid censorship issues and ensure 30-days of follow-up for outcome ascertainment, we excluded admissions beginning after December 1, 2013. Admissions of patients less than 18 years of age or receiving hospice care at the time discharge were excluded. We further limited to admissions with a length of stay of at least 48 hour. All admissions to VA facilities with fewer than 100 admissions per year or to facilities not reporting key laboratory data to the central data warehouse were also excluded. Finally, we filtered admissions by site, randomly selecting 50% of the sites within each Veterans Integrated Service Network (VISN) for inclusion in this analysis. Site-based filtering of the data was included for practical reasons, as the national sample of eligible admissions was so large as to be computationally infeasible given the complexity and volume of model fitting included in our analyses.

Temporal Data-Splitting

We temporally divided the data into development and validation periods. All admissions in the first year of each dataset (i.e., 2003 for AKI and 2005 for mortality) served as the development cohort. The remaining years of data served as the validation period, resulting in nine years of validation data for the AKI analysis and eight years of validation data for the mortality analysis. Admissions during the validation period were divided into sequential validation cohorts consisting of admissions within each consecutive 3-month period. Our AKI analysis included 36 such validation cohorts; our mortality analysis included 32 such validation cohorts. We labeled each 3-month validation period with the number of years between the end of the development period and end of the validation period.

Modeling Approach

Modeling Methods

Statistical and machine learning techniques approach prediction modeling from different perspectives, each with benefits and limitations. Regression models take a parametric or semiparametric approach, requiring pre-specification of predictor effects and interactions. The effect of each predictor in these models, however, is interpretable, which may be desirable and help support user confidence among some target clinical audiences.[47] Machine learning methods, on the other hand, use model-free algorithm-based approaches to develop predictions without the need to pre-specify the form of predictor-outcome associations or interactions between predictors. Such models may be able to leverage information from complex associations that are difficult to include in parametric models; [16, 47, 49] however, these models require large sample sizes, are not interpretable, and may be difficult to transfer across information systems.[17, 47] As both regression and machine learning methods may provide valid, clinically useful risk estimates, the appropriate modeling approach may vary by clinical use case and the robustness of each modeling method to the changing environment in which the model will be applied.[47]

We explored the robustness of seven common regression and machine learning methods to data shifts and performance drift. The regression models included logistic regression, L-1 penalized logistic regression, L-2 penalized logistic regression, and L-1/L-2 penalized logistic regression. The machine learning methods included naïve Bayes, neural networks, and random forests. Parallel models based on the same training data and predictor set were developed for each model, and the performance of each was assessed across each validation period. A short description of each method is provided below.

Regression Methods

Classical statistical regression techniques, such as logistic and Cox regression, have a long history of use for clinical prediction.[47-49] These parametric data modeling methods

incorporate subject matter knowledge to define the anticipated relationships between predictors and outcomes, as well as the interactions between predictors.[17, 47, 49] With the large predictor set available with EHRs, we may not have the detailed understanding required to fully pre-specify the complex, high-order interactions that exist between predictors and are informative for outcome prediction. While these methods are familiar and may be applied to relatively small cohorts, misspecification of predictor effects may bias risk estimates.[17, 47] Penalized regression methods extend traditional regression models to reduce overfitting by shrinking coefficients to reduce variability.

Logistic regression (LR). The logistic regression model is familiar and provides an interpretable model; however, requires pre-specification of all effects and does not provide in a parsimonious model.[40] Due to our large sample size, we need not be overly concerned with overfitting of our LR models as we have a large number of events per predictor variable. We may, however, include non-informative predictors or multiple highly correlated predictors. The basic LR model retains all specified effects and does not select key predictor effects to create a parsimonious model.

L-2 penalized logistic regression (L2). Often referred to as ridge regression, L-2 penalized regression extends traditional regression models by reducing overfitting by shrinking of coefficients. This is achieved by restricting the sum of the squared coefficients, which shrinks coefficients without allowing any coefficients to be reduced to 0.[137] This approach reduces overfitting but does not provide for variable selection. Groups of correlated predictors will tend to be assigned similar coefficients by the L2 model.[138]

L-1 penalized logistic regression (L1). Commonly known as lasso regression, L-1 penalized regression provides for both overfitting and variable selection. The L1 model allows some coefficients to shrink to 0 by restricting the sum of the absolute values of the coefficients.[139] While allowing for a more parsimonious model, a limitation of the L1 model approach is that among groups of correlated predictors, a single variable may be randomly selected for inclusion.[138]

L-1/L-2 penalized logistic regression (L1-L2). By combining both the L1 and L2 penalizations, L-1/L-2 penalized regression, known as the elastic net model, addresses issues of overfitting, variable selection, and groups of correlated predictors. In the L1-L2 model, groups of correlated predictors are included to excluded form the model as a group and, if included, tend to have similar coefficients.[138] The L1-L2 model, however, remains limited by the requirement to pre-specify all predictor affects and interactions.

Machine Learning Methods

With the increasing availability of high dimensional clinical datasets, machine learning techniques are increasingly being applied to clinical prediction tasks.[1, 16, 59] Machine learning methods are based on algorithmic approaches that do not require prespecification of effects and

leverage information in large datasets to capture previously unrecognized and complex associations for improved prediction.[16, 47, 49] These models may be better able to characterize the complex relationships in clinical datasets, but do so by constructing models that are not human-interpretable and thus may not be suitable for all use cases.[17, 47]

Naïve Bayes (NB). The naïve Bayes model takes an efficient approach to prediction by applying Bayes Rule under the assumption that all predictors are independent. This assumption allows predictions on new data based on simple summaries of the training set into prior probabilities and conditional probabilities calculated separately for each predictor.[140] This approach tends to work well for classification,[141] but can result in predicted probabilities that are extreme (i.e., pushed toward 0 or 1) when any predictors are rarely or frequently observed with the outcome, harming model calibration.[142]

Neural Networks (NN). Neural network models are composed of multiple layers of connected nodes, each having a value based on an activation function combining weighted information from input nodes.[143] The input layer consists of one node for each predictor variable; the output layer consists of nodes for the outcome variable. Any number of hidden layers is possible, however, one is typically adequate.[60] The weighted connections between the layers allow for complex non-linear associations and high-order interactions between predictors, without these relationships requiring prespecification.[60, 144] Computational requirements and the need to define the model structure without overfitting can be limitations of NN models; however, NN have a long history of utilization in the clinical prediction literature.[60]

Random Forest (RF). Classification and regression trees are graphical models that partition data by repeatedly splitting the sample. At each level of the tree, the sample is split on a single predictor selected to maximize the purity (i.e., consistency of the outcome) of the data in each branch.[145] The random forest model extends this approach by combining predictions across a large number of trees, each constructed with a subsample of predictors and a bootstrap sample of the training data.[146] The number of predictors considered for each tree affects model performance by balancing the strength of prediction for individual trees and correlation among trees.[146] The RF approach implicitly allows for complex predictor-outcome associations and interactions among predictors, while also allowing the consideration of any number of predictors. [16, 47, 49, 146]

Model Development

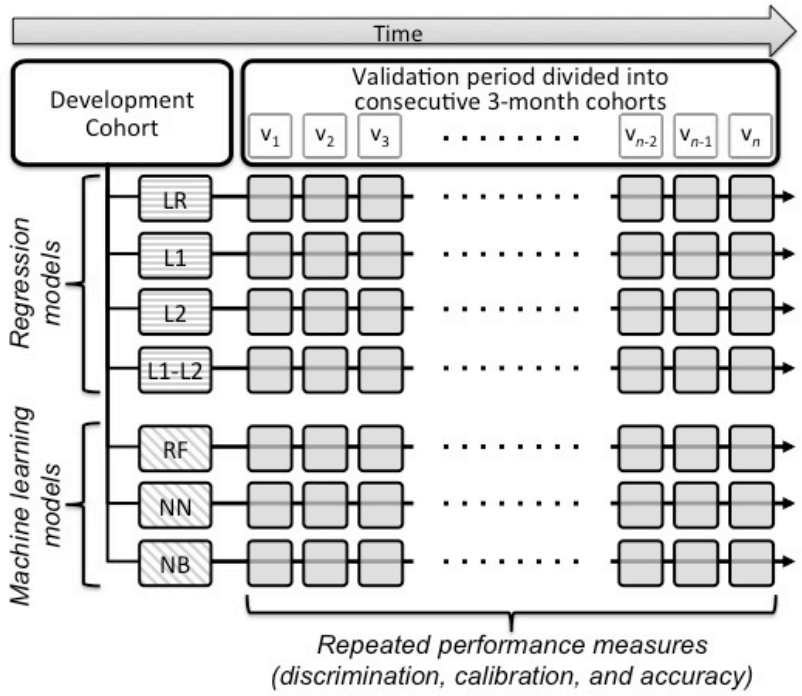
For both AKI and mortality, we fit parallel models using each of the seven methods based on a common predictor set and admissions in the one-year development cohort (see Figure 5). For the regression models, continuous predictors were fit with restricted cubic splines to capture nonlinear associations; however, no interaction effects were specified. We specified a NN model with one hidden layer.[60] For methods with hyperparameters (L1-L2: alpha penalty level; RF: number of predictors considered per split, number of trees, minimum node size/depth; NN: size of hidden layer), values were selected using 5-fold cross validation. Models were internally validated with the bootstrap (B=200). For each bootstrap iteration, we imputed missing

laboratory, demographic, and health care utilization values with predictive mean matching and used the same imputed dataset to fit each type of model. There were no missing values for predictors based on administrative diagnoses codes as we assumed the condition to be not present if no codes were recorded in the medical record. The lambda shrinkage parameter for the penalized regression methods was selected with cross-validation for each bootstrap iteration. We constructed a final version of each model based on the full development cohort.

Temporal Performance Evaluation

We assessed performance with measures of discrimination, calibration, and accuracy evaluated at development and in each quarterly validation period (see Figure 5). Metric definitions, ideal values, and interpretations are described below and summarized in Table 6.

Figure 5. Analysis structure with parallel models, temporal data-splitting, and repeated validations



Measures of Model Accuracy

Model accuracy describes overall model fit, capturing aspects of both discrimination and calibration. We evaluated accuracy with two proper scoring rules. The Brier score, a quadratic scoring rule, is the mean squared error in prediction (i.e., $Brier = \frac{1}{n}\sum_i^n (Y_i - p_i)$). The range of the Brier score, however, varies with the outcome rate, challenging interpretation across samples.[65] To simplify comparisons across our temporal cohorts, we implemented the scaled

Brier score, which measures model error as a proportion of maximum error and scaled such that 0 indicates maximum error and 1 indicates perfect accuracy (i.e., $Brier_{scaled} = 1 - \frac{Brier}{Brier_{max}}$

Table 6. Definitions and interpretations of model performance metrics

Metric	Description	Range	Ideal value
Accuracy			
Scaled Brier score	Mean squared difference between observed outcome and predicted probability, scaled by maximum possible value based on non-informative model.[65]	0 - 1	1
Nagelkerke's R^2	Variation of the logarithmic scoring rule, scaled such that higher values indicate a more accurate model.[65]	0 – 1	1
Discrimination			
Area under the Receiver Operating Characteristic curve (AUC)	An assessment of the probability that an observation with the outcome is assigned a higher risk estimate than an observation without the outcome. Values of 0.5 indicate an uninformative model, and values of 0.5 or 1.0 indicating perfect discrimination.[75]	0 - 1	1
Calibration			
Observed to expected ratio (O:E)	Ratio of outcome rate to mean predicted probability. Values greater than 1 indicate underprediction of risk, on average. Values less than 1 indicate overprediction of risk, on average.	0 - unbounded	1
Cox intercept	Intercept (α) of the logistic calibration curve model: $logit(y) = \alpha + \beta * lp$ where $\beta = 1$.[147] Values greater than 0 indicate systematic underprediction; values less than 0 indicate systematic overprediction. Also referred to as calibration-in-the-large.[22, 65]	unbounded	0
Cox slope	Slope (β) of the logistic calibration curve model: $logit(y) = \alpha + \beta * lp$.[147] Values less than 1 indicate overfitting, with predicted values having too much variability and predictor affects requiring shrinkage.[22, 65]	unbounded	1
Estimated calibration index (ECI)	Mean squared difference between predicted probabilities and observed probabilities estimated from flexible calibration curves, scaled to range from 0 to 100.[71, 72]	0 - 100	0

where $Brier_{max} = mean(Y) * \left(1 - mean(Y)\right)^2 + \left(1 - mean(Y)\right) * mean(Y)^2$).[65] Similarly, the logarithmic scoring rule can be difficult to interpret and have an undefined range, so we implemented the Nagelkerke's R^2, a normalized logarithmic scoring rule ranging from 0 for the least accurate model to 1 for the perfectly accurate model that can interpreted as a pseudo-R^2 describing the proportion of explained variation.[65]

Measures of Model Discrimination

Discrimination describes the ability of the model to distinguish events from non-events or correctly rank-order observations by risk.[65] The receiver operating curve plots model sensitivity against the false positive rate (1-specifity). The area under the curve (AUC) provides an assessment of the probability that an observation with the outcome is assigned a higher predicted probability than an observation without the outcome. AUC ranges from 0 to 1, with 0.5 indicating an uninformative model and 1.0 indicating perfect discrimination.[75]

Measures of Model Calibration

Calibration describes the agreement between observed and predicted risk, or how well the predicted probability aligns with the true probability that an individual will experience the outcome.[148] Van Calster et al[71] proposed a 4-tier hierarchy for assessing model calibration. The highest tier in this hierarchy may not be realistic, assessable, or necessary.[71] The third tier, moderate calibration, however, can be shown to ensure models have a net benefit greater than or equal to treat-all or treat-none strategies, thus ensuring predictions are nonharmful to clinical decision-making.[71] We thus characterized calibration across the first three tiers of the calibration hierarchy – mean, weak, and moderate calibration.

Mean calibration, the weakest form of calibration, requires agreement between the predicted and observed risk on average across all observations.[71] We characterized mean calibration with the observed to expected outcome ratio (O:E) and the intercept of the Cox recalibration model. The O:E ratio compares the mean predicted probability with the population event rate. O:E ratio values of 1 indicate perfect calibration, while values less than 1 indicate average overprediction of risk and values greater than 1 indicate average underprediction.[24] Systematic over- or underprediction is also captured in the intercept of the Cox recalibration model ($\propto$ in the model $logit(y) = \propto + \beta * lp$ where lp is the logit-scale prediction and $\beta = 1$).[147] Calibrated models will have an intercept of 0, while systematic overprediction will result in intercepts less than 0 and systematic underprediction will result in intercepts greater than 0.[22, 65]

Weak calibration builds on mean calibration by requiring neither systematic over- or underprediction nor over- or underfitting.[71] By extending calibration to include an assessment of over- and underfitting, weak calibration considers the appropriateness of the variability of the predictions rather than just the mean of the predictions. An overfit model will have predictions that are too extreme for low and high risk observations; an underfit model will have predictions that are too close to the mean for both low and high risk observations.[65] We assessed weak calibration with the intercept ($\propto$) and slope (β) of the Cox recalibration model.[71] For a model

meeting weak calibration standards, the intercept and slope would be 0 and 1, respectively. Over- and underfitting are measured by the slope, with values less the 1 indicating overfitting and values greater than 1 indicating underfitting.[65]

Moderate calibration, the most stringent level considered in this analysis, requires more detailed alignment of predicted and observed probabilities across the range of predictions.[71] For a binary outcome, we do not have true observed probabiltities, thus we may think of moderate calibration as comparing the observed outcome rate with the predicted probability for groups of observations with similar predicted risk. For a moderately calibrated model, this means that a plot of the observed proportion versus predicted probability would form approximately a 45° line. The Hosmer-Lemeshow test, which assesses calibration across predicted probabilities using a chi-squared test based on *g* equally-sized groups (typically *g*=10), is a commonly reported measure of moderate calibration.[71, 149] This test, however, is not well-suited to large datasets, as its power to detect miscalibration increases with sample size, resulting in significant tests even for very small deviations from perfect calibration when samples sizes are large.[150] This test statistic also lacks sensitivity to multiple common forms of miscalibration, including overfitting or systematic differences between datasets.[22] Thus, while a prominent metric in the literature, we do not evaluate the Hosmer-Lemeshow test. Instead, we assess moderate calibration with the recently proposed flexible calibration curve approach. These curves are constructed by fitting a logistic model for the observed outcome based on predicted probabilities fit with a restricted cubic spline (see Figure 6).[72, 151] As graphical comparison of multiple curves is difficult, we summarize curves with the estimated calibration index (ECI), the mean squared difference between predicted probabilities and estimated observed probabilities from the flexible calibration curves.[71, 72] The ECI is scaled to range between 0 and 100, with lower values indicating greater calibration.

We extended this analysis of moderate calibration to explore in more detail how model performance shifted in and out of calibration across the full range of predicted probability. As illustrated in Figure 6, using confidence intervals around the flexible calibration curves, we determined regions of calibration, overprediction, and underprediction. Ranges of predicted probability in which the 95% confidence interval (CI) included the 45° line were labeled as regions of calibration. Ranges of predicted probability in which 45° line was below or above the 95% CI were labeled as regions of overprediction and underprediction, respectively. We further divided regions of over- and underprediction into marginally miscalibrated (i.e., regions where the 99% CI included the 45° line but the 95% CI did not) and miscalibrated regions (i.e. regions where the 99% CI did not include the 45° line). We calculated the ECI within each region to assess the magnitude of miscalibration. For any regions with less than 50 admissions, we stabilized the regional ECI by borrowing information from adjacent regions. Since observations are not uniformly distributed across the range of predicted probability, we also rescaled the regions by the volume of observations with predicted probabilities within each region (see Figure 7). The original scale of the regions provides a sense of calibration across the entire range of probability, while the proportional regional volume assessment emphasizes calibration status based on data density and thus calibration status of the ranges of probability most relevant to the actual data.

Figure 6. Example flexible calibration curve with 95% (darker band) and 99% (lighter band) confidence intervals and regions of predicted probability by calibration status. The black 45° line indicates perfect calibration.

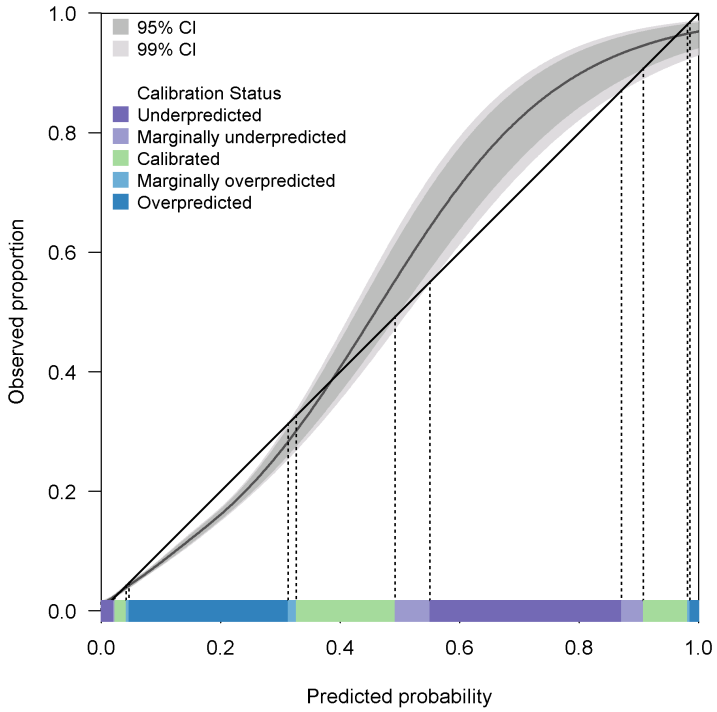
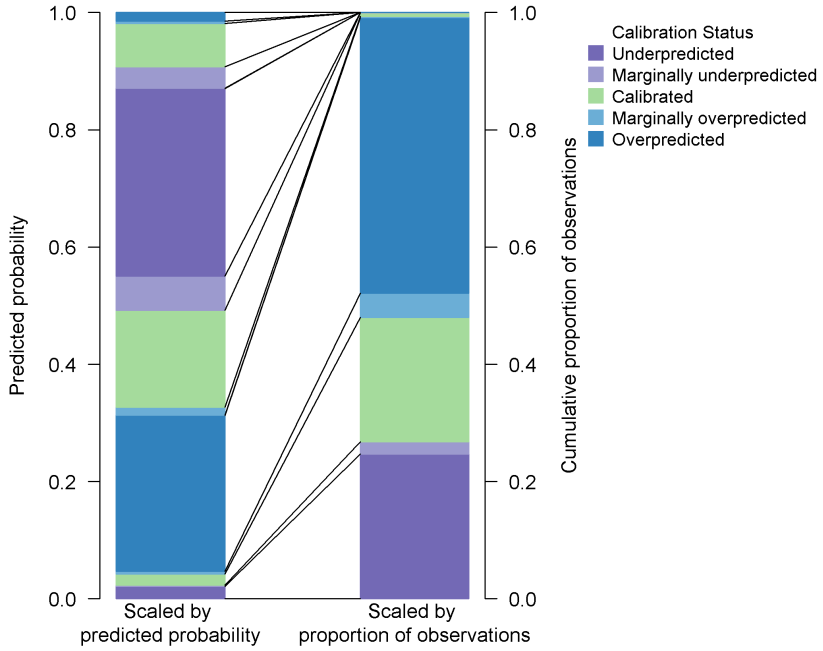


Figure 7. Illustration of rescaling regions of calibration by data density for proportional regional volume assessment



Temporal Validation

In order to assess the initial performance of each model, we recorded the performance of the models developed for each bootstrap iteration on those observations not selected for that iteration's training sample (i.e., holdout sample). We constructed 95% confidence intervals for each metric based on the mean and standard deviation across the holdout samples. These initial performance measures serve as a reference for assessing temporal changes in performance in subsequent validation cohorts. For the regions of calibration analysis, we combined predictions across all 200 holdout samples to construct the flexible calibration curves used to define the regions of probability over which each model was calibrated, overpredicted, and underpredicted. These combined holdout samples were also used for the proportional volume assessment.

We applied each model to all validation observations, providing multiple predicted probabilities of the outcome for every observation (seven for AKI observations; seven for readmission observations). Within each 3-month temporal validation cohort, we assessed performance with the percentile method by calculating mean performance and 95% confidence intervals for each model across 1,000 bootstrap samples. Flexible calibration curves for the regions of calibration and proportional volume analysis were based on all observations in each temporal cohort.

Data Shift Assessment

Performance drift results from data shifts, or differences that arise over time between the population on which the model was developed and the population on which the model is applied. Data shifts may take the form of changes in the underlying outcome rate, case mix of the patient population, and/or associations between predictors and outcomes,[6, 7, 23] which may result from changes in the health care system's patient population, treatment and diagnosis patterns, or measurement methods.[5, 6, 21] In order to better understand drivers of any observed performance drift and any differences in drift by modeling method, we evaluated our AKI and mortality cohorts for each type of data shift using multiple approaches.

Event Rate and Case Mix Shift Assessment

We explored the presence of event rate and case mix shift by assessing distributions of individual predictors, distribution of patient predicted risk, and membership models predicting whether observations belong to the development or validation cohorts.

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 temporal linear and non-linear changes in the distribution of each variable, adjusting for multiple comparisons with the Bonferroni method. This approach is straightforward and has been used to explore case mix changes affecting models for hospital mortality.[23] However, small changes in predictor distributions may be highlighted as significant trends;

simultaneous changes in multiple predictors can be difficult to interpret, particularly when predictors are correlated; and linking observed changes in predictor distributions directly to model performance can be challenging.[73, 148]

We 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 both the development and validation cohorts, these models predict whether an observation is from the validation cohort based on covariates that include all predictors and the outcome of the original model.[73] High levels of discrimination indicate case mix and event rate drift are present. Further, we can interpret the structure of the membership model to identify predictors contributing to case mix changes.[73] We fit separate membership models comparing the development cohort to each consecutive validation cohort using both logistic regression and random forest models. The AUC of the logistic regression membership models was recorded to determine the presence of case mix and event rate shift relative to the development set for each validation cohort.[73] AUCs were adjusted for optimism using the 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[73] and how the contribution of each predictor to case mix shift may have changed over time. Similarly, we documented the variable importance rank of each variable in the random forest membership models to explore how the relative importance of each predictor to case mix shift may have changed over time. Variable importance ranks were assigned based on the decrease in node impurity (Gini node impurity index) due to splitting on a particular variable averaged over all trees, with larger decreases indicating more important//higher ranked predictors. For each predictor, we tested for linear and non-linear changes in variable importance rank over time, adjusting for multiple comparisons with the Bonferroni method.

Finally, we also explored summary approach to characterizing case mix shift. We documented changes in the mean and standard deviation of the predicted probabilities over time, providing an indication of shifts in average patient severity and patient population heterogeneity of risk, respectively.[73, 148] Changes in heterogeneity of risk has been linked to changes in model discrimination, with increased heterogeneity leading to increased discrimination.[73] By combining these measures with the AUC of the membership models, we can assess both the degree of case mix shift and how this case mix shift is affecting the patient population risk distribution.

Predictor-Outcome Association Shift Assessment

We explored shifts in the strength of associations between predictors and the outcome, as well as shifts in the relative importance of predictors, by considering how the structure of models changed over time when refit for each validation cohort. We documented changes the odds ratio for each predictor from LR models and changes in variable selection patterns from L1 models fit in each validation cohort. Since variable selection may be unstable, we implemented the bootstrap with 200 iterations for each validation cohort when refitting the L1 model. For each predictor, we calculated the proportion of bootstrapped iterations in which the predictor was selected by the L1 models in each validation period [138, 152] and tested for linear and non-linear

changes in selection proportion over time. We also documented temporal patterns in the relative importance of each predictor as measured by changes in RF variable importance ranks. Predictors were ranked by the mean decrease in the Gini node impurity index, with larger decreases indicating more important//higher ranked predictors. These refit RF models were developed with the hyperparameters of the original RF model based on the development cohort. For each predictor, we tested for linear and non-linear changes in variable importance rank over time. We adjusted for multiple comparisons with the Bonferroni method within in component of the analysis.

CHAPTER 4

TEMPORAL EVALUATION OF MODELS PREDICTING HOSPITAL-ACQUIRED ACUTE KIDNEY INJURY

We explored performance drift and data shifts among models for hospital-acquired acute kidney injury in a national population of admissions to VA hospitals over the 10-year period from 2003-2012. Across seven parallel model fit with logistic regression, L-1 penalized logistic regression, L-2 penalized logistic regression, L-1/L-2 penalized logistic regression, naïve Bayes, neural networks, and random forests, we document diverging patterns of calibration drift. This performance deterioration occurred in the presence of complex, multi-form data shifts in the patient population.

Our national VA cohort for the AKI study consisted of 1,841,951 admissions, 170,675 during the development cohort (i.e., admissions beginning in 2003) and 1,671,276 during the validation period (i.e., admissions beginning in 2004 through 2012). Each of the 36 consecutive temporal validation cohorts included a mean of 46,424 admissions (range: 42,168-49,798). A brief summary of the patient population at select points across the study period is presented in Table 7 (See Appendix C for details for all predictors). Patients were primarily white males (96.1%), with a mean age of 66.1 years (standard deviation: 13.0) and mean body mass index of 27.8 (standard deviation: 7.5). Overall, 6.8% of admissions were complicated by AKI.

Model Development

Modeling Parameters

Prior to developing and internally validating the seven parallel models, we selected hyperparameters values for the L-1/L-2 penalized logistic regression, random forest, and neural network models. Table 8 provides the values selected through 5-fold cross validation. The shrinkage parameters for the three penalized regression models were selected during each bootstrap iteration.

Initial Model Performance

Bootstrap-corrected performance metrics for the seven models at development are presented in Table 9. We observed similar levels of overall accuracy among the regression models, with slightly lower levels of accuracy for the RF and NN models. Discrimination was modest, with AUCs ranging from 0.69-0.76. The LR, L1, and L1-L2 models were most discriminative; the NB model was least discriminative. The four regression models and the NN model were well-calibrated based on the O:E ratio and ECI. The RF model was well-calibrated based on ECI, while slightly underpredicting according to the O:E ratio (1.07, 95%CI:1.06-1.07).

Table 7. AKI Patient population at development (2003) and in three years of the validation period (2006, 2009, 2012)

	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: ACEi=angiotensin converting enzyme inhibitor; GFR=glomerular filtration rate; SD=standard deviation

Table 8. Hyperparameters selected for AKI models

Model	Hyperparameter	Value
L1-L2	Elastic-net mixing parameter (α)	0.8
Random forest	# predictors considered per tree	14
	Minimum # observations per node	15
	# trees	600
Neural network	Size of hidden layer	100

Table 9. Initial AKI model performance in development cohort

	Regression				Machine Learning		
	LR	**L1**	**L2**	**L1-L2**	**RF**	**NN**	**NB**
Accuracy							
Scaled Brier score	0.081 [0.080, 0.081]	0.078 [0.077, 0.078]	0.069 [0.068, 0.069]	0.078 [0.077, 0.078]	0.062 [0.062, 0.062]	0.049 [0.049, 0.049]	*
Nagelkerke's R^2	0.154 [0.154, 0.155]	0.150 [0.149, 0.150]	0.135 [0.135, 0.136]	0.150 [0.149, 0.150]	0.113 [0.112, 0.113]	0.102 [0.102, 0.103]	*
Discrimination							
AUC	0.764 [0.764, 0.765]	0.761 [0.761, 0.762]	0.750 [0.750, 0.751]	0.761 [0.761, 0.762]	0.734 [0.734, 0.735]	0.720 [0.719, 0.720]	0.692 [0.692, 0.693]
Calibration							
O:E ratio	1.001 [0.998, 1.003]	1.002 [0.999, 1.004]	1.002 [1.000, 1.005]	1.002 [0.999, 1.004]	1.066 [1.060, 1.072]	1.003 [1.000, 1.006]	0.225 [0.217, 0.233]
Cox intercept	-0.087 [-0.094, -0.079]	0.085 [0.074, 0.096]	0.231 [0.221, 0.241]	0.094 [0.082, 0.106]	-0.283 [-0.291, -0.274]	-0.211 [-0.221, -0.201]	-2.416 [-2.430, -2.402]
Cox slope	0.958 [0.955, 0.962]	1.039 [1.035, 1.043]	1.105 [1.100, 1.109]	1.043 [1.038, 1.048]	0.840 [0.836, 0.844]	0.903 [0.899, 0.907]	0.083 [0.082, 0.084]
ECI	0.004 [0.003, 0.004]	0.004 [0.004, 0.004]	0.007 [0.007, 0.008]	0.004 [0.004, 0.005]	0.006 [0.006, 0.007]	0.008 [0.008, 0.009]	21.028 [20.189, 21.867]

* Non-calculable due to extreme predicted probabilities of 0 and 1.

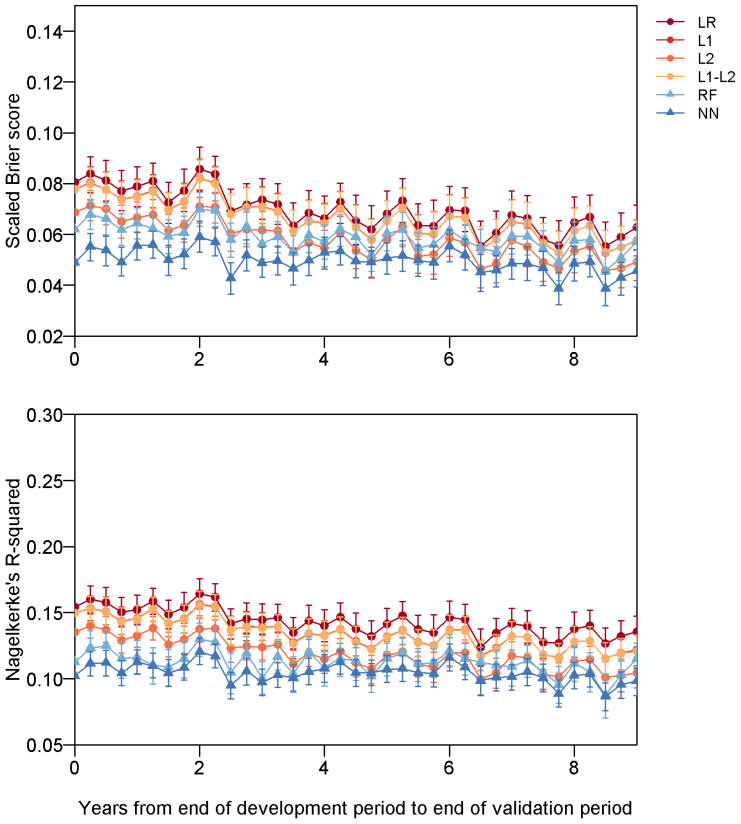
While no models were perfectly calibrated according to the Cox intercept and slope, these metrics approached their ideal values for the LR, L1, and L1-L2 models. The Cox intercept and slope of the RF and NN models indicated minor systematic overprediction and overfitting, while these metrics indicated minor underprediction and underfitting of the L2 model. The NB model lacked accuracy and calibration as measured by all metrics.

Model Performance Over Time

Accuracy

The accuracy of all models declined over time, with a larger magnitude of change observed for the regression models (see Figure 8). For both the scaled Brier score and Nagelkerke's R^2, smaller values indicate lower model accuracy. The scaled Brier score declined over time for all models (adjusted $p<0.001$), with the rate of change larger for the regression models than the RF and NN models. This metric initially indicated lower levels of accuracy for

Figure 8. Accuracy of AKI models over time by modeling method. Regression models are displayed with circular markers and red-orange colors. Machine learning models 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.

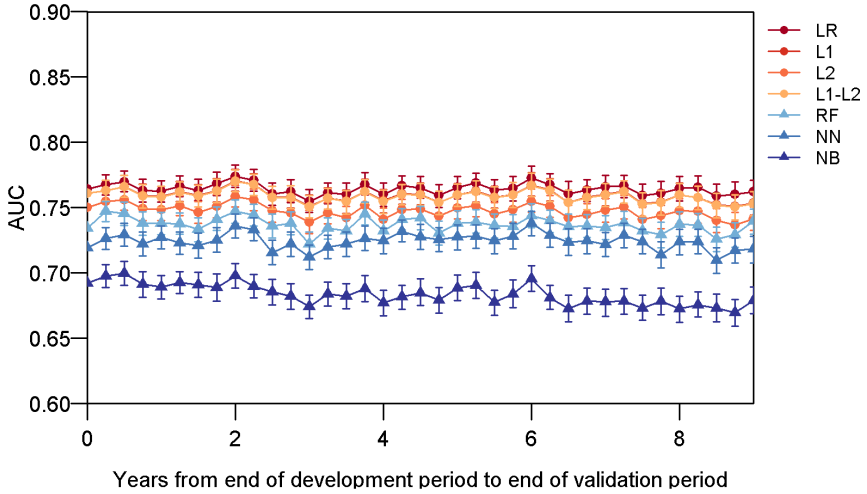


the NN model compared to the regression models, with the difference attenuating over the first 3-4 years of the validation period as the scaled Brier score declined more quickly for the regression models. The Nagelkerke's R^2 was stable for the RF and NN models, while declining over time for the regression models (adjusted $p<0.001$). The decline in accuracy among the regression models lead to the difference in the Nagelkerke's R^2 between the regression and machine learning methods being attenuated over time. The NB model consistently underperformed all other models on both accuracy metrics.

Discrimination

Over the 9-year validation period, discrimination was stable for all models except the NB (adjusted $p<0.001$; see Figure 9). The NB model experienced a small decline in AUC over time (slope: -0.002, 95% CI: -0.003, -0.002), which resulted in an overall change in AUC across the study of -0.01 (AUC of 0.69 at developed to 0.68 in the final validation cohort). The regression models generally maintained higher AUCs compared to the machine learning models. The L2 regression model, however, had comparable discrimination to the RF model and significantly lower AUCs than the LR model in eleven validation cohorts, primarily after five years.

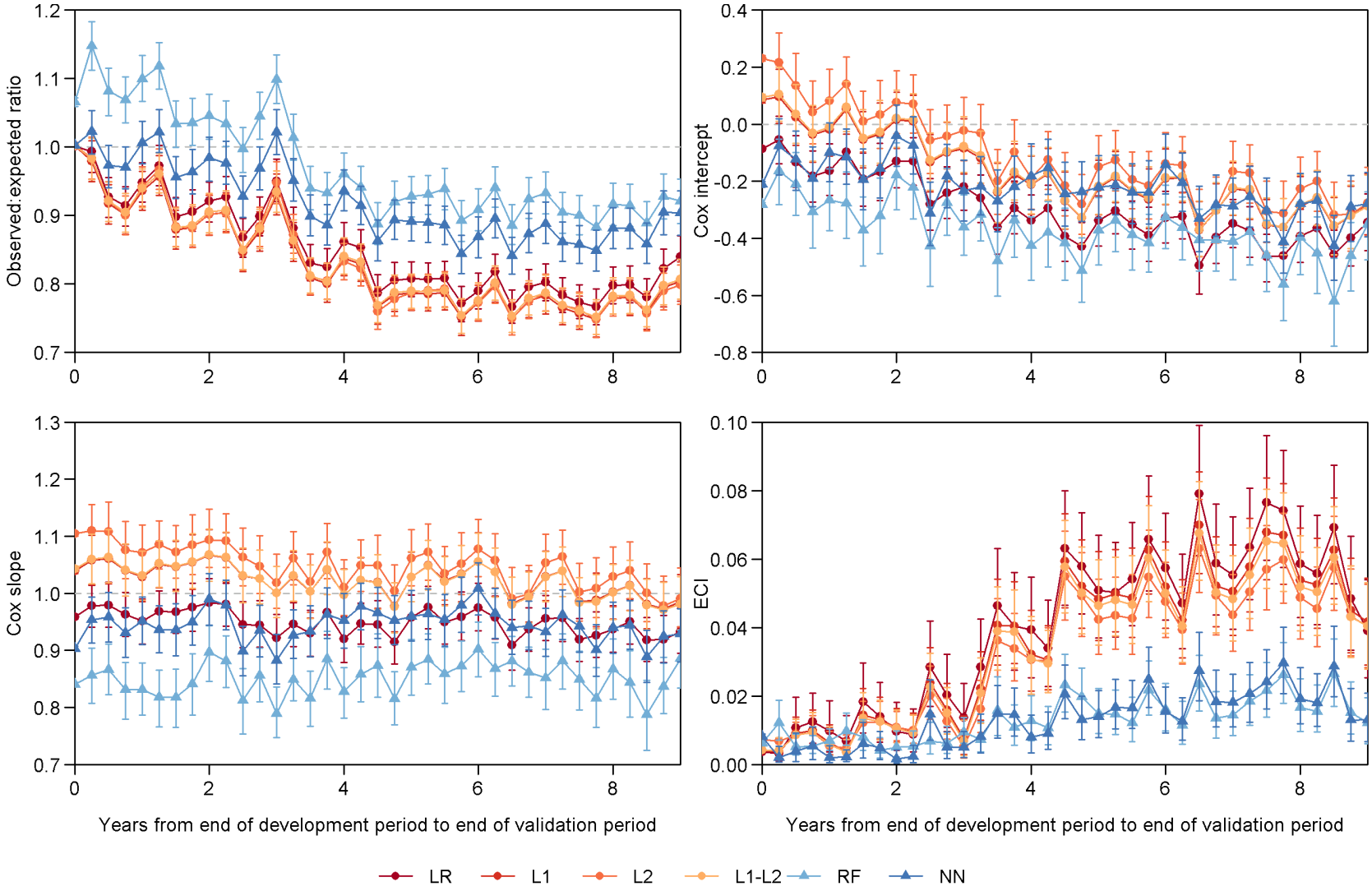
Figure 9. Discrimination of AKI models over time by modeling method. Regression models are displayed with circular markers and red-orange colors. Machine learning models are displayed with triangular markers and blue-purple colors.



Calibration

All models experienced calibration drift (see Figure 10). The models experienced similar temporal patterns of calibration drift for measures of mean (O:E ratios) and weak calibration (Cox calibration model intercepts and slopes). O:E ratios and Cox intercepts primarily declined over the first four years of the validation period. Within these four years, we observed several

Figure 10. Calibration of AKI models over time by modeling method. Regression models are displayed with circular markers and red-orange colors. Machine learning models 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.



periods of rapid calibration drift, particularly in the first and fourth years of validation. Cox intercepts continued to decline after the first four years of validation, although at a slower rate. In the second half of the validation period, all models had O:E ratios less than 1 and Cox intercepts less than 0, indicating overprediction. O:E ratios for the RF and NN models were significantly higher than O:E ratios for the regression models three years after development. Cox intercepts were higher for the penalized regression models than the NN and RF models for the first four years and across the validation period, respectively. Together, these two findings indicate a larger magnitude of average overprediction among the regression models compared to the RF and NN models. Cox slopes declined slightly over time for the regression models and were stable for the RF and NN models (adjusted $p<0.001$). The rate of change in the Cox slope among the regression models was small, with, for example, a slope of -0.004 (95%CI: -0.007, -0.002) for the LR model resulting in a change from 0.958 (95% CI: 0.955, 0.962) at development to 0.933 (95% CI: 0.895, 0.974) in the final validation period. Despite these changes in the Cox slope, the confidence intervals for the Cox slopes for the regression models captured the ideal value of 1.0 for most validation cohorts. The LR and NN models provided slightly overfit predictions with the Cox slope less than 1.0 and intermittently significant. The RF model exhibited overfitting across the validation period.

We observed diverging patterns of calibration drift among models for measures of moderate calibration base on flexible calibration curves. ECIs increased over the validation period for all models (adjusted $p<0.001$), indicating deteriorating calibration. For all models, the rate of drift in ECI was slow over the first three years of the validation period, after which we observed larger increases in ECIs over time. ECIs exhibited varying calibration drift patterns between the regression and machine learning models, particularly in the second half of the validation period. After three years, ECIs deteriorated more substantially and became significantly higher for the regression models compared to the RF and NN models. The ranges of predicted probabilities and proportion of admissions over which each model was calibrated also changed over time and varied by modeling method (see Figures 11 and 12). With the exception of the NB model, which strongly overpredicted for most predicted probabilities and strongly underpredicted for the lowest predicted probabilties models moved in and out of regions of calibration, overprediction, and underprediction across the range of predicted probability. The penalized regression models maintained calibration over time in the 0.4 – 0.9 range, while the RF and NN models tended to underpredict in this range with surrounding calibration regions for lower and higher probabilities. These patterns were fairly consistent over time. Rescaling each region by the volume of admissions captured with the range of predicted probability, we observe similar grouping of temporal patterns – one for regression models and one for the NN and RF models. For the regression models, the majority of admissions fell within regions of overprediction for most validation cohorts. Over the second half of the validation period, the magnitude of this overprediction increased, as indicated by the darkening of the blue region. In the first three years of validation, for the majority of admissions, the NN model was calibrated and the RF model slightly underpredicted or was calibrated. After three years, regions of overprediction captured a higher proportion of admissions for both the RF and NN models; however, the magnitude of overprediction remained low for most validation cohorts. The NB model significantly unpredicted for approximately 40% of admissions and significantly overpredicted for approximately 60% of admissions.

Figure 11. Regions of calibration of AKI models over time by modeling method. Areas of overprediction and underprediction are shaded based on based on within-region estimated calibration index (ECI) in order to highlight the magnitude of miscalibration.

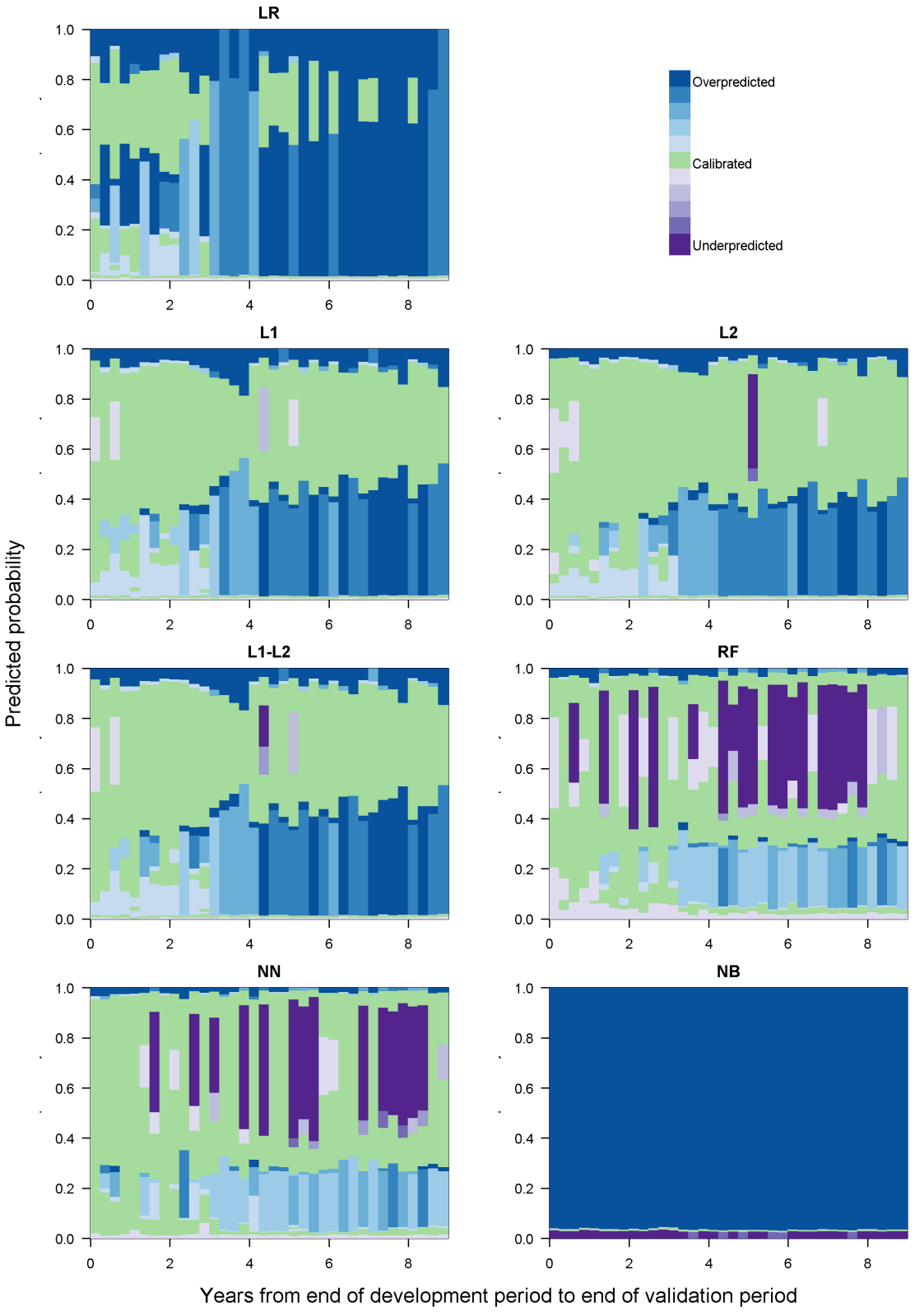
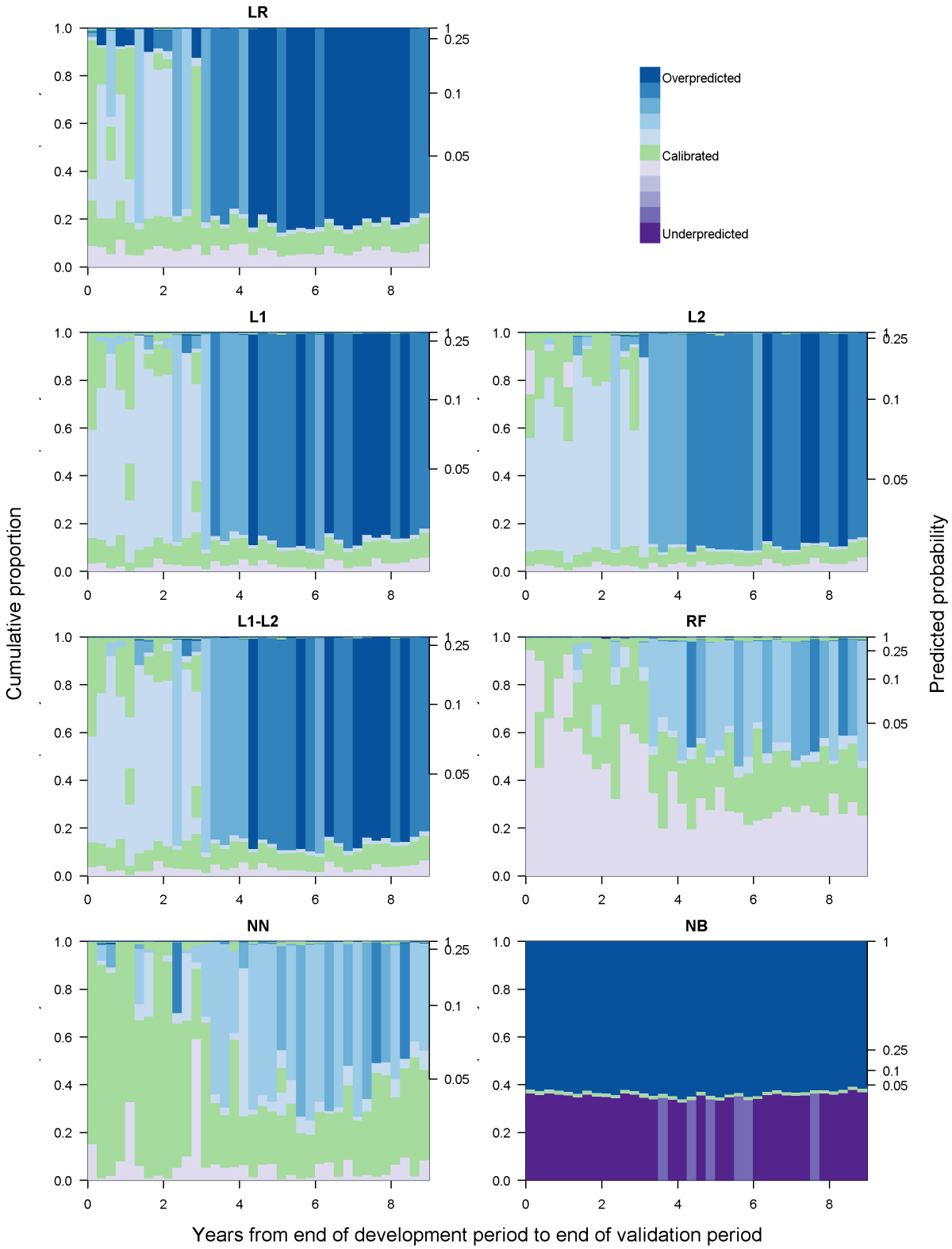


Figure 12. Proportional volume assessment regions of calibration of AKI models over time by modeling method. Regions of calibration are scale by proportion of observations in each region and shaded by the magnitude of the within region estimated calibration index (ECI).



Data Shifts Over Time

Event Rate Shift

We observed evidence of event rate shift across the three assessment methods. Across the validation period, the event rate declined from 7.7% in the development cohort to 6.3% in the final quarterly validation cohort (adjusted p<0.0002; see Figure 13). AKI was also a significant predictor in the logistic regression membership models discriminating between admissions from the validation and development cohorts (see Figure 14). The odds of an

Figure 13. Proportion of admissions complicated by AKI over time. Red lines indicate fitted values.

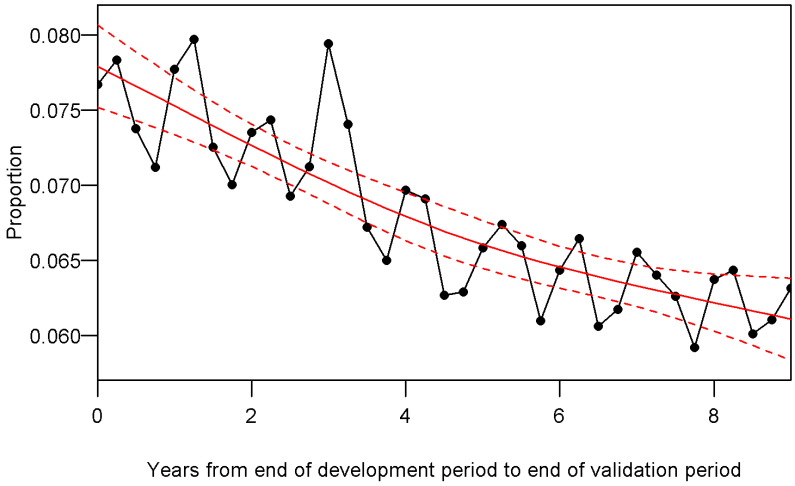
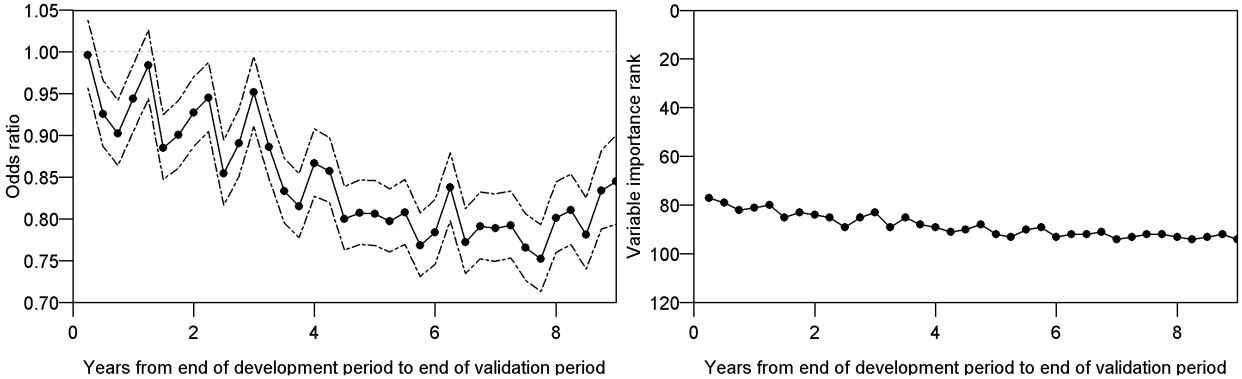


Figure 14. Membership model results for AKI. Odds ratio (left) and 95% confidence intervals for AKI based on logistic membership models fit for each 3-month temporal validation cohort. Grey line indicates null odds ratio of 1. Variable importance rank (right) for AKI based on random forest membership models fit for each 3-month temporal validation cohort. Higher ranks (smaller numbers) indicate greater importance to model performance.

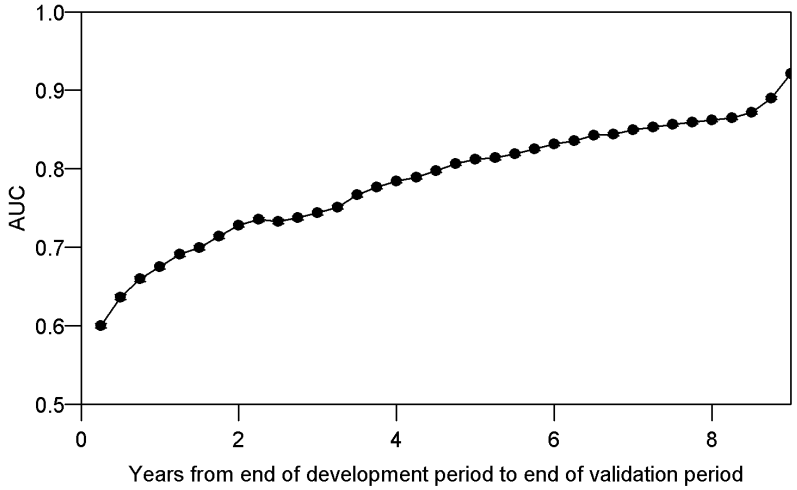


admission belonging to a validation sample were approximately 15-25% lower for admissions complicated by AKI than admissions without AKI in validation years 4 through 9. The relative importance of the outcome compared to predictors in the random forest membership model, however, declined over time from the 77th most important variable to 94th. The proportion of admissions complicated by AKI, magnitude of the odds ratio from the logistic regression membership model, and variable importance rank from the random forest membership model shifted at a higher rate in the first half of the validation period and slowed or stabilized.

Case Mix Shift

We observed case mix shift across the validation period. Changes over time distribution of the linear predictors revealed changing patient case mix in terms of increasing severity of the patient population (adjusted p<0.002). However, we did not observe changes in heterogeneity of risk in the patient population for most models, with the exception being the RF model which experienced a decline in the standard deviation to indicate decreasing heterogeneity of the patient population (adjusted p<0.002). The membership model approach also noted the presence of case mix shift. The logistic membership models increasingly discriminated between admissions from the development and each sequential validation cohort, as indicated by the AUC increasing from 0.601 (95% CI: 0.598, 0.603) to 0.921 (95% CI: 0.919, 0.922) (see Figure 15).

Figure 15. Discrimination of AKI logistic membership models over time

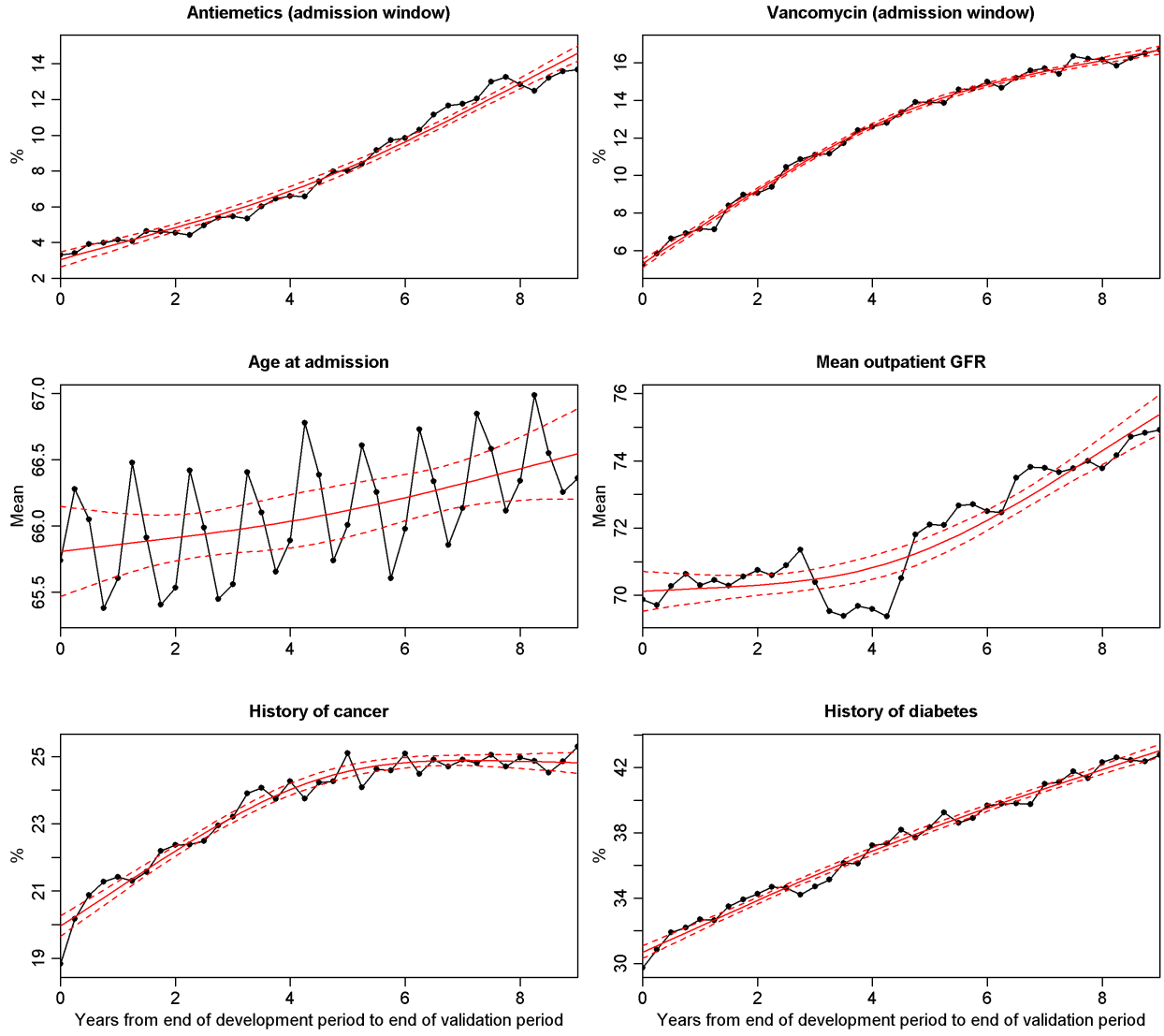


For predictor-level assessments of case mix shift, we present detailed graphical results for six consistent exemplar predictors – use of antiemetics and vancomycin during the admission window, age, mean outpatient glomerular filtration rate (GFR) prior to admission,

history of cancer, and history of diabetes. Findings for these predictors highlight several data shift patterns observed among the full set of predictor variables.

Temporal changes were observed in the distributions of 95.2% of predictors (adjusted p<0.0002), including linear, monotone nonlinear, and non-monotone changes (see Figure 16 for details of six exemplar predictors). Shifts were generally small in magnitude. For example, although a statistically significant linear trend in age was recorded across the study period, the overall increase in mean age was less than 1 years. The largest changes we observed were in the proportion of admissions involving patients receiving certain medications and diagnosed with chronic diseases (see Table 7 and Appendix C). For example, admissions to patient with a history of hypertension increased from 55.0% in 2003 to 76.7% in 2012, and admissions to

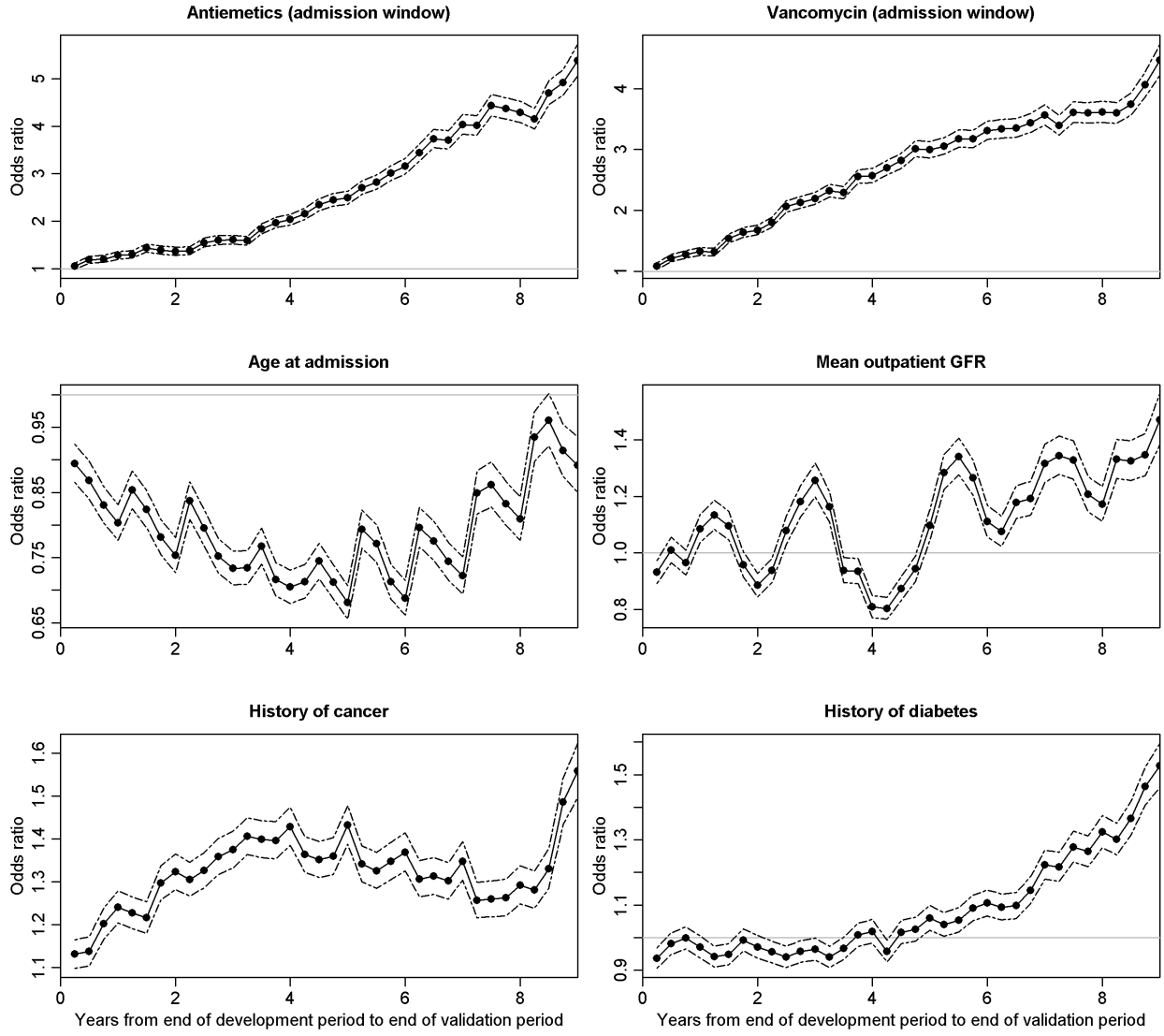
Figure 16. Distributions of select AKI predictors across over time. Continuous predictors summarized as means, categorical variables summarized as proportions. Red lines indicate fitted, smoothed values.



patients with diabetes increased from 29.7% to 42.6% over the same period. Similarly, patients receiving statins during the admission window increased from 27.9% in 2003 to 44.0% in 2012, and patients receiving vancomycin during the admission window increased from 5.3% to 16.3%. Six predictors – use of antifungals and monoamine oxidase inhibitors (MAOIs) during the 90 days prior to admission and use of non-steroidal anti-inflammatory drugs (NSAIDs), MAOIs, anhydrase diuretics, and lithium during the admission window – had stable distributions over time.

We also observed changes the adjusted contributions of predictors to discriminating between development and validation admissions. Odds ratios from the logistic membership models indicated significant and temporally varying strengths of associations between the

Figure 17. Odds ratios and 95% confidence intervals of select predictors from AKI logistic membership models. For continuous variables, odds ratios indicate effect of one IQR change in value. Grey lines indicate null odds ratio of 1.



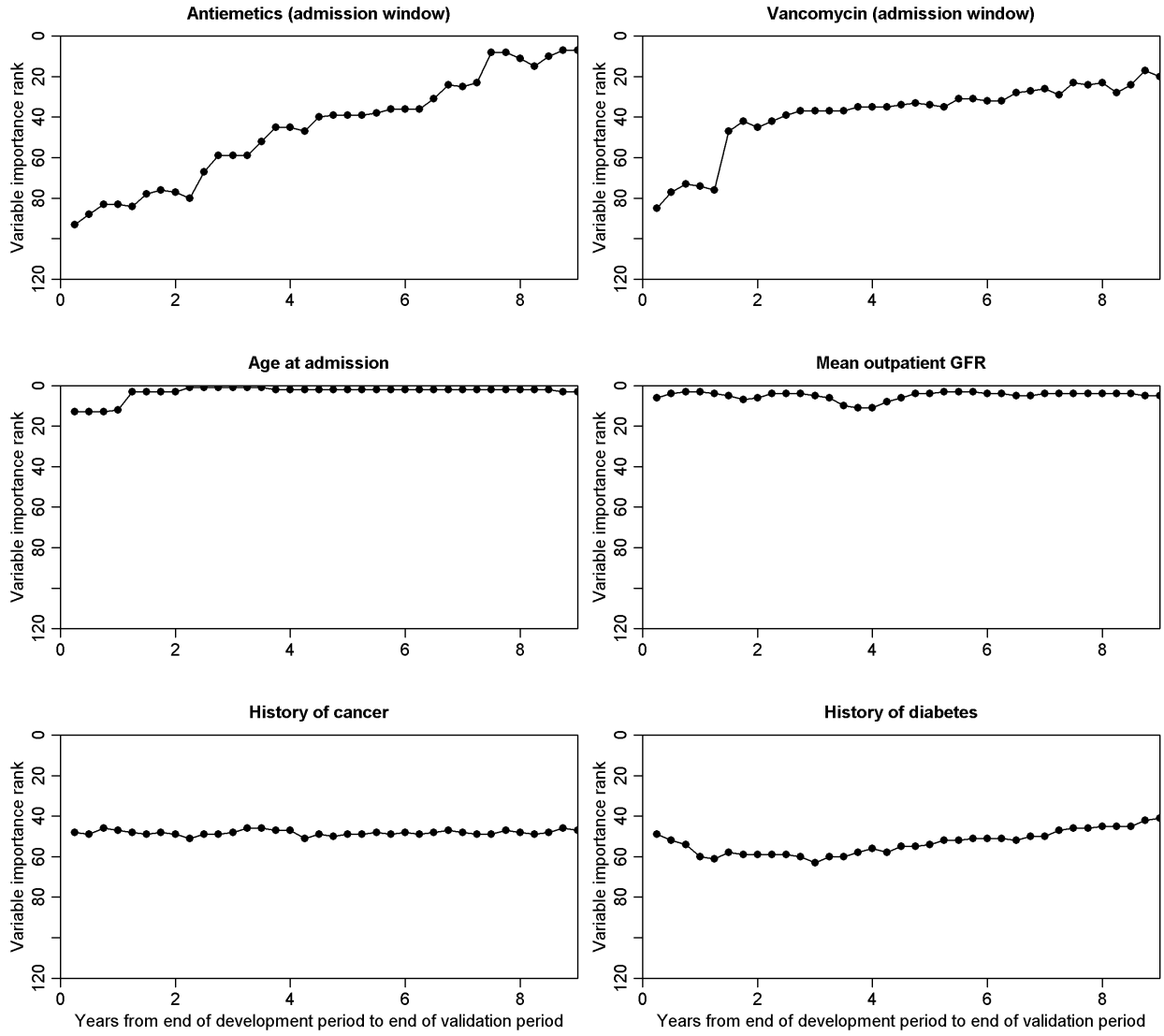
majority of predictors and whether admissions belonged to the validation or development cohorts (see Figure 17 for odds ratios of the six exemplar predictors). Temporal changes in membership model odds ratios did not consistently follow the temporal trends observed in the unadjusted distributions of individual predictors. For example, while the mean age at admission increased across the 9-year validation period, the odds ratio exhibited increasing strength of association between age and membership in the validation versus development cohort over the first five years and subsequently a decreasing strength of association over the remaining four years. The proportion of admissions with a history of diabetes increased almost linearly over time; however, the membership model odds ratio for this predictor was stable and non-significant over the first 3 years of the validation period. On the other hand, use of antiemetics during the admission window both increased in frequency and was associated with increased odds of being a validation observation over time. Among the six predictors with stable individual variable distributions, we observed stable logistic membership model odds ratios.

The relative importance of predictors to discriminating between validation and development admissions varied over time (see Figure 18 for details on the six exemplar predictors). In random forest membership models, we observed temporal changes in variable importance rankings for 96 of 121 predictors (79.3%; $p<0.0002$). Changes in variable importance ranks were generally small in magnitude (mean change=17, interquartile range= 10-19). Predictors with the greatest changes in importance rank included use of vancomycin, antiemetics, and fluoroquinolones during the admission window. The relative importance of use of vancomycin and antiemetics during the admission window increased from 85 to 17 and 93 to 7, respectively. The importance of antiemetics use increased steadily over time, while the vancomycin use experienced a sharp increase in importance during the second year of the validation period, followed by a more gradual increase in rank over the remaining years of the validation period. The variable importance rank of fluoroquinolones use during the admission window was stable in in the mid-30s for two years and slowly declined to 89 over the next seven years. A history of dyslipidemia was ranked as the 37th most important variable for the first 3-month validation cohort. The predictor's rank increased over the next 3 years, with a history of dyslipidemia becoming one of the three most important predictors by 3 years and being ranked the most important predictor during 5 validation cohorts. With the exception of use of NSAIDs during the admission window, no trends were observed in the variable importance ranks of the six predictors with stable distributions over time.

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. No temporal changes were observed for the majority of predictors. For those predictors with strongest and/or most changed predictor-outcome associations, we present results for four time points in Table 10. In addition, we present more detailed graphical results for the same six exemplar predictors included in the detailed case mix shift results above – use of antiemetics and vancomycin during the admission window, age, mean outpatient GFR prior to admission, history of cancer, and history of diabetes.

Figure 18. Variable importance ranks of select predictors from AKI random forest membership models. Higher ranks (smaller numbers) indicate greater importance to model performance.



In logistic regression models refit in each validation cohort, we observe few significant temporal changes in the strength of predictor-outcome associations (see Table 10 and Figure 19). Compared to odds ratios from the original model based on 2003 data, we observed tendencies toward temporal strengthening or weakening associations for a limited number of predictors. However, these changes generally did not reach statistical significance, with the exception of a few isolated 3-month validation periods. This pattern was most pronounced for use of vancomycin during the admission window, for which the odds ratio increased starting at approximately three years into the validation period and was significantly higher than the odds ratio of the original model for 7 of the last 12 validation cohorts. In the original model and for most of the first half of the validation period, patient age at admission and a history of diabetes were significantly associated with AKI. Declines in the odds ratio for these predictors beginning

Table 10. Select predictor-outcome associations in AKI models refit over time. Odds ratios from logistic regression models (LR OR), variable importance ranks from random forest models (RF rank), and proportion of times predictors was selected in 200 L-1 penalized logistic regression (L1) modeling iterations based on models fit at development and within select temporal validation cohorts.

	Development (2003)			2006 – Q4			2009 – Q4			2012 – Q4		
Predictor	**LR OR* (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 [1.20, 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, 1.10]	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, 1.30]	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

Table 10. (continued) Select predictor-outcome associations in AKI models refit over time. Odds ratios from logistic regression models (LR OR), variable importance ranks from random forest models (RF rank), and proportion of times predictors was selected in 200 L-1 penalized logistic regression (L1) modeling iterations based on models fit at development and within select temporal validation cohorts.

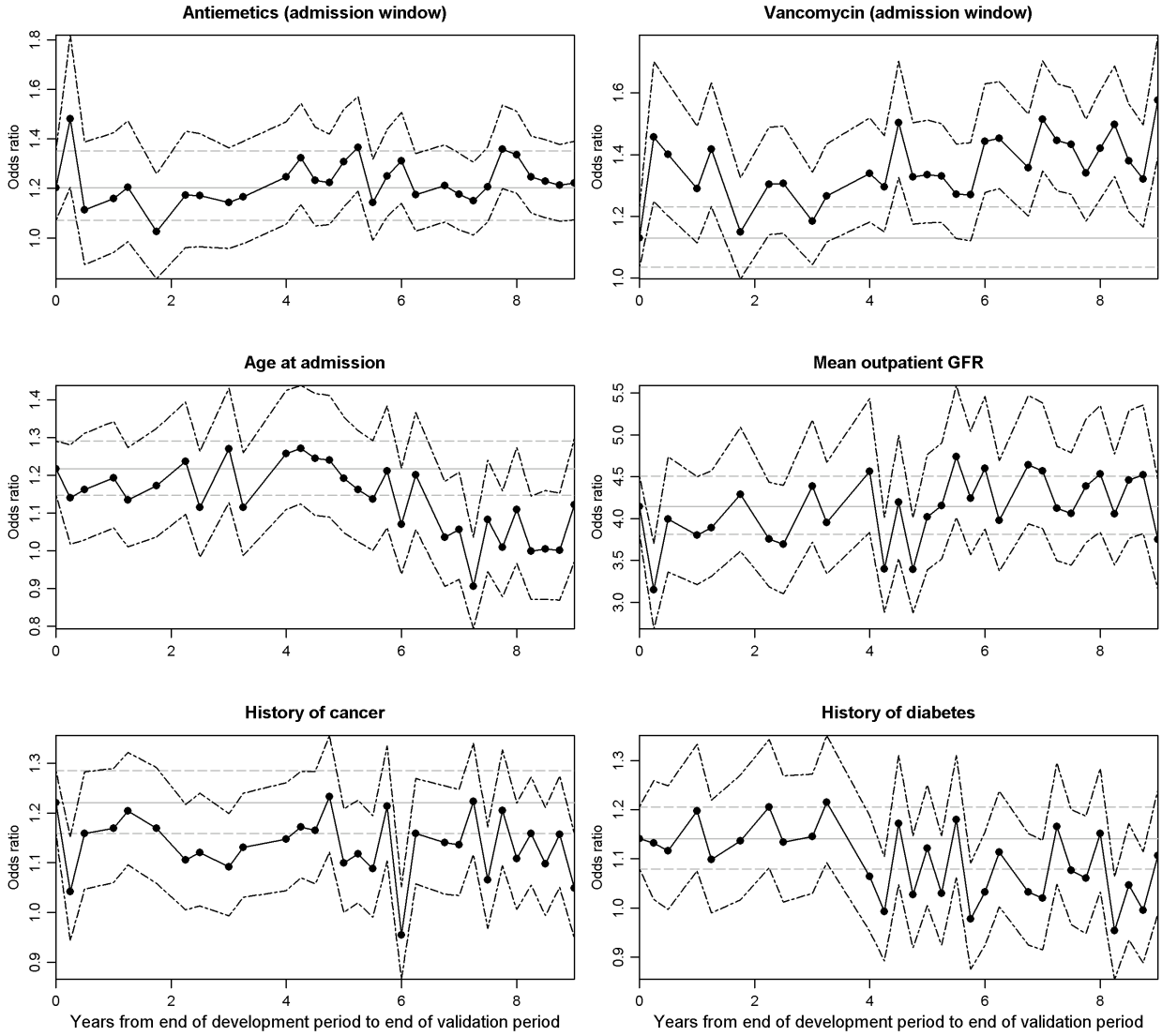
	Development (2003)			2006 – Q4			2009 – Q4			2012 – Q4		
Predictor	**LR OR* (95% CI)**	**RF Rank**	**L1 % selected**	**LR OR (95% CI)**	**RF Rank**	**L1 % selected**	**LR OR (95% CI)**	**RF Rank**	**L1 % selected**	**LR OR (95% CI)**	**RF Rank**	**L1 % selected**
History of diabetes mellitus	1.14 [1.08, 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 in 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 [0.90, 1.13]	65	47
CCB in 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, 1.20]	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
Lactate ringers 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
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 in 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

* Odds ratios for continuous predictors are for an interquartile range increase in value

Abbreviations::CCB=calcium channel blocker; COPD=chronic obstructive pulmonary disease; IVF=intravenous fluids

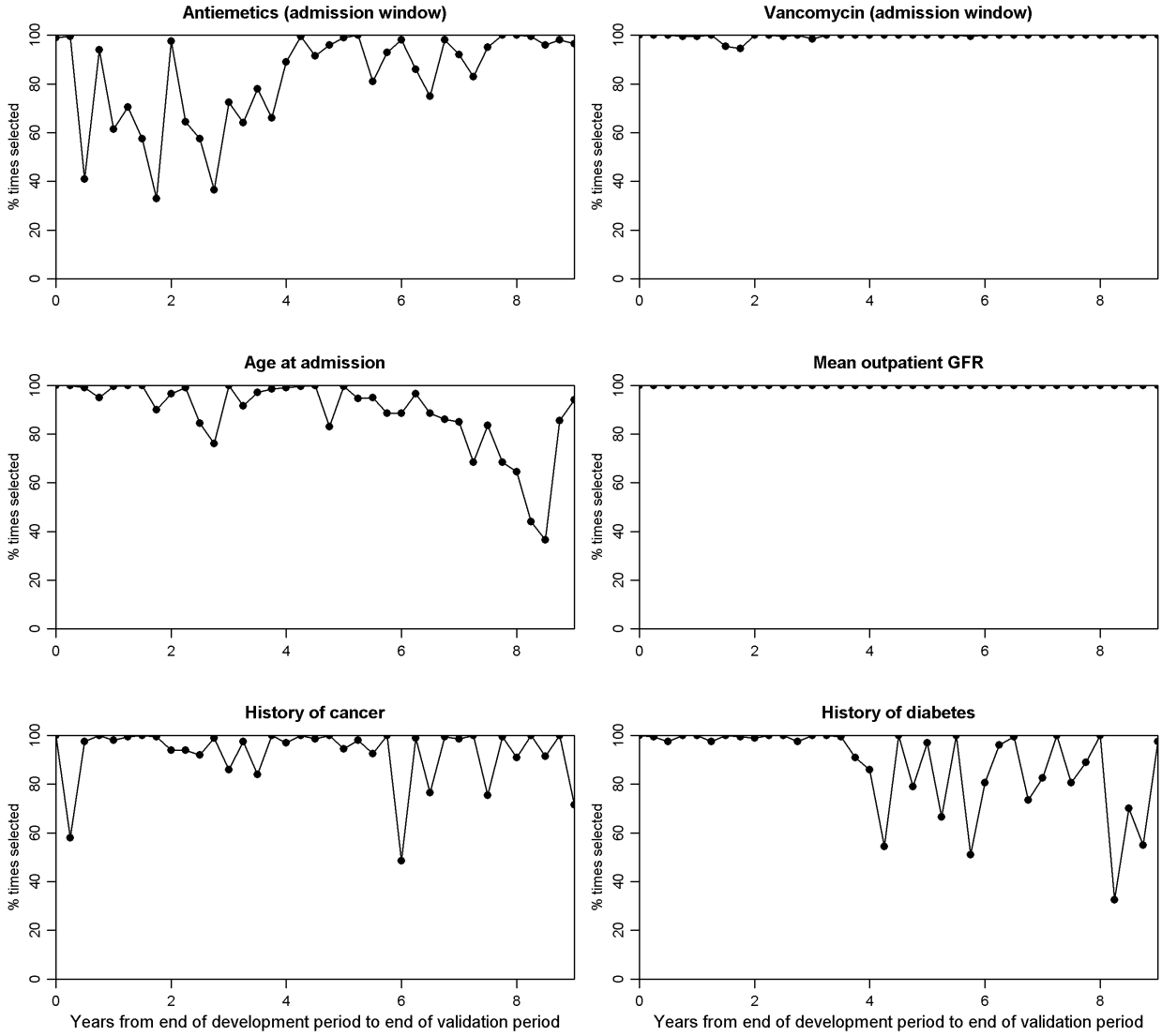
at validation year four resulted in the associations becoming non-significant; however, confidence intervals for the refitted odds ratios overlapped with confidence intervals of the original model in all but one validation cohort.

Figure 19. Odds ratios and 95% confidence intervals of select predictors from AKI logistic models refit over time. For continuous variables, odds ratios indicate effect of one IQR change in value. Grey lines indicate initial odds ratio and confidence interval.



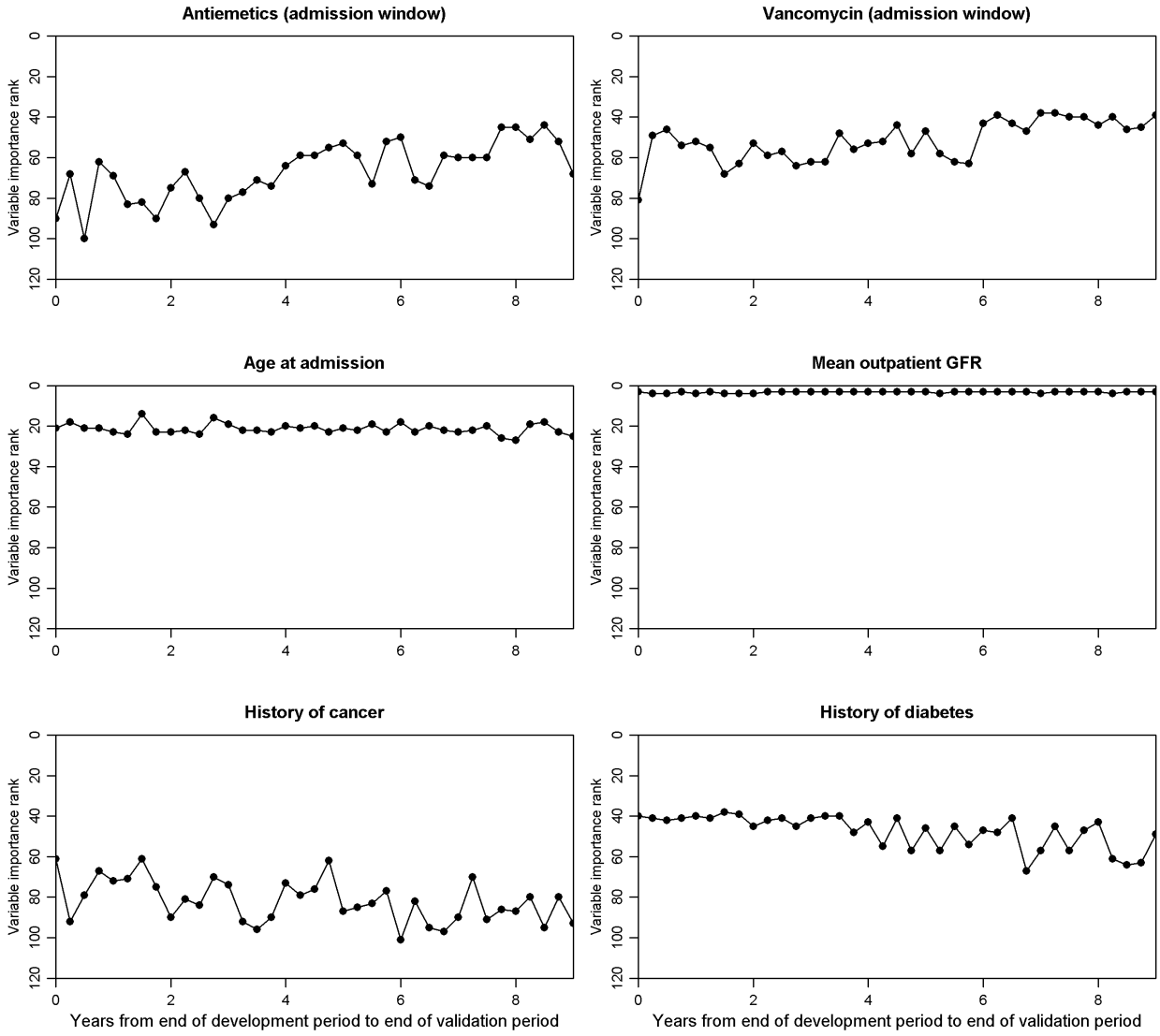
In L-1 penalized logistic regression models refit over time, we observed no temporal changes in selection patterns of the majority of predictors (see Figure 20 and Table 10). Six predictors—use of loop and thiazide diuretics in the admission window, mean and change in GFR in the admission window, mean outpatient GFR prior to admission, and patient race—were consistently selected for inclusion in the model in the development and every validation cohort.

Figure 20. Variable selection of select predictors from AKI L-1 penalized logistic regression models refit over time. Proportion of bootstrap iterations (B=200) in which select predictors were retained.



Use of vancomycin during the admission window, a history of hypertension, and hypertension at admission were selected in at least 90% of L1 modeling iterations in the development and every validation cohort. Two predictors—age at admission and use of fluoroquinolones during the admission window—exhibited significant temporal changes in the frequency of selection (adjusted $p<0.0002$). Age at admission became less frequently selected during the second half of the validation period. Selection frequency of use of fluoroquinolones during the admission window declined by three years into the validation period, followed by small increases in selection frequency over the remainder of the validation period. Although not statistically significant, we noted temporal changes in the variability of selection of additional predictors, such as history of diabetes and use of antiemetics during at the admission window (see Figure

Figure 21. Variable importance ranks of select predictors from AKI random forest models refit over time. Higher ranks (smaller numbers) indicate greater importance to model performance.



20), with the changes noticeably separating patterns of the first and second halves of the validation period.

Based on RF models refit in each temporal validation cohort, 11 predictors (9.3%) experienced significant temporal changes in variable importance ranking (adjusted p<0.0002; see Table 10 and Figure 21). The relative strength of association between AKI and admission GFR count, use of antiemetics in the 90 days prior to admission, lactate ringer intravenous fluids, and use of antiemetics, vancomycin, and penicillins during the admission window increased over time as the variable importance rank of each predictor increased over the validation period. We observed reduced variable importance ranks over time for history of dyslipidemia, use of ACEi and calcium channel blockers in the 90 days prior to admission, history of diabetes, and history of chronic obstructive pulmonary disease. The observed

temporal changes in variable ranks were not consistent over time. For example, the relative importance of history of diabetes was stable around 40th for almost four years before becoming more variable and trending toward lower ranks. Mean GFR during the admission window ranked as the most important variable in the original model and for models in every validation cohort.

Conclusions

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 one 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 did not result in practically meaningful changes in AUCs. For the most stringent calibration measures, machine learning models 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. Interpretation of these findings and their implications for model updating are discussed and integrated with the findings of our corresponding 30-day mortality analysis in Chapter 6.

CHAPTER 5

TEMPORAL EVALUATION OF MODELS PREDICTING 30-DAY ALL-CAUSE MORTALITY AFTER HOSPITAL ADMISSION

To more fully understand the impact of modeling methods on performance over time, we extended our analysis beyond the acute kidney injury study and performed corresponding analysis among models for 30-day all-cause mortality after hospital admission. Applying the same seven modeling methods – logistic regression, L-1 penalized logistic regression, L-2 penalized logistic regression, L-1/L-2 penalized logistic regression, naïve Bayes, neural networks, and random forests – we observed multiple patterns of calibration drift that both reinforced and expanded on the findings of our AKI analysis. The combinations of data shifts in our mortality study population also differed from those observed in our AKI population, further making this analysis complementary to the findings described in the previous chapter.

Nationwide, 3,467,142 admissions to VA facilities met all eligibility criteria for our 30-day all-cause mortality analyses. Restricting to the 50% of sites randomly selected by stratified Veterans Integrated Service Network, our analysis included 1,893,284 admissions (54.6% of all eligible admissions), 235,548 in the 2006 development cohort and 1,657,736 in the 7-year validation period. The final validation cohort (i.e., 2013-Q4) was smaller than the other validation cohorts (n=37,442) as it was restricted to admissions beginning on or before December 1, 2013 to allow for sufficient follow-up time for outcome ascertainment. The remaining 27 validation cohorts consisted of 60,011 admissions on average (range 57,367 – 62,139). A brief summary of the patient population at select points across the study period is presented in Table 11 (See Appendix D for details for all predictors). Admitted patients were primarily male (95.0%), white (72.1%), in their early 60s (mean age: 63.4; standard deviation: 14.0), and diagnosed with at least one chronic medical condition (93.9% with one diagnosis, 86.8% diagnosed with multiple conditions). Overall, the 30-day all-cause mortality rate after admission was 4.9%.

Model Development

Modeling Parameters

Table 12 provides the hyperparameter values selected through 5-fold cross validation for the L-1/L-2 penalized logistic regression, random forest, and neural network models. These hyperparameters were used for all internal validation bootstrap iterations. The shrinkage parameters (lamba), however, for the three penalized regression models were selected during each bootstrap iteration.

Table 11. 30-Day mortality patient population at development (2006) and in three years of the validation period (2007, 2010, 2013)

	2006	2007	2010	2013
N	235,548	235,734	243,631	214,798
% 30-day mortality	5.0	4.9	4.9	4.7
Age in years (mean and SD)	62.9 (13.7)	63.0 (13.8)	63.6 (14.0)	63.9 (14.3)
% Female	4.5	4.7	4.9	5.5
Race				
% White	71.7	71.6	72.3	72.1
% Black	19.8	20.0	19.6	19.8
% American Indian/Alaskan	1.3	1.4	1.5	1.6
% Asian/Pacific Islander	1.1	1.2	1.2	1.3
% Unreported	6.0	5.9	5.5	5.3
BMI at admission (mean and SD)	28.2 (7.1)	28.3 (7.3)	28.7 (7.2)	28.8 (7.1)
Health care utilization (prior year)				
Inpatient visits (mean and SD)	1.3 (2.0)	1.3 (2.0)	1.3 (2.0)	1.3 (2.1)
Outpatient visits (mean and SD)	36.4 (43.6)	37.1 (43.3)	42.0 (48.2)	43.5 (48.9)
% Unplanned readmission	10.3	10.5	10.5	10.4
Select diagnoses (preadmission)				
Cardiac arrhythmias	8.4	10.6	15.7	19.3
Chronic pulmonary disease	28.4	32.4	38.5	41.2
Congestive heart failure	17.3	19.1	22.0	23.7
Depression	20.1	24.5	32.6	38.4
Drug abuse	12.4	14.9	18.8	22.0
Dyslipidemia	41.8	49.6	61.5	66.4
Fluid and electrolyte disorders	19.0	23.9	32.8	38.1
Hypertension	61.7	67.4	74.2	76.4
Renal failure	12.3	15.3	19.5	21.9

Abbreviations: SD=standard deviation

Table 12. Hyperparameters selected for 30-day mortality models

Model	Hyperparameter	Value
L1-L2	Elastic-net mixing parameter (α)	0.2
Random forest	# predictors considered per tree	6
	Minimum # observations per node	5
	# trees	600
Neural network	Size of hidden layer	51

Initial Model Performance

Initial performance of each of the seven models is presented in Table 13. For both measures of accuracy, performance was similar for the four regressions and the RF model, and

Table 13. Initial 30-day mortality model performance in development cohort

	Regression				Machine Learning		
	LR	**L1**	**L2**	**L1-L2**	**RF**	**NN**	**NB**
Accuracy							
Scaled Brier score	0.131 [0.131, 0.132]	0.128 [0.128, 0.129]	0.124 [0.124, 0.125]	0.129 [0.128, 0.129]	0.119 [0.119, 0.119]	0.073 [0.072, 0.074]	*
Nagelkerke's R^2	0.253 [0.252, 0.254]	0.248 [0.247, 0.249]	0.241 [0.241, 0.242]	0.248 [0.247, 0.249]	0.220 [0.219, 0.221]	0.171 [0.170, 0.172]	*
Discrimination							
AUC	0.847 [0.846, 0.847]	0.844 [0.844, 0.844]	0.842 [0.841, 0.842]	0.844 [0.844, 0.844]	0.834 [0.833, 0.834]	0.794 [0.794, 0.795]	0.768 [0.768, 0.769]
Calibration							
O:E ratio	0.998 [0.996, 1.001]	0.998 [0.996, 1.001]	0.998 [0.996, 1.001]	0.998 [0.996, 1.001]	0.929 [0.927, 0.931]	0.997 [0.993, 1.000]	0.339 [0.338, 0.340]
Cox intercept	-0.048 [-0.055, -0.042]	0.088 [0.080, 0.096]	0.215 [0.207, 0.223]	0.096 [0.088, 0.105]	0.074 [0.065, 0.082]	-0.122 [-0.133, -0.112]	-2.548 [-2.551, -2.546]
Cox slope	0.980 [0.977, 0.982]	1.039 [1.036, 1.042]	1.093 [1.090, 1.096]	1.043 [1.040, 1.046]	1.072 [1.069, 1.076]	0.951 [0.947, 0.955]	0.113 [0.113, 0.114]
ECI	0.013 [0.013, 0.014]	0.010 [0.010, 0.010]	0.011 [0.010, 0.011]	0.010 [0.010, 0.010]	0.034 [0.033, 0.034]	0.008 [0.007, 0.008]	7.783 [7.758, 7.808]

* Non-calculable due to extreme predicted probabilities of 0 and 1.

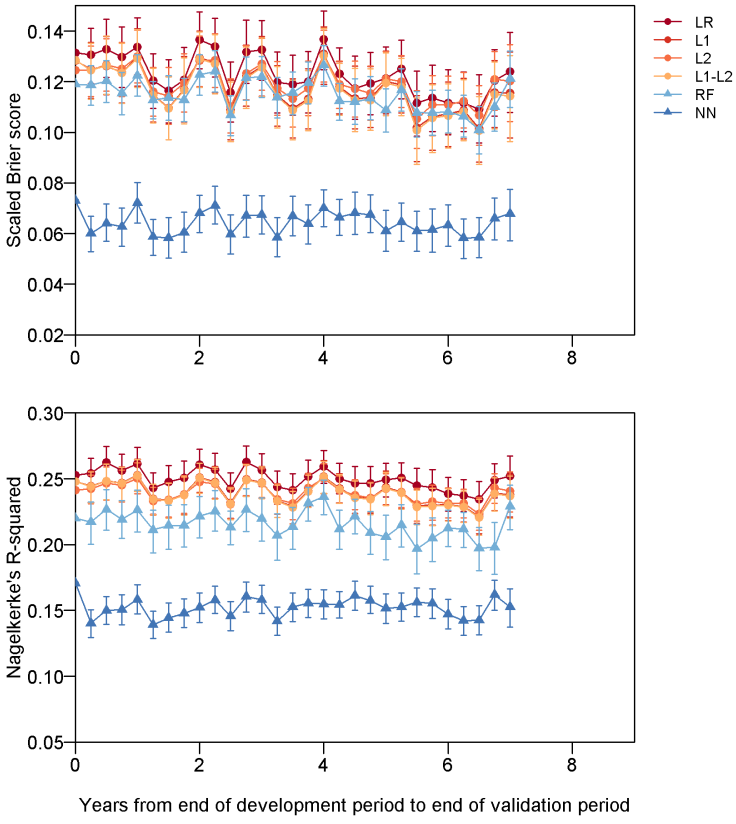
lower for the NN model. Discrimination was generally good, with AUCs ranging from 0.77 to 0.85. The NN and NB models had slightly lower AUCs that the regression and RF models. The regression models and the NN model were calibrated based on O:E ratios and ECIs. The RF model, with an O:E ratio of 0.93 (95% CI: 0.93, 0.93), slightly overpredicted on average. No models were perfectly calibrated according to Cox intercepts and slopes; however, these metrics generally approached their ideal values and indicated similar levels of over/underpredicting and over/underfitting across models. The NB model lacked accuracy and calibration as measured by all metrics.

Model Performance Over Time

Accuracy

Accuracy was stable over time for the L2, RF, and NN models, and declined slightly for the LR, L1, L1-L2, and NB models (adjusted p<0.001; see Figure 22). Temporal changes in the

Figure 22. Accuracy of 30-day mortality models over time by modeling method. Regression models are displayed with circular markers and red-orange colors. Machine learning models 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.

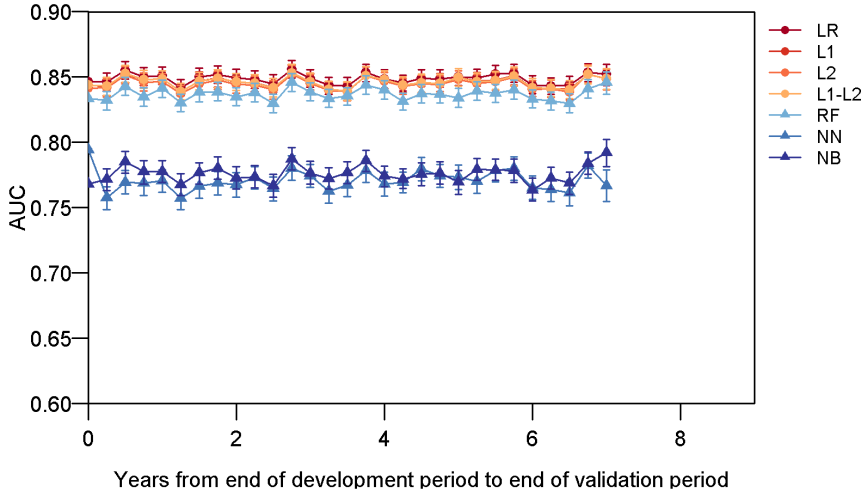


accuracy of the LR, L1, and L1-L2 regression models were observed for the scaled Brier score but not the Nagelkerke's R^2. For these models, overall changes in the scaled Brier score were small in magnitude. For example, the rate of change for the LR model was -0.002 (95% CI: -0.004, -0.001) and resulted in an overall change from 0.131 (95% CI: 0.131, 0.132) to 0.124 (95% CI: 0.108, 0.139). The accuracy of the RF and regression models was similar, initially and over time. On both accuracy measures, the NN model exhibited lower accuracy than all other models, with the exception of the NB model, across the entire study period. The NB model consistently underperformed all other models based on both the scaled Brier score and Nagelkerke's R^2.

Discrimination

We observed stable discrimination for all models over the seven-year validation period (adjusted p>0.001; see Figure 23). The regression and RF models had comparable AUCs and maintained higher discrimination than the NB and NN models. The NB and NN models exhibited similar levels of discrimination, with significant differences that were small in magnitude at a couple time points.

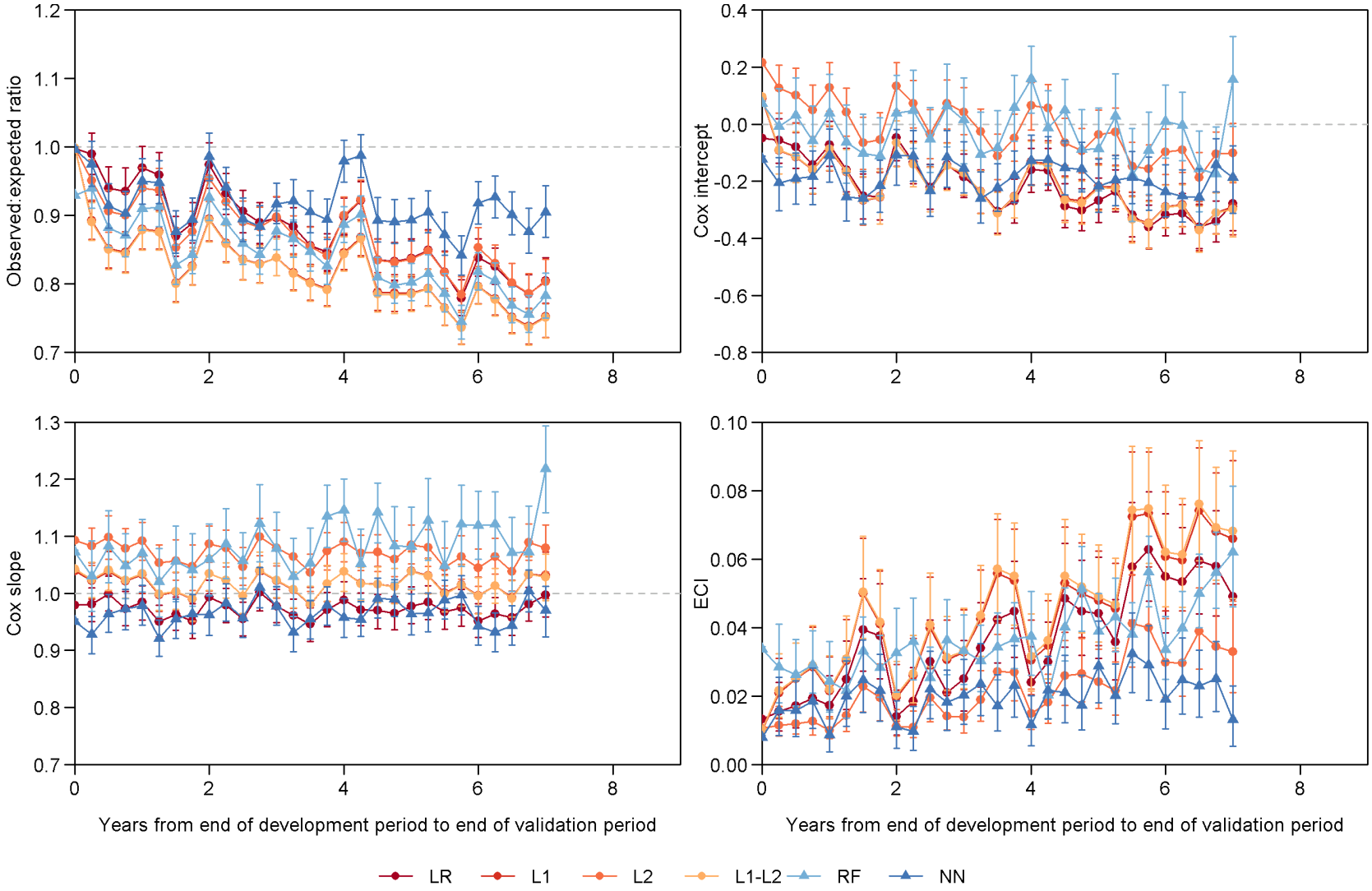
Figure 23. Discrimination of 30-day mortality models over time by modeling method. Regression models are displayed with circular markers and red-orange colors. Machine learning models are displayed with triangular markers and blue-purple colors.



Calibration

We observed calibration drift for all models, with the magnitude and pattern of drift varying by modeling method and calibration metric (see Figure 24). The NB model substantially underperformed all other models in terms of calibration due to extreme predictions, and thus the performance of this model is not considered further.

Figure 24. Calibration of 30-day mortality models over time by modeling method. Regression models are displayed with circular markers and red-orange colors. Machine learning models 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.



O:E ratios declined immediately after development for all models and across the study period for all models except the NN, indicating increasing average overprediction. At the first quarterly validation period, the O:E ratio included the null values of 1.0 for only the NN and LR models. These two models achieved calibration according to the O:E ratio at 2-4 additional time points, however, overpredicted on average for most of the validation period. The trajectory of O:E ratios was similar for most models, with the exception of the NN model which did not exhibit a significant slope in the O:E ratio over time (adjusted $p<0.001$ for all models except the NN model). The NN model demonstrated significantly less overprediction than the RF and regression models, particularly in the last three years of the validation period. In addition to the overall drift, a seasonal pattern was apparent in the O:E ratios. In the first and fourth quarters of most validation years, O:E ratios peaked for all models.

Assessments of weak calibration also documented different patterns of drift across methods. Cox intercepts declined across the validation period for each of the regression models (adjusted $p<0.001$), indicating increasing overprediction, while remaining stable for the RF and NN models. The L2 regression model exhibited a smaller decline over time in the Cox intercept than the other regression models and did not systematically over or underpredict in 16 of 28 validation cohorts. The RF did not systematically over or underpredict in the majority of validation cohorts (24 of 28), and the NN model systematically overpredicted to a stable degree for the entire study period. Additionally, we observed a seasonal pattern in Cox intercepts similar to that of the O:E ratios, although to a lesser degree. Cox slopes were stable over time (adjusted $p>0.001$). No significant overfitting was observed for the LR, L1, and L1-L2 regression models. The L2 regression and RF models exhibited some underfitting (i.e., Cox slope>1.0). This underfitting was consistent over time for the L2 model and demonstrated a nonsignificant tendency toward increasing for the RF model. The NN model had Cox slopes less than 1.0, indicating overfitting, for 13 of the 28 validation cohorts; however, there was no significant change in the level of Cox slopes over time for the NN model.

While the regression and RF models experienced calibration drift at the moderate level of calibration, the NN model exhibited stable overall moderate calibration with some seasonal variation. ECIs of the regression and RF models increased across the validation period (adjusted $p<0.001$), indicating declining calibration, and exhibited no changes in the trajectory or rate of ECI drift during the seven years. The L2 regression model experienced a smaller magnitude of drift in the ECI compared to the other regression models and a similar magnitude of drift to the RF model. Compared with the L2 regression and RF models, the rate of change in ECI was 50% higher for the LR regression model (0.006 [95%CI: 0.005, 0.008] vs 0.004 [95% CI: 0.003, 0.005] and 0.004 [95% CI: 0.002, 0.005] for the LR vs L2 and RF, respectively) and 75% higher for the L1 and L1-L2 regression models (0.007 [95%CI: 0.006, 0.009] for both). For each model, seasonal corrections of the ECI were observed in the first and fourth quarters of each year. ECIs were stable over time for the NN model (adjusted $p>0.001$). Although not significantly different from the surrounding time periods, in most validation years, the ECI of the NN model was markedly lower (i.e., closer to the ideal value of 0) during the first and fourth quarters.

The ranges of predicted probabilities and proportion of admissions over which each model was calibrated also changed over time and varied by modeling method (see Figures 25 and 26). With the exception of the NB model, which strongly overpredicted for most predicted

Figure 25. Regions of calibration of 30-day mortality models over time by modeling method. Areas of overprediction and underprediction are shaded based on based on within-region estimated calibration index (ECI) in order to highlight the magnitude of miscalibration.

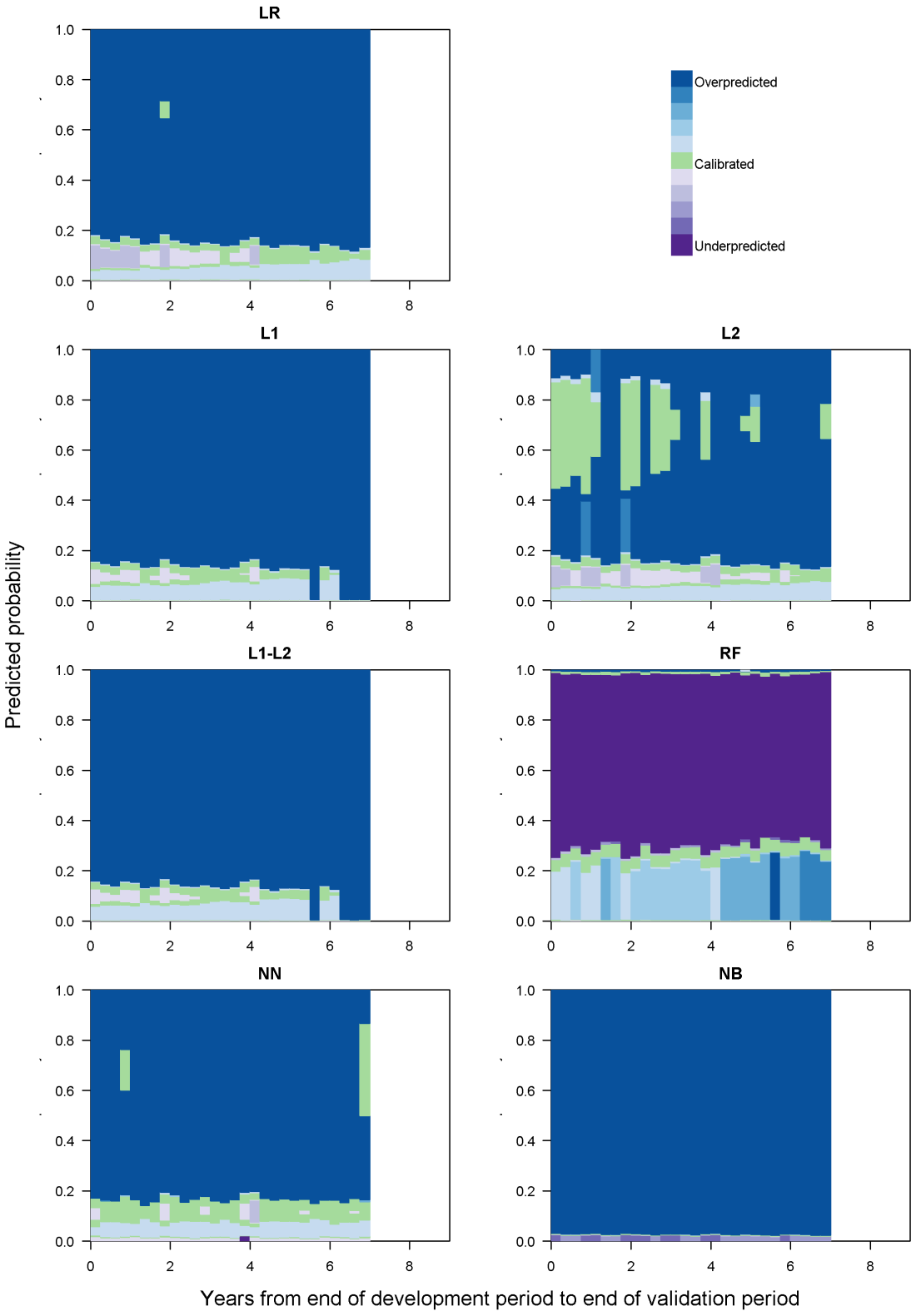
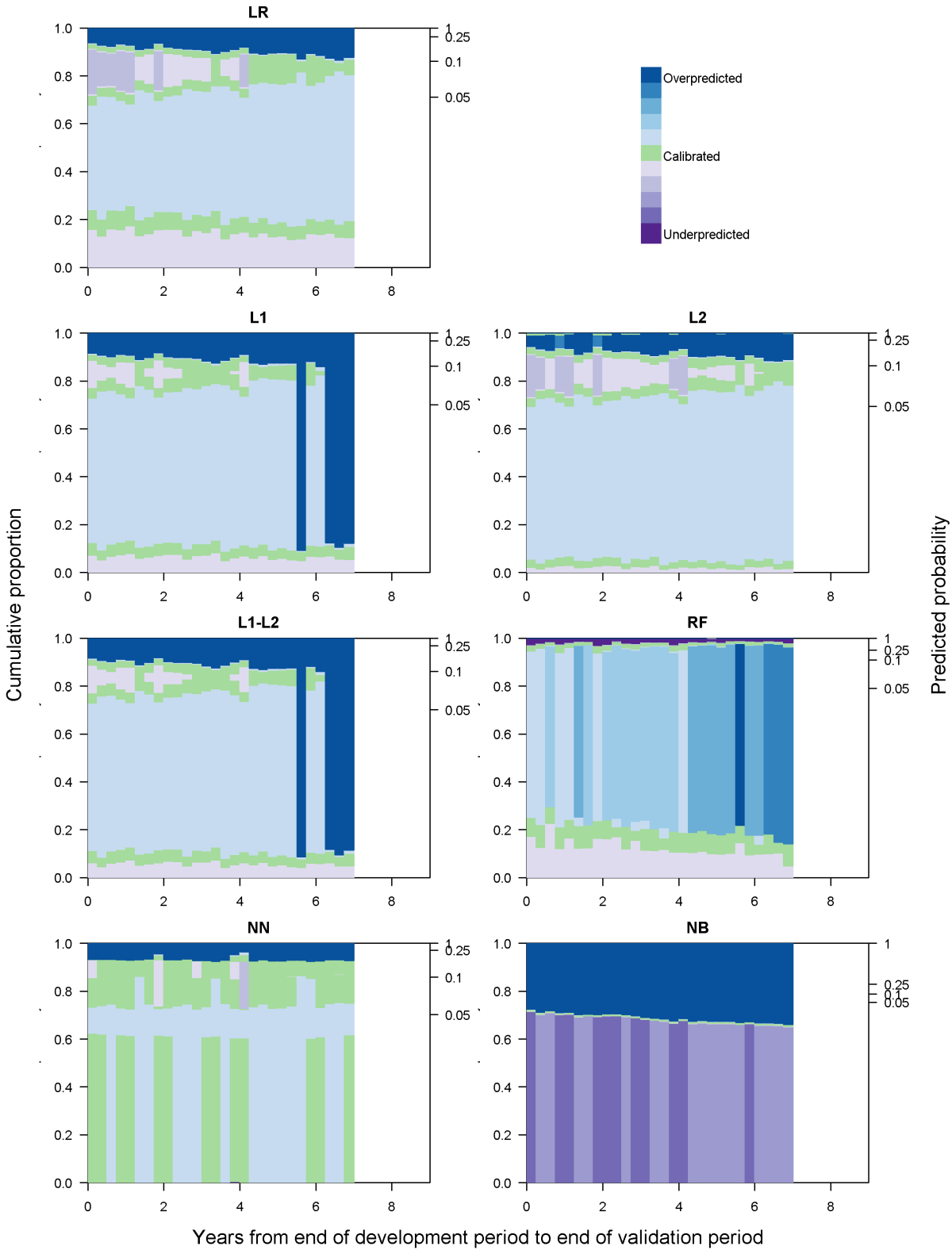


Figure 26. Proportional volume assessment regions of calibration of 30-day mortality models over time by modeling method. Regions are scaled by proportion of observations in each region and shaded by the magnitude of the within region estimated calibration index (ECI).



probabilities, each model moved in and out of regions of calibration, overprediction, and underprediction across the range of predicted probability. The RF model was the only model with a large range of probabilities over which the model strongly underpredicted, as highlighted by the darkest purple shade dominating the RF panel of Figure 25. During the first half of the validation period, the L2 regression model tended to be calibrated for predictions in the 50% to 90% ranges of risk. The remaining models and the L2 regression model during the second half of the validation period tended to strongly overpredict for predicted probabilities starting at approximately 20% risk. For each regression model, across the validation period the proportional volume assessment indicated that nearly half of admissions were in regions of overprediction, with the proportion increasing slowly over time. The majority of these admissions were minimally overpredicted, falling in areas with regional ECI values near the ideal value of 0, as highlighted by the lightest blue shades in Figure 26. For example, in the LR model, a low risk region of overprediction captured at least 40% of admissions in each validation cohort and had a mean ECI of 0.005 (range: 0.002 – 0.009). Each regression model also experienced growth in volume of an overpredicted region with a larger magnitude of miscalibration among higher predicted probabilities. For the L1 and L1-L2 regression models, the vast majority of admissions were strongly overpredicted in the last year of the study period. The overpredicted region of the RF model captured a slowly growing proportion of admissions over time and increased modestly in the magnitude of overprediction across the validation period. The proportional volume assessment highlighted a seasonal pattern in the calibration of the NN model for admissions with predicted probabilities under 3%. On average, 61.3% of all admissions were in this low risk region for which the NN model was calibrated during the first and fourth quarters of most years and minimally overpredicted during the second and third quarters.

Data Shifts Over Time

Event Rate Shift

Over the seven-year validation period, there was a statistically significant decline in the 30-day mortality rate (adjusted $p<0.0003$). The event rate declined from 5.0% in the 2006 development year to 4.8% in the final validation period (see Figure 27). Seasonal changes within each validation year were three times larger than the overall change in the mortality rate (mean within year change: 0.6%; overall change: 0.2%). Two years after development, the mortality rate was a significant predictor in logistic membership models discriminating between development and validation admissions (see Figure 28). The odds ratios for mortality in these models decreased over time from 0.95 (95% CI: 0.91, 0.99) in the first quarter of 2009 (i.e., two years from development) to 0.81 (95%CI: 0.76, 0.87) in the final validation cohort. While this overall change accrued over five years, we observed an annual pattern of odds ratios generally moving toward the null during the 1st and 4th quarters. The relative importance of 30-day mortality in the random forest membership models changed over time (adjusted $p<0.0004$); however, this small change occurred over the first validation year in which the rank changed from 59 to 63, and the variable importance rank was stable at a mean of 63 (range: 62 – 64) out of 67 over the next six years.

Figure 27. Proportion of admissions resulting in 30-day mortality over time. Red lines indicate fitted values.

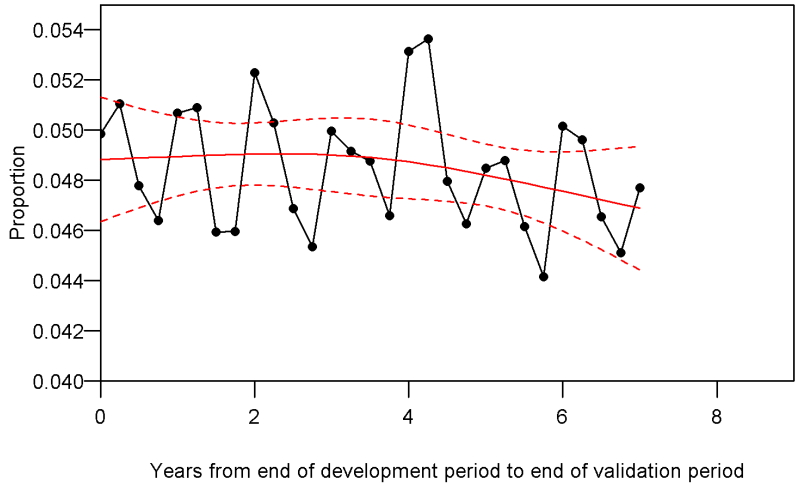
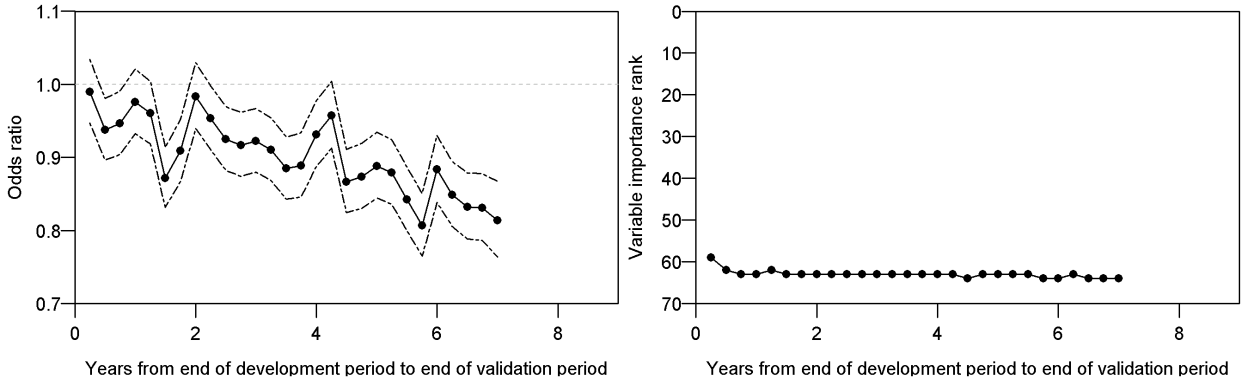


Figure 28. Membership model results for 30-day mortality. Odds ratio (left) and 95% confidence intervals for 30-day mortality based on logistic membership models fit for each 3-month temporal validation cohort. Grey line indicates null odds ratio of 1. Variable importance rank (right) for 30-day mortality based on random forest membership models fit for each 3-month temporal validation cohort. Higher ranks (smaller numbers) indicate greater importance to model performance.



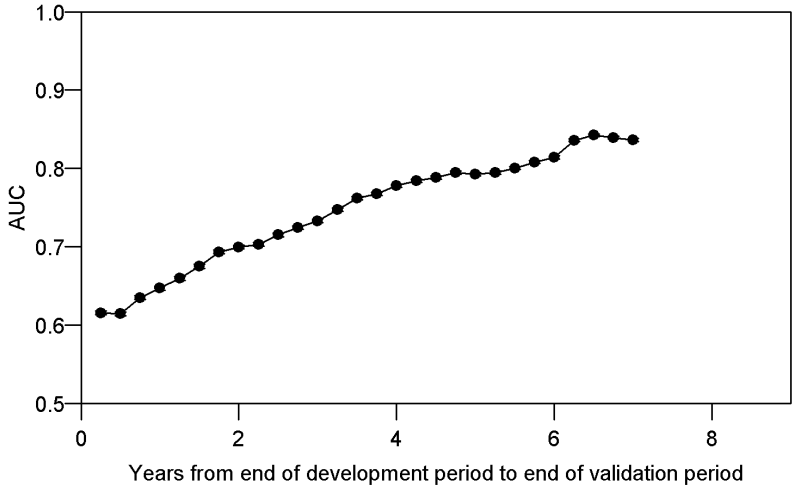
Case Mix Shift

We observed case mix shift across the validation period. Distributions of predicted probabilities indicated changes in severity and heterogeneity of risk in the patient population over time (adjusted $p<0.002$). Predictions from the NN model were the exception, exhibiting no change in the variability of predictions. Predictions from all models indicated an increase in mean patient severity over time. The standard deviation of the predictions from regression models increased 4-5% over the seven validation years, indicating more heterogeneity of risk among patients. The RF model, on the other hand, indicated a 5% decrease in heterogeneity of risk. Membership models also noted the presence of overall case mix shift. The logistic

membership models increasingly discriminated between admissions from the development and each sequential validation cohort, with AUCs increasing from 0.616 (95% CI: 0.613, 0.618) to 0.836 (95% CI: 0.834, 0.839) (see Figure 29).

For predictor-level assessments of case mix shift, we present detailed graphical results for six exemplar predictors – history of liver disease, depression, and dyslipidemia, as well as the most recent sodium, mean corpuscular hemoglobin concentration (MCHC), and serum creatinine values recorded during the admission window. Findings for these predictors highlight common data shift patterns observed among the full set of predictor variables.

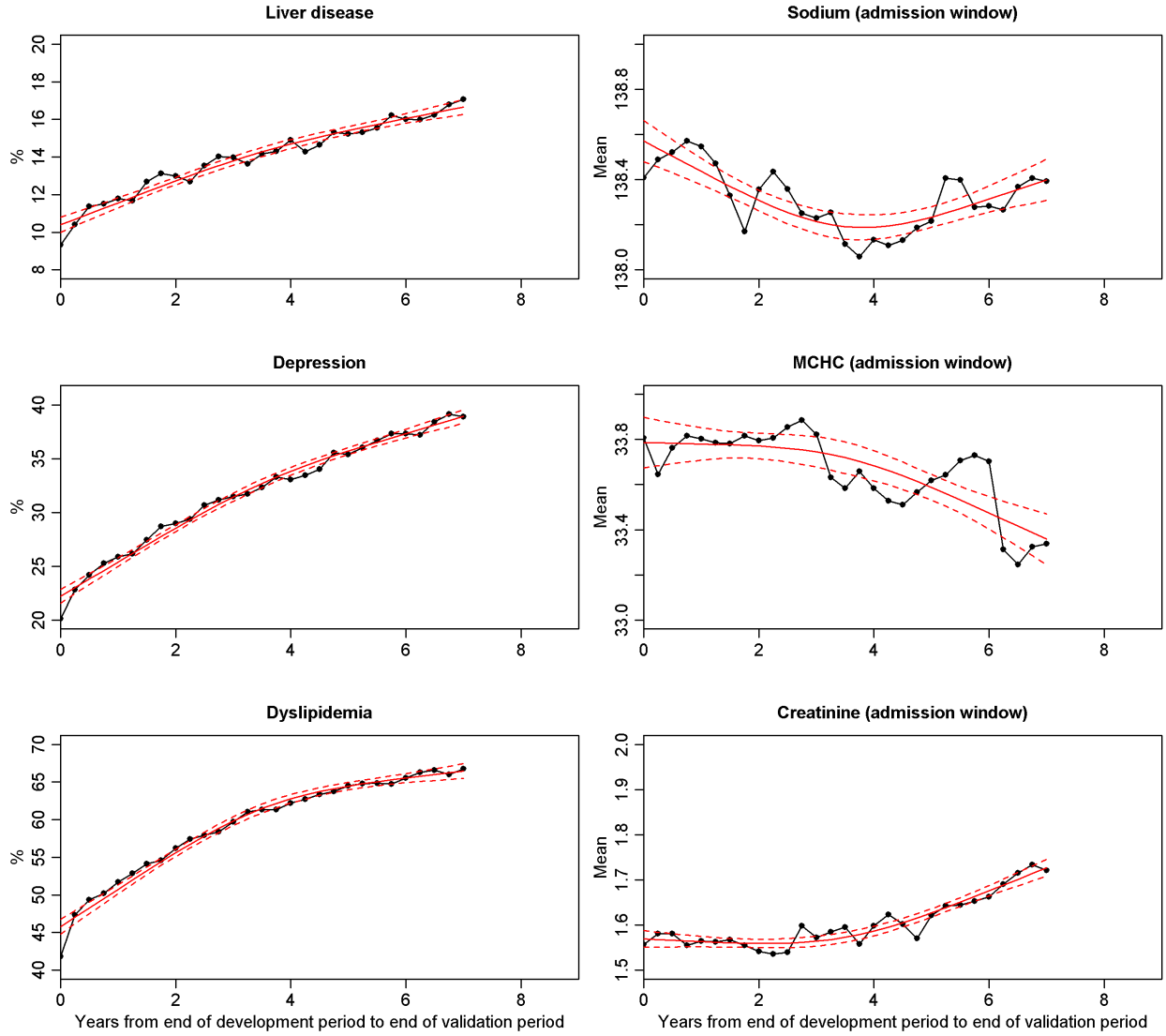
Figure 29. Discrimination of 30-day mortality logistic membership models over time



We observed changes in the distributions of 94.5% of predictors during the study period (adjusted $p<0.0003$; see Figure 30 for details of six exemplar predictors). The proportion of admissions to black patients, admissions that were planned, and admissions that were unplanned but not readmissions did not change over time. In addition, the mean of the most recent blood urea nitrogen level in the admission window was constant over. Changes in vital signs, laboratory values, and body mass index were generally small in magnitude. The forms of these changes were variable, with some having an inflection points at three to four years after model development. The largest changes were observed among health history variables. With the exception of HIV, which declined by less than 0.5%, the proportion of admissions involving patients with each health condition increased across the validation period. The rates of these increases were generally constant over time. Among health history variables, the largest changes occurred for dyslipidemia (41.8% to 66.8%), fluid and electrolyte disorders (19.0% to 38.9%), and depression (20.1% to 39.9%), and the smallest change for lymphoma (1.2% to 1.6%).

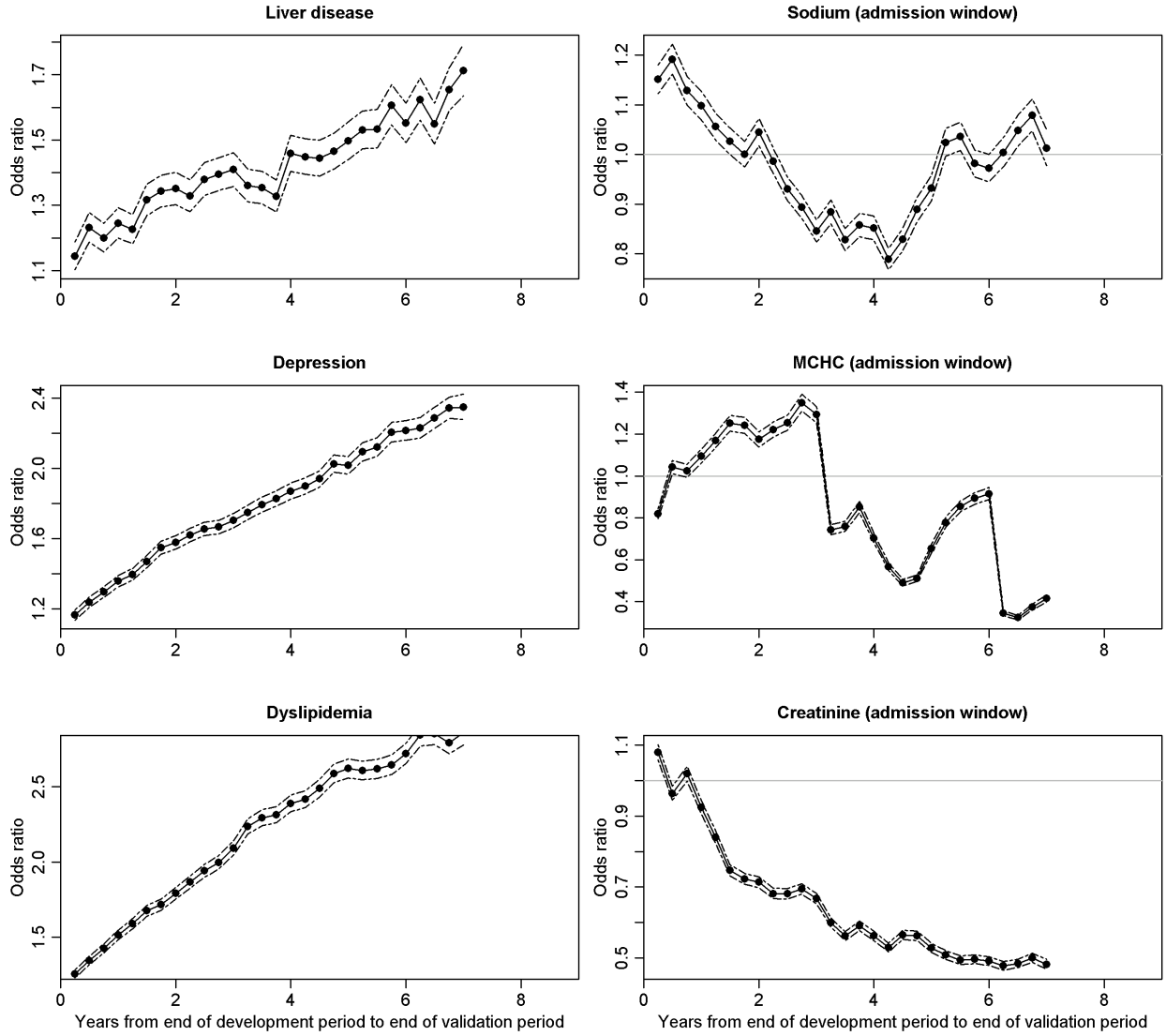
For most predictors, we observed temporal changes in the adjusted contributions of predictors to discriminating between development and validation admissions based on logistic membership models (see Figure 31 for details on the six exemplar predictors). Most laboratory

Figure 30. Distributions of select 30-day mortality predictors across over time. Continuous predictors summarized as means, categorical variables summarized as proportions. Red lines indicate fitted, smoothed values.



value predictors were variable in terms of both their strength and direction of association with membership in the validation versus development cohorts, indicating variable contribution to case mix differences. The health history variables generally either increased in strength of association as membership model predictors across the study period or during the first two to three years followed by stable odds ratios for the remainder of the study period. The largest changes in odds ratios, indicating the most substantial shift in contribution to case mix difference after adjustment for other variables, were observed for dyslipidemia, depression, drug abuse and fluid and electrolyte disorders. The increasing magnitude of these odds ratios resulted in these variables becoming the most predictive variables for distinguishing development and validation observations by the end of the study period. A history of blood loss

Figure 31. Odds ratios and 95% confidence intervals of select predictors from 30-day mortality logistic membership models. For continuous variables, odds ratios indicate effect of one IQR change in value. Grey lines indicate null odds ratio of 1.

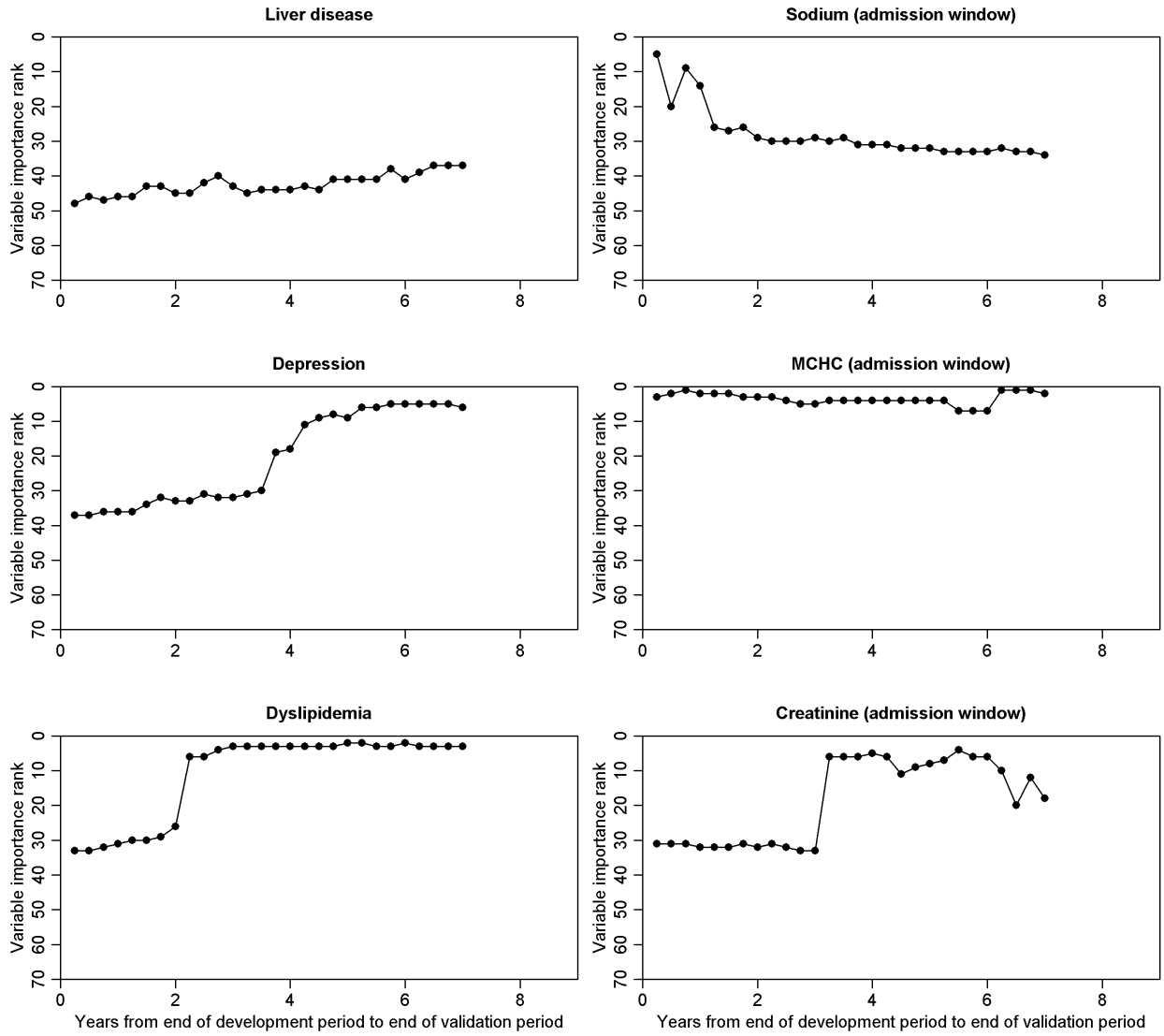


anemia, race, and unplanned readmission in the previous year were not significant predictors in the logistic membership models.

Temporal changes in the relative importance of predictors in random forest membership models were observed in 74.6% of predictors (adjusted $p<0.0004$; see Figure 32 for details on the six exemplar predictors). Among variables with significant temporal changes in variable importance rank, the mean change in rank was 14 (interquartile range: 7 – 17). Consistent with the logistic membership models, the largest contribution to case mix shifts based on the magnitude of change in random forest membership model variable importance rank were dyslipidemia, depression, drug abuse, and fluid and electrolyte disorders. The rank of drug abuse increased slowly over the first six years and then more quickly during the final study year,

leading to an overall increase in rank from 43 at the first validation period to 13 in the final validation cohort. For dyslipidemia, depress, and fluid and electrolyte disorders, variable importance ranks were initially fairly stable in the 30s and 40s (out of 67), increased quickly three to four year after development, and then remained stable in the top 10 most important predictors for the remainder of the validation period. Among predictors with smaller changes in variable importance rank over time, shifts in rank generally followed similar patterns with either consistent changes over time or an inflection in the trajectory of rank at roughly three or four years. Mean corpuscular hemoglobin concentration and platelet count were consistently among the highest ranking variables in the random forest membership models.

Figure 32. Variable importance ranks of select predictors from 30-day mortality random forest membership models. Higher ranks (smaller numbers) indicate greater importance to model performance.



Predictor-Outcome Association Shift

Changes in the strength of associations between predictors and 30-day mortality were measured by changes in the structure of models refit in each validation cohort. For the majority of predictor, we observed no temporal changes in association. For those predictors with the strongest and/or most changed predictor-outcome associations, we present results for four time points across the study period in Table 14. In addition, we present more detailed graphical results for the same six exemplar predictors included in the detailed case mix shift results above – history of liver disease, depression, and dyslipidemia, as well as the most recent sodium,

Figure 33. Odds ratios and 95% confidence intervals of select predictors from 30-day mortality logistic models refit over time. For continuous variables, odds ratios indicate effect of one IQR change in value. Grey lines indicate initial odds ratio and confidence interval.

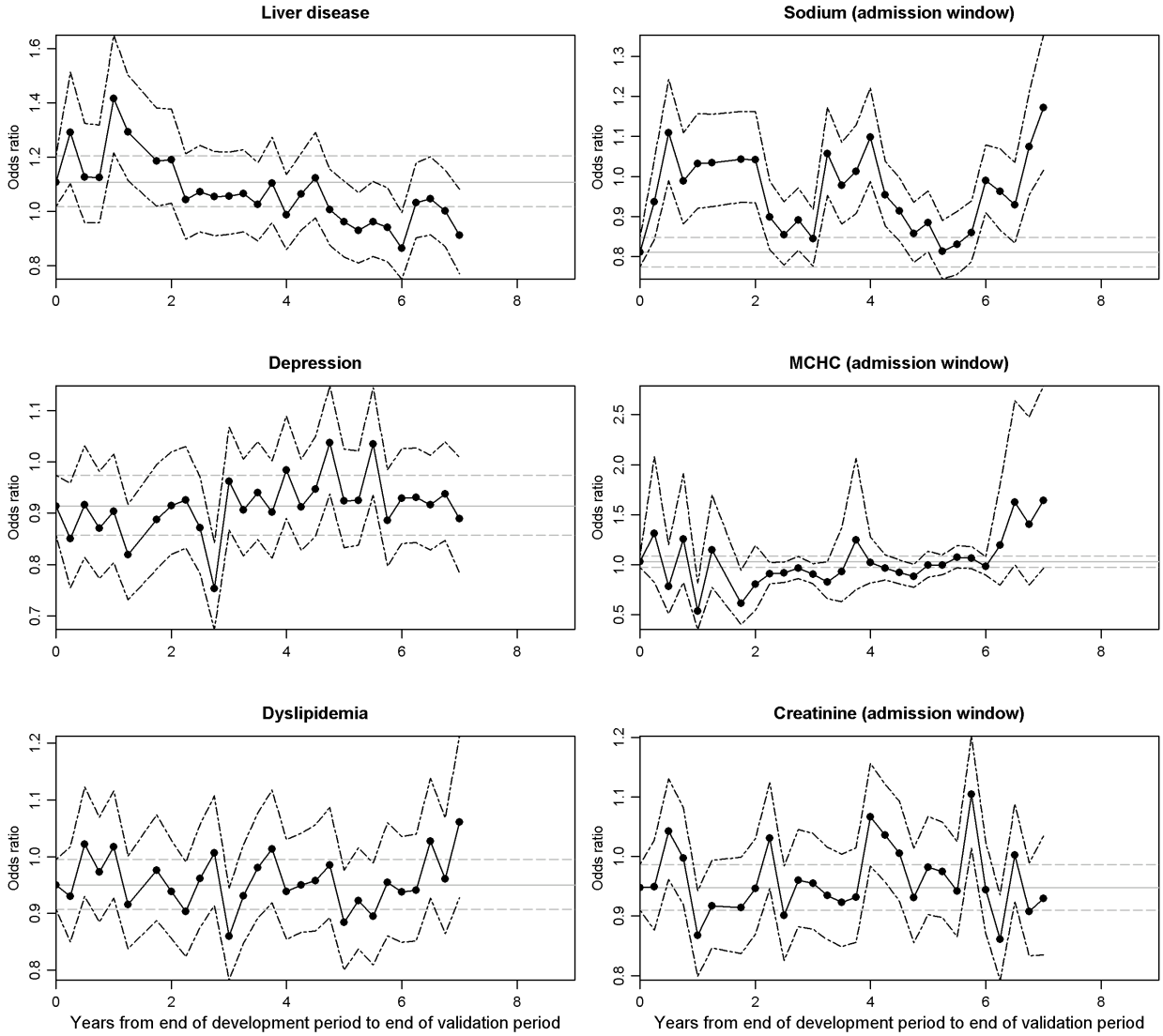


Table 14. Select predictor-outcome associations in 30-day mortality models refit over time. Odds ratios from logistic regression models (LR OR), variable importance ranks from random forest models (RF rank), and proportion of times predictors was selected in 200 L-1 penalized logistic regression (L1) modeling iterations based on 30-day mortality models fit at development and within select temporal validation cohorts.

	Development (2006)			2007 – Q4			2010 – Q4			2013 – Q4		
Predictor	**LR OR* (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												
Age (years)	2.47 [2.30, 2.65]	1	100	2.54 [2.19, 2.94]	1	100	2.57 [2.21, 2.99]	1	100	2.38 [1.93, 2.94]	2	100
Body mass index	0.80 [0.76, 0.84]	2	100	0.70 [0.62, 0.78]	4	100	0.75 [0.67, 0.83]	2	100	0.72 [0.62, 0.83]	5	100
Albumin	0.63 [0.60, 0.67]	3	100	0.67 [0.59, 0.75]	2	100	0.54 [0.48, 0.61]	3	100	0.53 [0.45, 0.63]	1	100
Pulse (minimum)	1.42 [1.34, 1.50]	4	100	1.46 [1.31, 1.64]	3	100	1.31 [1.18, 1.47]	4	100	1.54 [1.33, 1.78]	3	100
Pulse (maximum)	1.75 [1.65, 1.85]	5	100	1.45 [1.29, 1.63]	8	100	1.54 [1.38, 1.72]	7	100	1.51 [1.30, 1.75]	10	100
Alkaline phosphatase	1.15 [1.10, 1.20]	6	100	1.11 [1.01, 1.22]	5	100	1.18 [1.09, 1.29]	8	100	1.17 [1.04, 1.32]	9	100
White blood cell count	1.26 [1.20, 1.33]	7	100	1.30 [1.17, 1.45]	9	100	1.20 [1.08, 1.34]	9	100	1.30 [1.13, 1.49]	8	100
Systolic blood pressure (maximum)	0.80 [0.75, 0.85]	8	100	0.75 [0.66, 0.85]	6	100	0.67 [0.59, 0.75]	5	100	0.69 [0.59, 0.81]	4	100
Platelet count	0.94 [0.89, 0.99]	9	100	0.96 [0.86, 1.08]	10	100	0.88 [0.79, 0.98]	10	100	1.04 [0.90, 1.19]	6	99.5
Systolic blood pressure (minimum)	0.98 [0.92, 1.05]	10	100	0.86 [0.76, 0.99]	7	100	0.92 [0.81, 1.05]	6	100	0.94 [0.79, 1.11]	12	96
Variables with shifts in association												
Chloride	0.95 [0.90, 0.99]	28	99.5	0.77 [0.68, 0.87]	26	98.5	0.65 [0.58, 0.73]	23	100	0.61 [0.53, 0.71]	25	100
Depression	0.91 [0.86, 0.97]	51	99.5	0.90 [0.80, 1.02]	55	84.5	0.98 [0.89, 1.09]	48	60	0.89 [0.78, 1.01]	43	85
Dyslipidemia	0.95 [0.91, 0.99]	39	100	1.02 [0.93, 1.12]	40	51	0.94 [0.85, 1.03]	36	85.5	1.06 [0.93, 1.21]	50	39

Table 14. (continued) Select predictor-outcome associations in 30-day mortality models refit over time. Odds ratios from logistic regression models (LR OR), variable importance ranks from random forest models (RF rank), and proportion of times predictors was selected in 200 L-1 penalized logistic regression (L1) modeling iterations based on 30-day mortality models fit at development and within select temporal validation cohorts.

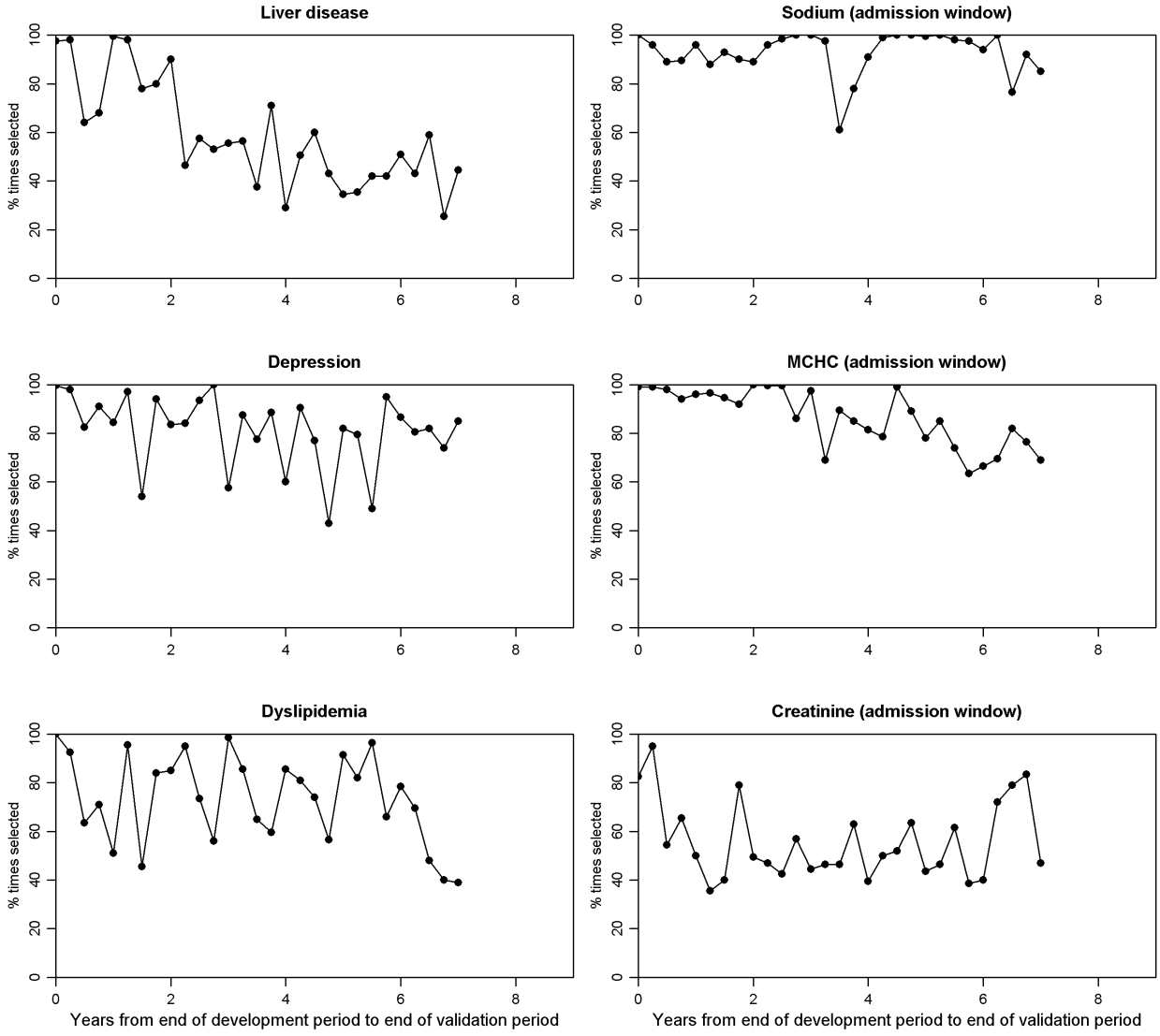
	Development (2006)			2007 – Q4			2010 – Q4			2013 – Q4		
Predictor	**LR OR* (95% CI)**	**RF Rank**	**L1 % selected**	**LR OR (95% CI)**	**RF Rank**	**L1 % selected**	**LR OR (95% CI)**	**RF Rank**	**L1 % selected**	**LR OR (95% CI)**	**RF Rank**	**L1 % selected**
Liver disease	1.11 [1.02, 1.21]	56	97.5	1.42 [1.22, 1.65]	42	99.5	0.99 [0.86, 1.14]	57	29	0.91 [0.77, 1.08]	57	44.5
Liver disease	1.11 [1.02, 1.21]	56	97.5	1.42 [1.22, 1.65]	42	99.5	0.99 [0.86, 1.14]	57	29	0.91 [0.77, 1.08]	57	44.5
Mean corpuscular hemoglobin concentration	1.03 [0.97, 1.09]	25	99	0.54 [0.35, 0.82]	24	96	1.02 [0.82, 1.28]	27	81.5	1.64 [0.96, 2.79]	24	69
Serum creatinine	0.95 [0.91, 0.99]	29	82.5	0.87 [0.80, 0.94]	29	50	1.07 [0.98, 1.16]	20	39.5	0.93 [0.84, 1.03]	17	47
Sodium	0.81 [0.77, 0.85]	23	100	1.03 [0.92, 1.16]	28	96	1.10 [0.99, 1.22]	28	91	1.17 [1.02, 1.35]	21	8525

* Odds ratios for continuous predictors are for an interquartile range increase in value

mean corpuscular hemoglobin concentration, and serum creatinine values recorded during the admission window.

We observed little evidence of temporal changes in the strength of predictor-outcome associations based on logistic regression models refit in each validation cohort (see Table 14 and Figure 33). For the majority of predictors, odds ratios remained within the confidence intervals of the odds ratio from the original model based on 2006 data. Some variables did exhibit a non-significant tendency toward strengthening or weakening associations. For example, a history of liver disease seemed to reveal declining odds ratio over time (see Figure 33), although the confidence intervals overlapped with those from the original model and

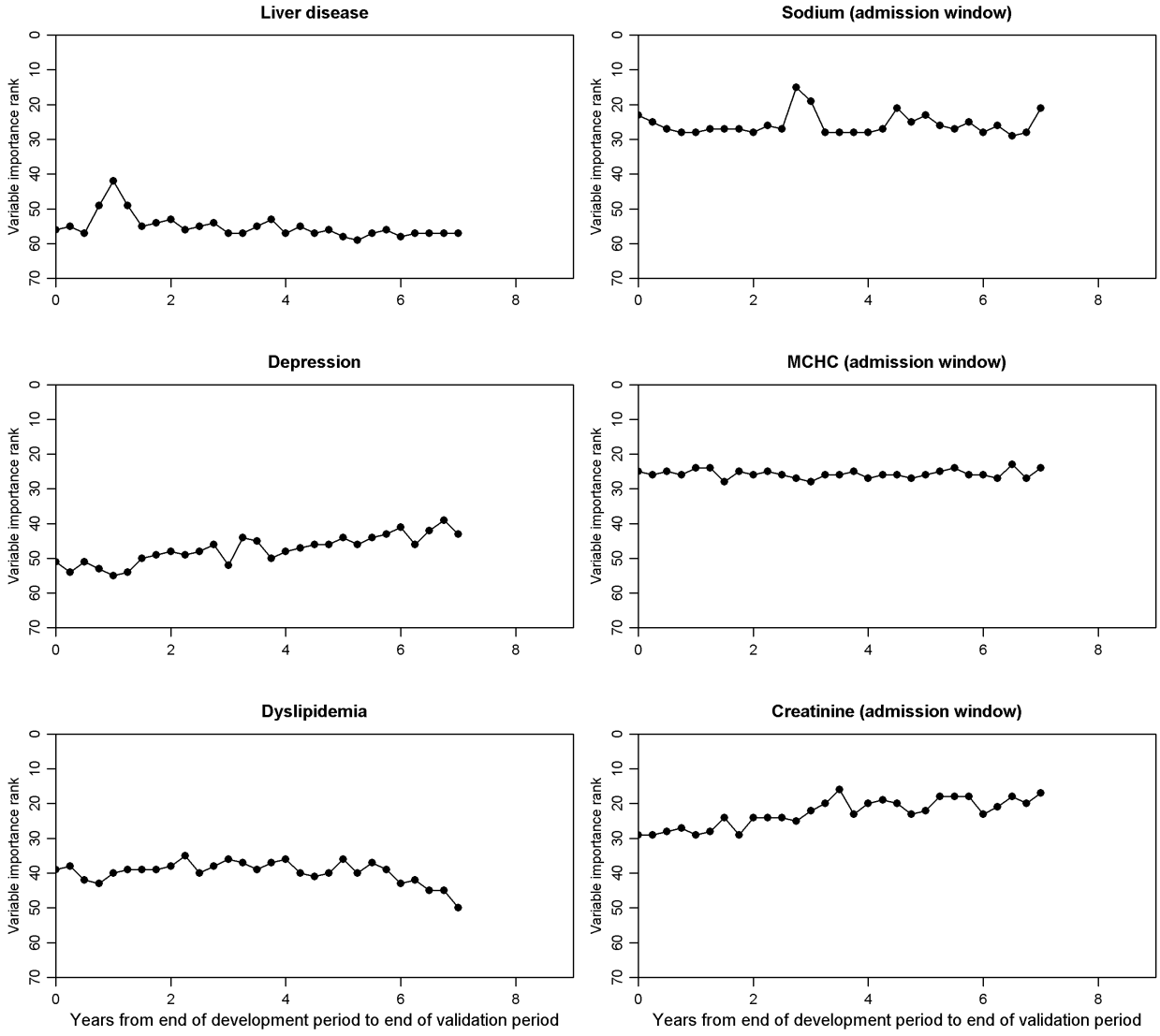
Figure 34. Variable selection of select predictors from 30-day mortality L-1 penalized logistic regression models refit over time. Proportion of bootstrap iterations (B=200) in which select predictors were retained.



captured the null value in nearly all validation cohorts. For some laboratory values, such as chloride and sodium levels during admission, odds ratios were less stable, moving in and out of significance in both directions of association.

In L-1 penalized logistic regression models refit over time, we observed significant temporal changes in selection patterns for two predictors (adjusted p<0.0004; see Table 14 and Figure 34). The frequency of selection for inclusion in L1 regression models began to decline starting after approximately two years of validation for both a history of liver disease and mean corpuscular hemoglobin concentration. The magnitude of decline in selection frequency was larger for liver disease, which was selected in 98.0% of bootstrap iterations at development and

Figure 35. Variable importance ranks of select predictors from 30-day mortality random forest models refit over time. Higher ranks (smaller numbers) indicate greater importance to model performance.



44.5% in the final validation period. Twenty-one of the 66 predictors were consistently selected for inclusion in the L1 regression models across the study period. These predictors characterized laboratory values, vital signs, demographics, and health history. Among the remaining of predictors, the proportion of times each was selected was variable, with no clear or statistically significant patterns emerging over time.

Refitting random forest models in each validation cohort revealed temporal changes in the variable importance ranking of three predictors (adjusted $p<0.0004$; see Table 14 and Figure 35). Ranks for a history of depression and serum creatinine during the admission window increased steadily across the validation period, increasing from 51 to 43 for depression and 29 to 17 for creatinine. The variable importance rank of dyslipidemia was steady through the first five years and then declined from 39 to 50 over the last two years of the validation period. Variable importance ranks were variable for most predictors, but did not experience short or long term tendencies toward increasing or decreasing rank. Age at admission was consistently identified as the most important predictor, with the exception of the final validation cohort in which it was ranked second. Body mass index and minimum pulse in the admission window were also consistently ranked highly across the study period.

Conclusions

In repeated validations over seven years, we observed varying patterns of performance among regression and machine learning models for 30-day all-cause mortality after hospital admission. Among all models, discrimination was stable over time. With the exception of the neural network model, calibration drifted across the entire validation period as all other models increasingly overpredicted risk. Seasonal changes in the mortality rate were correlated with cyclical fluctuations in the calibration of all models, including the neural network model despite its maintaining stable calibration overall. Case mix shift dominated temporal changes in the patient population. Taken together with the observed calibration drift, the data shift assessments highlight robustness of the neural network model, moderate susceptibility of the random forest and L-2 penalized logistic regression models, and high susceptibility of the other regression models to case mix shift. In the next chapter, the implications of these findings are interpreted in conjunction with the findings of our acute kidney injury analysis, in which different patterns of performance drift and data shift were observed (see Chapter 4), to provide a more complete assessment of the impact of modeling methods on model performance over time.

CHAPTER 6

IMPACT OF MODELING METHODS ON PERFORMANCE DRIFT AND IMPLICATIONS FOR MODEL UPDATING PROTOCOLS

Our literature review highlighted the need for additional research into whether and how modeling methods exacerbate or alleviate calibration drift under varying data shift scenarios. In this rigorous comparison of performance drift in regression and machine learning models, we demonstrated that discrimination remained reasonably stable over time and all modeling methods were susceptible to deteriorating calibration. The degree of susceptibility of each modeling method to calibration drift varied based on the types of data shifts occurring in the patient population. All methods were susceptible to changes in the underlying event rate, while shifting case mix and predictor-outcome associations had a greater impact on regression models than the machine learning approaches. These findings highlight the need to tailor model updating protocols to modeling methods in terms of both the approach to and timing of recalibration or revision. Here we integrate findings from both our hospital-acquired acute kidney injury and 30-day all-cause mortality analyses, discuss implications of these findings, note limitations of our approach, and provide recommendations for updating strategies.

Performance Drift across Modeling Methods

We observed stable discrimination over time for modeling methods. The exception to this finding was the naïve Bayes model for AKI which experienced a statistically significant decline in discrimination; however, the magnitude of this change was minimal, resulting in the AUC weakening from 0.69 to 0.68 over 9 years, a difference unlikely to have practical implications. The stability of discrimination in our models was consistent with findings of our literature review, which highlighted various logistic models having stable discrimination even up to 20 years after model development.[23, 24, 78, 82, 90, 92, 99, 100]

Calibration deteriorated over time for all modeling methods. The calibration metrics being considered directly impacted whether we observed differences in calibration drift across methods. For our mortality models, we detected differences in calibration drift by modeling method at all levels of calibration. For our AKI models, calibration drift was similar across methods when characterized with mean and weak calibration measures, yet machine learning models exhibit superior stability in calibration compared to regression models when considering the most stringent calibration measures.

Mean calibration, as measured with O:E ratios, deteriorated over time for all modeling methods. In our AKI analysis, O:E ratios declined over the first four validation years, indicating increasing overprediction, with several periods of rapid drift, particularly in the first and fourth years after development. This pattern was consistent across AKI models. In our mortality analyses, O:E ratios declined across the seven years of the validation period for all models except the neural network model, which exhibited a cyclical pattern in O:E ratios and initial drift

into overprediction but did not demonstrate a statistically significant declining trend over time. Our findings of increasing overprediction on average for both outcomes were consistent with previous studies of calibration drift.[23, 24, 76, 82, 90, 95, 96] While prior work focused on logistic regression models, Minne et al[24, 92] directly compared calibration drift based on O:E ratios of logistic regression and tree-based versions of the rSAPS-II model. O:E ratios indicated increasing levels of overprediction within four years of development for the logistic model, while remaining relatively stable for the tree-based model.[24, 92] In contrast, we did not observe more stability of O:E ratios for our random forest models compared to logistic models. Minne et al[24, 92] did not report more details measures of calibration or evaluations of data shifts, preventing us from fully understanding the factors driving the differences in our findings.

At the level of weak calibration, we observed increasing overprediction and generally stable levels of overfitting over time, although findings varied by modeling method in some cases. Among all AKI models, Cox intercepts declined over time, indicating increasing overprediction. The rate of this change, however, was greater in the first half of the validation period than the second. For regression models for mortality, Cox intercepts also showed increasing overprediction across the validation period. The decline in Cox intercepts was smaller for the L-2 penalized logistic regression model than other regression models. For the mortality models built with random forest and neural network approaches, Cox intercepts were stable over time. The level of overfitting, as measured by the Cox slope, was stable over time for all mortality models, as well as the AKI random forest and neural network models. Statistically significant, but small in magnitude, declines in Cox slopes were recorded for the AKI regression models. Our findings for the logistic models for both AKI and mortality were consistent with an assessment of the logistic APACHE-III model over a 10-year period by Paul et al[95] which reported declining Cox intercepts and stable Cox slopes.

For the most stringent calibration measures considered in these analyses, the machine learning methods exhibited more stability in performance than the regressions methods. In our AKI models, temporal deterioration in ECIs over time was substantively greater for regression compared to machine learning models. These diverging patterns of calibration drift were particularly apparent during the second half of the validation period. The neural network and random forest AKI models maintained calibration over probability ranges covering more admissions than the regression models and maintained a more consistently low magnitude of overprediction compared to an increasing magnitude of overprediction among the regressions. In our mortality analysis, the L-2 penalized logistic regression and random forest models demonstrated increasing ECIs (i.e., deteriorating moderate calibration) over time, although to a lesser degree than the other regression models. While the neural network mortality model had stable ECIs over time and was the best calibrated mortality model throughout the validation period, it also exhibited seasonal performance patterns in which calibration was markedly improved in the first and fourth quarters, corresponding to Autumn and Winter months. No prior studies of calibration drift were available for comparison with our moderate calibration findings.

Linking Data Shifts and Calibration Drift

Performance drift results from data shifts in patient populations, including changes in the outcome rate, patient case mix, clinical practice, and documentation practices.[5-7, 21, 23, 73] Our literature review underscored the limited availability of studies directly linking performance drift with temporal changes in patient populations. McCormick et al[89] and Hickey et al,[86] however, noted complex, multifaceted forms of data shifts, such as those we observed in both our AKI and mortality cohorts. Event rate shift across the study period, predictor-outcome association shift during the second half of the validation period, and complex case mix shift throughout the study period were documented in our 10-year AKI cohort. Over the eight years of admissions in the mortality cohort, we observed seasonal variation in the event rate, case mix shift across the study period, and minimal evidence of predictor-outcome association shift. These shifts resulted in performance drift at varying rates across the validation period and disparate patterns of calibration drift across models. Event rate and predictor-outcome association shifts straightforwardly link to patterns in performance drift. However, linkages between strong, complex case mix shift, which also influenced performance drift, are more difficult to directly link with performance. Despite such challenges, we integrated results across components of the analyses and across our two clinical domains to identify differences in the susceptibility of modeling methods to each form of data shift. These findings are summarized in Table 15.

Table 15. Relative susceptibility of modeling methods to calibration drift under each form of data shifts in patient populations. Note, the calibration of the naïve Bayes model was insufficient in all cases, and so we do not include the model here.

Modeling method	Event rate shift	Association shift	Case mix shift
Logistic regression	High	High	High
L-1 penalized logistic regression	High	High	High
L-2 penalized logistic regression	High	High	Moderate
L-1/L-2 penalized logistic regression	High	High	High
Random forest	High	Low	Moderate
Neural network	High	Low	Low

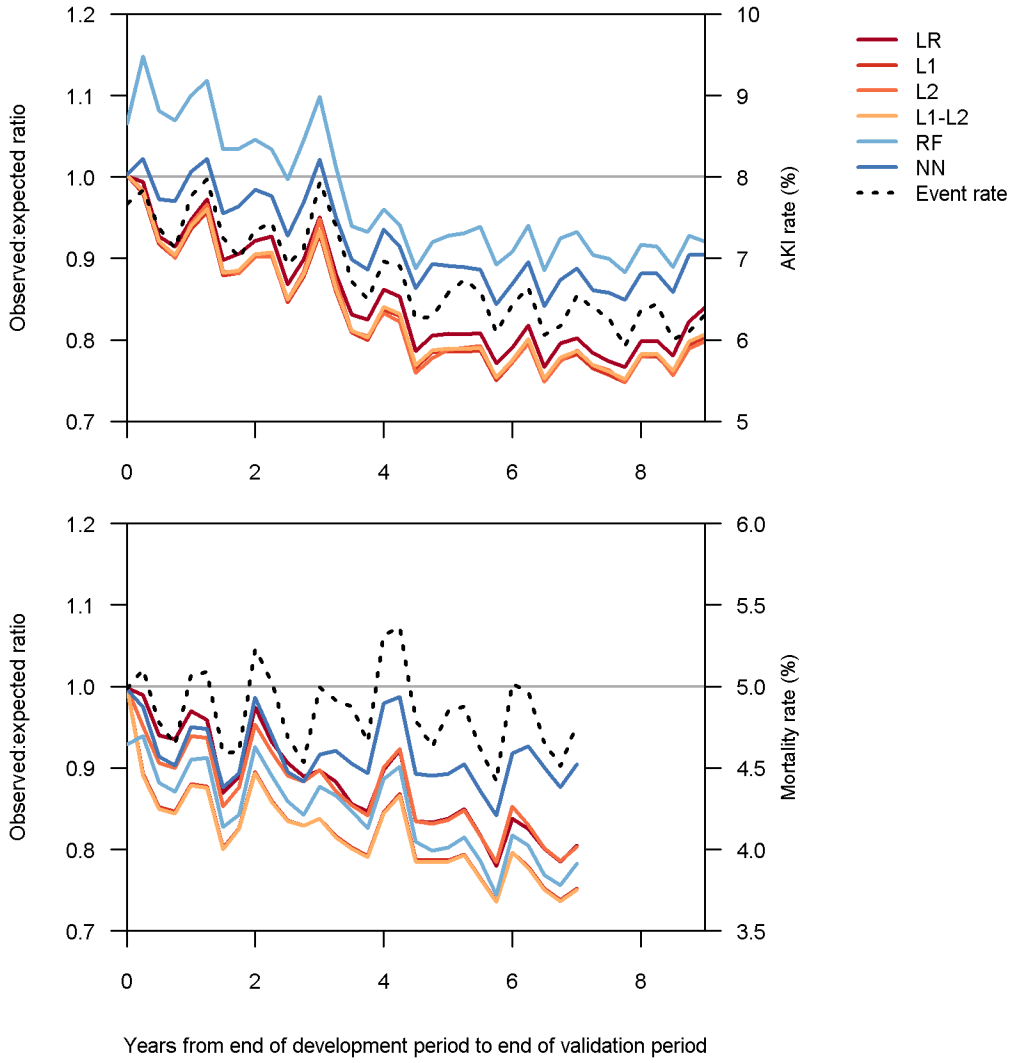
Note: Susceptibility of calibration for each model is relative to other models considered and is not intended as an absolute measure of model susceptibility to calibration drift.

Event Rate Shift and Calibration

Different patterns of event rate shift were observed in the AKI and mortality cohorts. The event rate in our AKI cohort declined from 7.7% in the development cohort to 6.3% across the study period. While we observed a small overall decline in the mortality rate (5.0% at development and 4.8% in the final validation cohort), seasonal variability within each study year was three times larger than this overall change.

Mean, weak, and moderate calibration metrics were susceptible to event rate shift for all modeling methods. Figure 36 illustrates the correlations between O:E ratios and outcome rates. In our AKI models, O:E ratios were strongly positively correlated with the event rate (Spearman rho: 0.92-0.95; adjusted p<0.001). In our mortality models, O:E ratios were strongly positively correlated with the event rate for the neural network model (Spearman rho: 0.92; adjusted p<0.001) and were moderately positively correlated for all other models (Spearman rho: 0.66-0.72; adjusted p<0.001). The mortality panel in Figure 36 highlights that although the O:E ratios of the neural network model did not decline over time (adjusted p>0.0003), the seasonal pattern of higher mortality rates in the first and fourth quarters of each year (i.e., Autumn and Winter months) was strongly correlated with seasonal fluctuations in O:E ratios. This seasonal fluctuation in O:E ratios can be observed for the other mortality models. However, there is an additional temporal decline in these models' O:E ratios beyond that attributable to event rate

Figure 36. Observed to expected outcome ratios and event rates over time for AKI (top) and 30-day mortality (bottom) cohorts



shift, which is also reflected in the lower correlation coefficients for these models compared to the neural network model. These linkages between O:E ratio drift and event rate shift correspond with previous studies.[23, 76, 93] At more refined levels of calibration, Cox intercepts were strongly positively correlated with event rate for the AKI regression models (Spearman rho: 0.93-0.94; adjusted p<0.001), and were moderately correlated for the AKI random forest and neural network models (Spearman rho: 0.64 and 0.79, respectively; adjusted p<0.001). Among mortality models, Cox intercepts were moderately positively correlated with the event rate for the penalized regression models only (Spearman rho: 0.59-0.62; adjusted p<0.001). ECI was strongly negatively correlated with event rate for all AKI models (Spearman rho: -0.83 to -0.95; adjusted p<0.001), but was not correlated for any of the mortality models.

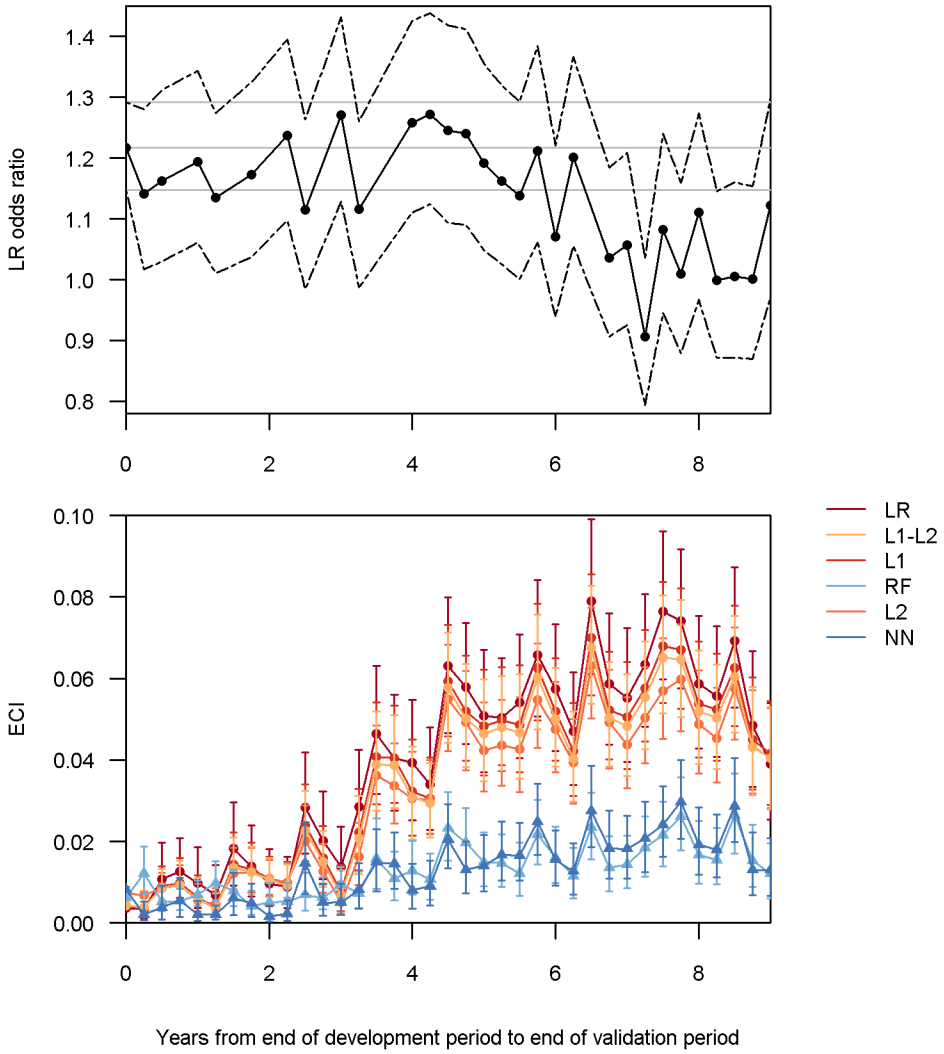
Declines in event rates, whether over extended periods of time or during specific seasons, were associated with increasing overprediction. The link between reduced event rates and overprediction is most clearly illustrated by the results for the neural network mortality model. The mortality rate negligibly declined over time, and the neural network mortality model did not exhibit a tendency toward either increasing overprediction or increasing underprediction. Even small seasonal variations in the mortality rate, however, impacted both the model's O:E ratios and proportional volume regions of calibration. Among patients predicted to be low risk, the neural network mortality model marginally overpredicted during the second and third quarters of most years when the mortality rate tended to be lower than average, but provided more calibrated predictions during the first and fourth quarters when the mortality rate tended to be higher than average. On the other hand, in the AKI cohort, in which the event rate decreased across the validation period, the magnitude of overprediction increased over time for all models. This pattern was apparent at various levels of calibration from declining O:E ratios less than 1.0 to temporally increasing ECIs within regions of overprediction. The random forest and neural network models for AKI exhibited smaller absolute changes in each calibration metric compared to the regression models, and in the regions of calibration analysis did not reveal large temporal increases in the magnitude of overprediction as was observed for the regression models. These findings may indicate some difference in susceptibility of modeling methods to event rate shift. However, similar levels of correlation between AKI rates and O:E ratios, and the isolation of the differences in regional calibration patterns to the second half of the validation period rather than accruing across the study period may indicate difference across methods are due to something other than event rate shift.

Predictor-Outcome Association Shift and Calibration

Predictor-outcome association shifts were assessed by refitting the logistic regression, L-1 penalized logistic regression, and random forest models within each consecutive 3-month validation period. The mortality cohort did not appear to be experiencing significant shifts in predictor-outcome associations, and we thus focus on the AKI results for an understanding of how association changes may influence calibration drift across modeling methods. In the AKI cohort, shifts in predictor-outcome associations coincided with diverging patterns of calibration drift between the machine learning and regression models (see Figure 37 for illustrative

example). Although we observed limited evidence of changes in the strength of associations between predictors and AKI, the shifts we did document tended to occur over the second half of the validation period. This temporally corresponded with the years during which the regression models experienced more substantial calibration drift than the random forest and neural network models as measured by ECIs and regions of calibration. This may indicate greater susceptibility of regression methods to predictor-outcome association shift compared to random forest and neural network models, even in the presence of few or small association changes (see Table 15). Additionally, this link between predictor-outcome association shift and model calibration appeared to impact moderate measures of calibration to a greater extent than mean and weak calibration measures, which experienced consistent patterns of drift across AKI models.

Figure 37. Example of temporal concurrence of predictor-outcome association shift and calibration drift in AKI models. Odds ratios for an interquartile range change in age at admission from refit logistic models for AKI (top) and estimated calibration index of AKI models over time (bottom).



Case Mix Shift and Calibration

Membership models clearly indicated the presence of case mix shift in both our AKI and mortality cohorts. We noted statistically significant changes over time in the distribution of approximately 95% of predictors for both AKI and mortality. The complexity of concurrent changes across the majority of predictors prevents the linkage of changes in individual predictors with performance drift, and the concurrence of multiple forms of population data shift, particularly in the AKI analysis, complicates our interpretation of the case mix results.

In our AKI cohort, most of changes in predictor distributions were relatively small in magnitude and may not have been substantial enough to impact performance. Changes in the distribution of the suite of predictors, however, lead to an increase in mean patient severity. These changes did not lead to a change the variability of AKI risk in the patient population. This may indicate that shifts among correlated variables increased patient severity while keeping the case mix balanced in terms of risk distribution. On the other hand, case mix shifts leading to changes in patient severity, as observed in our study population, may affect mean and weak calibration,[73] both of which experienced drift over the validation period. We might expect, however, that increasing severity of the patient population would result in underprediction,[73] while we observed increasing overprediction. These results parallel those previously reported in three studies documenting both increasing patient severity and declining O:E ratios.[23, 88, 90] This underscores a high degree of susceptibility of all modeling methods to event rate shift as a driver of calibration drift even in the presence of case mix shifts. It may also be the case that our predictor set is not capturing certain risk factors that may also be shifting over time ways that actually reduce mean patient severity and negatively impacting our ability to characterize case mix changes in the patient population.

Focusing on our mortality models may provide a clearer connection between case mix shift, performance drift, and modeling methods. In our mortality cohort, we did not detect substantial predictor-outcome association shifts and observed primarily seasonal changes in the event rate. As previously noted, this seasonal event rate shift was highly correlated with a cyclical pattern in calibration. However, with the exception of the neural network model, our mortality models experienced calibration drift that could not be explained by the seasonal variation in the mortality rate alone and is likely, therefore, associated with the documented changes in the patient case mix. The stability of calibration at all levels for the neural network model suggests this method is robust to case mix changes, at least to the extent seen in our mortality cohort. At the weak and moderate levels of calibration, the L-2 penalized logistic regression and random forest models experienced less deterioration in calibration over time than did the logistic, L-1 penalized logistic, and L-1/L-2 penalized logistic regressions. The basic logistic and penalized logistic regression methods that include variable selection, therefore, appear to be the most susceptible methods to calibration drift in the presence of case mix shift (see Table 15). The L-2 penalized logistic regression and random forest models appear to be moderately susceptible, falling between the other regression approaches and the neural network model.

Additionally, we note that case mix shifts leading to changes in the variability of risk in the patient population have been noted to be of particular concern for discrimination drift as patients become more difficult to distinguish or more easily separable as they become more

homogenous or more heterogeneous.[73] Although we observed increased heterogeneity of risk among patients in our mortality cohort, we did not observe changes in discrimination of our mortality models over time. This may indicate that the degree of change in the variability of the risk distribution was not sufficient to trigger discrimination drift. Studies of additional cohorts with greater changes in case mix variability would be necessary to understand whether modeling methods can withstand similar degrees of change prior to experiencing discrimination drift.

Informatics Implications

Our findings have important implications for the design and implementation of e-HPA systems and prediction model maintenance protocols.[1] We recommend e-HPA systems incorporate active surveillance components to track model performance and characterize changes in patient populations over time. We further recommend the incorporation of both modeling goals and methodologies into such systems in order to tailor the performance metrics of concern, frequency of updating, and updating approach.

The metrics most relevant to assessing performance over time and informing decisions regarding the need to update a clinical prediction model will vary by the context in which predictions are used. For use cases focused providing risk categories or for which only granular estimates are needed, the ability of a model to separate observations with and without the outcome is sufficient, while accuracy of individual predictions is not imperative. The stable discrimination observed for our AKI and mortality models, as well as in previous performance drift studies,[23, 24, 82, 88, 95, 99, 100] may alleviate model updating concerns for such use cases. In these settings, frequently assessments of discrimination may not be necessary; however, model updating protocols must still be in place to address instances of change in local clinical or organizational practices which may require model revision.[153] On the other hand, when individual predictions are of interest, discriminating but inaccurate predicted probabilities may lead to over-confidence, inappropriately alter treatment choices, or misappropriate resources.[1, 23, 67, 68] Precision medicine tools supporting individually tailored decisions must thus be concerned with calibration and attentive to drifting performance.

In our study, not all calibration metrics captured important features of performance drift. Moderate calibration, evaluated with flexible calibration curves, is particularly relevant to the implementation of individual-level prediction systems. With moderate calibration, models have a net benefit greater than or equal to treat-all or treat-none strategies, thus ensuring predictions are nonharmful to decision-making.[71] Among our AKI models, traditional weaker calibration assessments (i.e., O:E ratios and Cox recalibration model intercepts and slopes) failed to reveal differences in calibration drift between modeling methods that were readily apparent with measures of moderate calibration. Flexible calibration curves revealed substantially more deterioration in performance over time among regression models for AKI than corresponding random forest and neural network models. Based on these findings, we recommend surveillance systems focusing on calibration should implement moderate calibration measures, including the ECI and regional calibration assessments. Although the familiarity of O:E ratios and Cox intercepts and slopes is appealing and might also be tracked, reporting more stringent

metrics will enhance understanding of potentially clinically-relevant calibration issues and underscore methods-based performance vulnerabilities that may inform updating strategies.

The frequency of repeated performance evaluations and model updating should be balanced to maintain acceptable levels of performance and ensure sufficient information content in validation and updating cohorts. In our two study populations, models began drifting toward overprediction within one year after development. We also noted periods of rapid change in calibration, particularly during the first and fourth AKI validation years. The neural network model for mortality, however, maintained stable calibration over time and did reveal a clear need for updating during the study period. Our results thus raise concern over the common approach of simply updating at scheduled annual or biannual intervals,[46, 154, 155] as this may allow periods of reduced accuracy in the interim or result in the updating of well-performing models. While avoiding unnecessary model updating may conserve analytical and computational resources, particularly when refitting complex models, measuring calibration even at stringent levels is not especially computationally intensive. We thus recommend active model surveillance systems conduct frequent performance assessments at the shortest time interval that provides enough observations to support performance measurement. Prior research suggests at least 10 outcome events per predictor is necessary to ensure sufficient information content for model validations.[46] We elected to assess performance every 3 months, which provided more than enough data in each validation cohort based this rule of thumb (recommended minimum sample sizes vs mean sample size in validation cohorts: ~17,500 vs ~46,000 for AKI and ~13,500 vs ~60,000 for mortality). In settings with large patient cohorts or common outcomes, monthly or bi-monthly performance measures may be possible and preferable. Dynamic assessment as new observations become available may also be a feasible consideration for some systems. The timing of model recalibration or revision may then be data-driven based on observed performance degradation and would inherently be tailored to the susceptibility of various modeling methods to calibration drift.

We further recommend active model surveillance systems monitor data shifts to provide early warning of the need for model updating and insight into the updating approach required to correct deteriorations in performance. There is a range of approaches to model updating, from simple recalibration to complex recalibration to full model revision (i.e., refitting) and even model extension with the incorporation of new predictors.[6, 7, 45, 46, 156] Recalibration techniques retain information in existing models and improve generalizability, making these approaches preferable to revision when recalibration is sufficient to improve performance to acceptable levels.[7, 45, 153, 156] Recalibration would be indicated when event rate shift dominates data shift and both stable discrimination and deteriorating calibration are recorded. However, in the presence of predictor-outcome association shifts, significant case mix shift, or changes in discrimination, full model revision should be recommended.[7, 45, 153, 156] Surveillance systems tracking both model performance and population data shifts could leverage information about the susceptibility of modeling methods to identify key data shift components driving observed changes in performance in order to recommend an updating approach. As our study suggests, models based on different methods have variable updating needs, so the choice of updating approach should be tailored to the distinct susceptibilities modeling methods to performance drift. For example, while the detection of case mix shift may motivate model revision for most models, our mortality results suggest neural networks may not require updating as a result of isolated case

mix shift. Similarly, in our AKI analyses, regression models required revision in the presence of limited predictor-outcome association shifts, while under the same circumstances the random forest and neural network models maintained performance without updating. Of course, we also emphasize the importance of local knowledge, as changes in local clinical practice, organizational guidelines, or data definitions should always trigger model updates.[153]

Clinical Implications

Users of clinical prediction models, including providers, patients, and health administrators, should be aware of and attentive to calibration drift. Poor calibration can lead to poor decision-making at the level of both individual patients and health systems. Patients may be dissuaded from pursing potentially effective treatments when presented with erroneously elevated estimates of complication risks or may elect to undergo difficult treatments when presented with inflated estimates of negative disease prognosis.[23, 67] Drifting calibration may also mislead benchmarking quality assessments that utilize prediction models to risk-adjust quality metrics for differences in patient case mix and severity between hospitals or care units. Minne et al[24] found calibration drift of the logistic rSAPS-II model resulted in overly optimistic quality assessments. Fifteen percent of hospitals were identified as underperforming by the original rSAPS-II model affected by calibration drift, while 35% of the hospitals were identified as underperforming when the model was recalibrated to correct for overprediction.[24] In another study comparing model recalibration and revision techniques, however, model updating did not influences quality assessments based on EuroSCORE predictions.[46] Our study highlights that all models, regardless of the underlying modeling method, are susceptible to calibration drift in some data shift circumstances and calibration drift can occur quickly. Analysts managing models for patient-level decision-making and health system assessment should be proactive in developing policies and procedures for model updating.

Limitations and Paths to Address Them

Limited, Complex Data Shift Scenarios

Although we utilized a large, national cohort, we explored model performance under just two data shift scenarios in our acute kidney injury and 30-day mortality study populations. All three forms of data shift were observed in our AKI population. Our mortality population experienced case mix shift and seasonal event rate shift, but not predictor-outcome association shift. These differing data shift combinations allowed for an initial understanding of the relative susceptibility of modeling methods to each form of data shift. With only two illustrative examples and co-occurring forms of shift, we are limited in our ability to draw conclusions about the extent of data shift each modeling method can tolerate. Extending our analyses with additional illustrative clinical outcomes may provide more evidence regarding in what circumstances certain modeling methods are more or less susceptible to performance drift. However, the results of such an approach would remain complicated by complex, co-occurring forms of data

shift. Simulation studies, on the other hand, would permit systematic control of the form and extent of data shifts. Such studies based on simple simulated datasets with defined variable relationships could be complemented by parallel analyses using synthetic datasets built from existing clinical cohorts to better reflect the depth of complexity in real-world variable relationships. With such simulations, we can consider model performance patterns under each form of shift in isolation and in predefined combinations. We may also explore other components of data shift that cannot be characterized with illustrative use cases, such as case mix or predictor-outcome association shifts in risk factors omitted from models. A simulation study would thus allow a more detailed characterization of the susceptibility of each modeling method. Additionally, simulations would provide a platform for testing hypotheses regarding the technical aspects of each model that convey susceptibility or robustness of calibration to data shifts in the patient population. By integrating results from illustrative use cases and simulation studies considering both fully simulated and synthetic datasets, we could construct a more fully formed understanding of the links between modeling methods, data shifts, and performance drift.

Characterization of Case Mix Shift

We observed complex case mix shift involving changes in the distribution of multiple predictors in both cohorts. Despite implementing a variety of methods to characterize case mix changes, directly linking shifting patterns among a multitude of individual predictors with calibration metrics remains challenging. Identifying those predictors driving the case mix changes that impact performance is further complicated by correlations between predictors and the potential for changes in omitted variables. While the existing methods for characterizing case mix shift focus on individual predictors, both in isolation and adjusted in multivariable membership models, the features of case mix shift most relevant to performance drift may not require such a detailed focus. Instead, we might consider clustering groups of similar patients and exploring how the density of our population moves between clusters over time. If model predictions are more accurate for patients in some clusters and less accurate in for those in others, then performance drift may be directly linkable to increased or decreased data density in high and low performing clusters. Identifying those clusters driving performance drift may allow us to better understand why some modeling methods are more robust to case mix shift than others by exploring the characteristics of key clusters and how models capture associations within these subgroups.

Statistical versus Practical Miscalibration

Our study explored model calibration from a statistical perspective rather than that of clinical utility. Calibration drift is of most concern when it is sufficient to influence the clinical impact of a model. A recent study recommended assessment of flexible calibration curves to ensure nonharmful predictions,[71] making our findings regarding the divergent patterns of ECI and regions of calibration across models an important consideration in model implementation.

However, statistically significant calibration drift may still not translate directly to clinically important changes in model performance. This is particularly true for large study populations, such as our AKI and mortality cohorts, in which confidence intervals can be quite narrow and cause even small deviations from perfect calibration to be identified as significant miscalibration. We observed an example of this issue with our mortality flexible calibration curves. In our mortality regression models, the flexible calibration curves identified a large proportion of admissions being contained in ranges where predicted probabilities were statistically significantly too high; however, the degree of overprediction was minimal and unlikely to be important from a practical standpoint. For example, in the logistic regression model for mortality, a low risk region of overprediction in each validation period captured more than 40% of admissions and these regions had a mean ECI of 0.005 (range: 0.002 – 0.009). Compared to the ideal ECI of 0 in the case of perfect calibration, an ECI of 0.005 is near perfect and essentially calibrated in a practical sense. Understanding whether, when, and how calibration drift affects the clinical utility of predictions for decision-making is one of the most important considerations for informing recalibration guidelines. Extending this study with the incorporation of clinical utility measures or the adaptation of calibration metrics to account for clinically acceptable margins of error would improve our ability to comment on how modeling methods influence performance drift in clinically meaningful ways.

Additional Modeling Methods

Finally, there are a multitude of machine learning methods applicable to clinical prediction problems, and we have only included a limited number of commonly used methods in this study. Other modeling methods – such as support vector machines, k-nearest neighbors, averaged one-dependence estimators, and Bayesian networks – may reveal different patterns of susceptibility to data shifts and calibration drift. This study thus serves as an initial exploration documenting the influence of modeling methods on performance drift in the presence of various data shifts. As computational resources permit, we could extend our findings with additional modeling approaches.

Conclusions

Growing opportunities to leverage predictive analytics and 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 and the influence of modeling methods on performance drift. Our finding of stable discrimination 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 personalized predicted probabilities. Our findings can inform recommendations regarding the timing of and approach to model updating. We recommend frequent validation of all models, with careful consideration of the timing of repeated assessment, sample sizes supporting each

assessment, and whether discrimination or stringent calibration metrics should be measured. In addition, routinely monitoring data shifts may 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, stable discrimination, and event rate shift; however, in the presence of predictor-outcome association shifts, significant case mix shift, or changes in discrimination, full model revision would be indicated. Tracking of model performance and population shifts may be managed through the implementation of active surveillance systems. As our study suggests that models based on different methods have variable updating needs, such surveillance systems should be tailored to the distinct susceptibilities modeling methods to performance drift. Of course, we also emphasize the importance of local knowledge for triggering updating as local clinical and organizational practice changes require. Finally, we recommend flexibility of updating protocols in order to address miscalibration as it occurs rather than at scheduled intervals.

Efficient and effective updating protocols will be essential for maintaining accuracy of and user confidence in personalized risk predictions integrated into clinical decision support for hospital-acquired acute kidney injury, 30-day all-cause hospital mortality, and other clinical outcomes. Development of automated model surveillance tools remains an open area of methods development, user-centered design, and best practices research. This work supports the advancement of predictive analytics for personalized clinical decision support by laying ground work for such automated, EHR-embedded surveillance frameworks implementing models based on advanced regression and machine learning techniques. While the 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 updated.

APPENDIX A

ACUTE KIDNEY INJURY PREDICTOR SET

	Time window for data capture			
	Not time dependent	**Preadmission (12 months)**	**Preadmission (90-days)**	**Admission window [a]**
Demographics				
Age at admission (years)	X			
Gender	X			
Race	X			
Health history				
Advanced liver disease		X		
Alcoholism		X		
Anemia		X		
Cancer		X		
Cardiovascular disease		X		
Cerebrovascular accident		X		
Chronic obstructive		X		
pulmonary disease		X		
Congestive heart failure		X		
Dementia		X		
Diabetes mellitus		X		
Dyslipidemia		X		
Hemiplegia		X		
Hepatitis		X		
HIV		X		
Hypertension (admission)		X		
Hypertension (preadmission)		X		
Hypotension(admission)		X		
Mitral valve regurgitation		X		
Peptic ulcer disease		X		
Peripheral vascular disease		X		
Rheumatoid arthritis		X		
Medications				
ACEi			X	X
Acyclovir				X
Aminoglycosides			X	X

	Time window for data capture			
	Not time dependent	**Preadmission (12 months)**	**Preadmission (90-days)**	**Admission window** [a]
Angiotensin II receptor blocker			X	X
Anhydrase diuretic			X	X
Anti tuberculosis			X	X
Antiemetics			X	X
Antifungals			X	X
Benzodiazepines			X	X
Beta blockers			X	X
Calcium channel blockers			X	X
Cephalosporins			X	X
Cimetidine				X
Cyclosporine				X
Fluoroquinolones			X	X
Glucocorticoids			X	X
Insulin			X	X
K-sparing diuretics			X	X
Lincomycin			X	X
Lithium				X
Loop diuretics			X	X
Macrolides			X	X
MAOIs			X	X
Nacetylcysteine				X
Nitrofurantoin			X	X
NSAIDs			X	X
Opioids			X	X
Penicillins			X	X
Statins			X	X
Sulfa antibiotics			X	X
Tetracyclines			X	X
Thiazides			X	X
Tricyclics			X	X
Trimethoprim				X
Vancomycin				X
Laboratory values [b] Glomerular filtration rate				X
Count			X	X
Mean			X	X

	Time window for data capture			
	Not time dependent	**Preadmission (12 months)**	**Preadmission (90-days)**	**Admission window [a]**
Standard deviation			X	X
Delta				X
Alanine aminotransferase				X
Albumin				X
Alkaline phosphatase				X
Aspartate aminotransferase				X
Bicardbonate				X
Blood urea nitrogen				X
Calcium				X
Chloride				X
Direct Bilirubin				X
Glucose				X
Hematocrit				X
Hemoglobin				X
Delta hemoglobin				X
Mean corpuscular hemoglobin				X
Mean corpuscular hemoglobin concentration				X
Mean corpuscular volume				X
Platelets				X
Sodium				X
White blood cell count				X
Other Body mass index (mean)		X		X
Intravenous fluids				
Normal saline				X
Half normal saline				X
Lactate ringers				X
Water				X
Temperature (max)			X	X

[a] 24 hours before admission to 48 hours after admission for vital signs, GFR, and medications; 5 days before admission to 48 hours after admission for other laboratory values

[b] Most recent value in window, except for glomerular filtration rate

APPENDIX B

30-DAY ALL-CAUSE MORTALITY AFTER HOSPITAL ADMISSION PREDICTOR SET

Predictor	Time window for data capture				
	Not time dependent	**Preadmission (12 months)**	**Preadmission (all available history)**	**Admission window [a]**	**Admission (last recorded value)**
Demographics Age at admission (years)	X				
Gender	X				
Race	X				
Health history AIDS/HIV (Elixhauser)			X		
Alcohol abuse (Elixhauser)			X		
Blood loss anemia (Elixhauser)			X		
Cardiac arrhythmias (Elixhauser)			X		
Chronic pulmonary disease (Elixhauser)			X		
Coagulopathy (Elixhauser)			X		
Congestive heart failure (Elixhauser)			X		
Deficiency anemia (Elixhauser)			X		
Depression (Elixhauser)			X		
Diabetes, complicated (Elixhauser)			X		
Diabetes, uncomplicated (Elixhauser)			X		
Dialysis		X			
Drug abuse (Elixhauser)			X		
Dyslipidemia		X			

Predictor	Time window for data capture				
	Not time dependent	Preadmission (12 months)	Preadmission (all available history)	Admission window [a]	Admission (last recorded value)
Fluid and electrolyte disorders (Elixhauser)			X		
Hypertension, complicated (Elixhauser)			X		
Hypertension, uncomplicated (Elixhauser)			X		
Hypothyroidism (Elixhauser)			X		
Liver disease (Elixhauser)			X		
Lymphoma (Elixhauser)			X		
Metastatic cancer (Elixhauser)			X		
Other neurological disorders (Elixhauser)			X		
Paralysis (Elixhauser)			X		
Peptic ulcer disease (Elixhauser)			X		
Peripheral vascular disorder (Elixhauser)			X		
Psychoses (Elixhauser)			X		
Pulmonary circulation disorder (Elixhauser)			X		
Renal failure (Elixhauser)			X		
Rheumatoid arthritis/collagen vascular diseases (Elixhauser)			X		
Solid tumor without metastasis (Elixhauser)			X		
Valvular disease			X		

Predictor	Time window for data capture				
	Not time dependent	**Preadmission (12 months)**	**Preadmission (all available history)**	**Admission window [a]**	**Admission (last recorded value)**
(Elixhauser)					
Weight loss (Elixhauser)			X		
Admission characteristics Admission type	X				
Body mass index				X	
Health care utilization Inpatient visits (count)		X			
Outpatient visits (count)		X			
Unplanned readmissions		X			
Vitals SBP, minimum				X	
SBP, maximum				X	
DBP, minimum				X	
DBP, maximum				X	
Pulse, minimum				X	
Pulse, maximum				X	
Laboratory values [b] Alanine transaminase					X
Albumin					X
Alkaline phosphatase					X
Aspartate aminotransferase					X
Blood sugar					X
Blood urea nitrogen					X
Calcium					X
Chloride					X
Hematocrit					X
Hemoglobin					X
Mean corpuscular hemoglobin					X

Predictor	Time window for data capture				
	Not time dependent	Preadmission (12 months)	Preadmission (all available history)	Admission window [a]	Admission (last recorded value)
Mean corpuscular hemoglobin concentration					X
Mean corpuscular volume					X
Platelets					X
Potassium					X
Serum bicarbonate					X
Serum creatinine					X
Sodium					X
Total bilirubin					X
White blood cell count					X

[a] 24 hours before admission to 48 hours after admission

[b] Most recent value in window

APPENDIX C

SUMMARY OF HOSPITAL-ACQUIRED ACUTE KIDNEY INJURY PREDICTORS OVER TIME

Summary of all predictors in hospital-acquired acute kidney injury models at development (2003) and for three years across the 9-year validation period.

	2003			2006			2009			2012		
Total admissions	170,675			176,341			193,917			184,827		
Variable	**%**	**Median (IQR)**	**% Missing**	**%**	**Median (IQR)**	**% Missing**	**%**	**Median (IQR)**	**% Missing**	**%**	**Median (IQR)**	**% Missing**
Outcome												
Acute kidney injury	7.7			7.4			6.5			6.2		
Demographics												
Age at admission (years)		66 (56-77)	0.0		64 (57-76)	0.0		64 (58-76)	0.0		65 (59-76)	0.0
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		
Asian/Pacific Islander	0.9			1.1			1.2			1.1		
American Indian/Alaskan	0.8			0.9			0.9			0.9		
Unreported	3.2			3.1			3.5			4.0		
Health history												
Advanced liver disease	2.7			3.6			4.2			4.7		
Alcoholism	12.1			18.9			23.5			26.4		
Anemia	14.5			23.3			28.6			31.1		
Cancer	18.8			22.8			24.6			24.9		
Cardiovascular disease	19.8			28.3			30.9			31.3		
Cerebrovascular accident	10.7			15.8			17.7			18.7		

	2003			2006			2009			2012		
Total admissions	170,675			176,341			193,917			184,827		
Variable	**%**	**Median (IQR)**	**% Missing**	**%**	**Median (IQR)**	**% Missing**	**%**	**Median (IQR)**	**% Missing**	**%**	**Median (IQR)**	**% Missing**
Chronic obstructive pulmonary disease	24.6			30.9			34.0			35.1		
Congestive heart failure	15.2			18.6			19.7			20.0		
Dementia	5.1			5.9			6.1			5.9		
Diabetes mellitus	29.7			34.6			39.1			42.6		
Dyslipidemia	28.8			49.7			59.7			65.4		
Hemiplegia	3.0			4.2			4.5			4.7		
Hepatitis	5.8			9.2			11.0			12.1		
HIV	1.4			1.3			1.3			1.1		
Hypertension (admission)	10.8			10.5			10.3			11.1		
Hypertension (preadmission)	55.0			69.4			74.5			76.7		
Hypotension(admission)	7.5			9.1			8.7			7.2		
Mitral valve regurgitation	1.2			2.5			3.2			3.4		
Peptic ulcer disease	3.6			5.5			6.1			6.3		
Peripheral vascular disease	11.9			17.2			19.5			20.6		
Rheumatoid arthritis	2.1			2.7			3.0			3.3		
Medications												
Preadmission												
ACEi	36.9			37.9			35.8			32.5		
Aminoglycosides	1.6			1.4			1.0			0.8		
Angiotensin II receptor blocker	3.8			5.5			6.6			7.2		
Anhydrase diuretic	0.2			0.2			0.2			0.2		
Antiemetics	3.4			3.8			4.4			5.7		
Antifungals	2.7			2.8			2.7			2.7		
Anti tuberculosis	0.4			0.5			0.4			0.3		
Benzodiazepines	13.1			13.6			13.3			12.0		

	2003			2006			2009			2012		
Total admissions	170,675			176,341			193,917			184,827		
Variable	**%**	**Median (IQR)**	**% Missing**	**%**	**Median (IQR)**	**% Missing**	**%**	**Median (IQR)**	**% Missing**	**%**	**Median (IQR)**	**% Missing**
Beta blockers	37.1			43.3			43.5			42.3		
Calcium channel blockers	21.7			21.5			23.1			23.6		
Cephalosporins	5.4			5.6			4.8			4.7		
Fluoroquinolones	5.6			10.8			8.7			8.0		
Glucocorticoids	10.5			11.2			12.2			12.3		
Insulin	10.7			12.0			13.5			14.3		
K-sparing diuretics	7.8			7.5			7.2			6.7		
Lincomycin	1.5			1.8			2.0			2.3		
Loop diuretics	23.9			23.2			22.1			21.4		
Macrolides	6.1			6.2			5.9			5.8		
MAOIs	0.0			0.0			0.0			0.0		
Nitrofurantoin	0.6			0.8			1.0			1.0		
NSAIDs	20.2			18.2			17.5			17.6		
Opioids	42.9			47.7			51.1			51.7		
Penicillins	9.0			9.3			8.3			8.1		
Statins	31.9			41.7			46.2			46.4		
Sulfa antibiotics	3.8			4.4			5.3			5.2		
Tetracyclines	2.4			3.3			3.6			4.1		
Tricyclics	6.2			5.5			4.3			3.8		
Thiazides	12.7			15.0			15.3			13.5		
Admission												
ACEi	32.9			34.1			31.4			27.7		
Acyclovir	0.7			1.0			1.2			1.4		
Aminoglycosides	2.7			2.0			1.3			0.9		
Angiotensin II receptor blocker	2.9			4.5			5.3			5.7		

	2003			2006			2009			2012		
Total admissions	170,675			176,341			193,917			184,827		
Variable	**%**	**Median (IQR)**	**% Missing**	**%**	**Median (IQR)**	**% Missing**	**%**	**Median (IQR)**	**% Missing**	**%**	**Median (IQR)**	**% Missing**
Anhydrase diuretic	0.2			0.2			0.2			0.2		
Antiemetics	3.3			5.1			9.3			13.2		
Antifungals	2.3			2.4			2.3			2.3		
Anti tuberculosis	0.5			0.5			0.3			0.3		
Benzodiazepines	21.3			22.7			21.3			19.9		
Beta blockers	40.1			48.6			48.4			47.1		
Calcium channel blockers	17.0			18.3			20.9			21.5		
Cephalosporins	17.6			18.6			17.7			19.0		
Cimetidine	0.4			0.2			0.1			0.1		
Cyclosporine	0.2			0.2			0.2			0.2		
Fluoroquinolones	5.0			8.2			6.9			5.9		
Glucocorticoids	12.5			13.4			13.9			13.8		
Insulin	21.4			27.1			29.6			28.9		
K-sparing diuretics	6.2			6.2			5.7			5.4		
Lincomycin	2.7			2.4			2.0			2.1		
Lithium	0.8			0.8			0.8			0.8		
Loop diuretics	25.8			26.0			24.6			23.4		
Macrolides	6.3			6.9			6.4			6.7		
MAOIs	0.0			0.0			0.0			0.0		
Nacetylcysteine	2.6			3.8			3.5			1.2		
Nitrofurantoin	0.2			0.2			0.3			0.2		
NSAIDs	8.6			8.6			8.5			8.7		
Opioids	50.8			59.2			63.3			64.3		
Penicillins	13.0			15.0			16.2			17.8		
Statins	27.9			38.9			43.8			44.0		
Sulfa antibiotics	1.7			1.9			1.8			1.5		

	2003			2006			2009			2012		
Total admissions	170,675			176,341			193,917			184,827		
Variable	**%**	**Median (IQR)**	**% Missing**	**%**	**Median (IQR)**	**% Missing**	**%**	**Median (IQR)**	**% Missing**	**%**	**Median (IQR)**	**% Missing**
Tetracyclines	1.4			1.8			1.6			1.7		
Thiazides	7.9			9.5			9.5			8.2		
Tricyclics	3.8			3.7			3.1			2.6		
Trimethoprim	1.4			1.4			1.4			1.1		
Vancomycin	5.3			10.4			14.5			16.3		
Laboratory values												
Glomerular filtration rate												
Count, preadmission		3 (2-6)	20.3		4 (2-6)	18.1		4 (2-7)	16.6		2 (1-5)	32.5
Mean, preadmission		67.8 (53.3-83.1)	17.5		68.8 (54.1-84.4)	14.9		70.8 (55.5-86.2)	12.6		72.8 (56.9-89.5)	12.3
SD, preadmission		8.02 (4.78-12.62)	34.8		8.06 (4.83-12.66)	31.8		8.38 (5.13-13.01)	29.7		8.32 (4.61-13.36)	53.0
Count, admission		3 (2-3)	1.5		3 (2-3)	1.2		3 (2-3)	1.1		3 (2-3)	1.1
Mean, admission		75.7 (58.8-94.73)	1.5		76.77 (59.73-96.25)	1.2		79.25 (61.7-98.73)	1.1		82.1 (63.8-102.7)	1.1
SD, admission		6.68 (3.04-11.46)	20.5		6.79 (3.15-11.62)	16.8		6.87 (3.46-11.66)	14.0		7.35 (3.76-12.7)	12.4
Delta, admission		0 (0-9.9)	1.5		0 (0-11.4)	1.2		0 (0-12.1)	1.1		1.2 (-0.5-13.6)	1.1
Alanine aminotransferase		26 (17-40)	27.6		24 (16-38)	24.5		23 (16-36.25)	23.6		23 (16-37)	24.0
Albumin		3.4 (2.9-3.8)	27.0		3.4 (2.9-3.9)	24.7		3.4 (2.9-3.9)	23.3		3.4 (2.9-3.8)	22.1
Alkaline phosphatase		86 (67-115)	25.5		83 (65-110)	24.0		81 (64-108)	23.3		79 (62-105)	23.7
Aspartate aminotransferase		25 (18-40)	26.8		26 (19-39)	25.6		25 (19-39)	24.8		25 (18-39)	25.3
Bicardbonate		26 (24-29)	0.5		26.1 (24-29)	0.3		26 (24-29)	0.2		26 (24-28)	0.1
Blood urea nitrogen		15 (11-22)	6.3		15 (10.7-22)	4.7		15 (10-21)	6.9		15 (10-21)	7.1
Calcium		8.7 (8.3-9.1)	12.3		8.7 (8.3-9.1)	6.2		8.7 (8.3-9.1)	5.1		8.6 (8.2-9)	4.3

	2003			2006			2009			2012		
Total admissions	170,675			176,341			193,917			184,827		
Variable	**%**	**Median (IQR)**	**% Missing**	**%**	**Median (IQR)**	**% Missing**	**%**	**Median (IQR)**	**% Missing**	**%**	**Median (IQR)**	**% Missing**
Chloride		103 (100-106)	0.6		103 (100-106)	0.2		103 (100-106)	0.2		103 (100-106)	0.2
Direct Bilirubin		0.6 (0.4-1)	25.8		0.7 (0.42-1)	24.2		0.7 (0.4-1)	23.3		0.7 (0.4-1)	23.7
Glucose		114 (96-148)	2.0		115 (96-148)	0.8		116 (97-150)	0.5		115 (96-149)	0.4
Hematocrit		36.4 (31.9-40.8)	0.8		36.1 (31.7-40.4)	0.5		35.5 (31.2-39.8)	0.5		35.6 (31.1-39.7)	0.4
Hemoglobin		12.2 (10.6-13.7)	1.4		12.1 (10.6-13.7)	1.2		11.9 (10.4-13.4)	1.1		11.9 (10.3-13.4)	1.2
Delta hemoglobin		-0.3 (-1.3-0)	1.5		-0.4 (-1.3-0)	1.2		-0.5 (-1.4-0)	1.1		-0.6 (-1.4-0)	1.1
Mean corpuscular hemoglobin		30.5 (28.9-31.9)	1.9		30.5 (28.9-31.9)	1.2		30.6 (29-32.1)	0.9		30.5 (28.9-32.1)	1.0
Mean corpuscular Hemoglobin concentration		33.6 (32.9-34.3)	0.9		33.8 (33-34.4)	0.7		33.7 (32.9-34.4)	0.6		33.6 (32.8-34.3)	0.7
Mean corpuscular volume		90.4 (86.5-94.2)	0.9		90.3 (86.4-94.1)	0.7		90.7 (86.8-94.6)	0.6		90.9 (86.9-94.8)	0.6
Platelets		214 (162-278)	1.1		221 (167-288)	0.9		202 (153-262)	0.8		192 (145-248)	0.8
Sodium		138 (135-140)	0.2		138 (135-140)	0.1		138 (135-140)	0.1		138 (135-140)	0.1
White blood cell count		8.1 (6.2-10.8)	0.9		8.1 (6.2-10.8)	0.7		8.01 (6.1-10.7)	0.6		7.94 (6-10.5)	0.7
Other												
Body mass index (mean, 365 days preadmission)		27 (23.5-31.3)	13.2		27.3 (23.7-31.6)	10.8		27.6 (23.9-32.1)	9.4		27.9 (24.2-32.5)	9.4
Body mass index (mean, admission)		26.5 (22.74-30.93)	40.9		26.68 (22.91-31.25)	30.7		27.03 (23.15-31.73)	24.7		27.34 (23.4-32.07)	23.5
Intravenous fluids												
Normal saline		0 (0-0.5)	0.0		0 (0-1)	0.0		0 (0-1)	0.0		0.1 (0-1.3)	0.0

	2003			2006			2009			2012		
Total admissions	170,675			176,341			193,917			184,827		
Variable	**%**	**Median (IQR)**	**% Missing**	**%**	**Median (IQR)**	**% Missing**	**%**	**Median (IQR)**	**% Missing**	**%**	**Median (IQR)**	**% Missing**
Half normal saline		0 (0-0)	0.0		0 (0-0)	0.0		0 (0-0)	0.0		0 (0-0)	0.0
Lactate ringers		0 (0-0)	0.0		0 (0-0)	0.0		0 (0-0)	0.0		0 (0-0)	0.0
Water		0 (0-0.1)	0.0		0 (0-0.14)	0.0		0 (0-0.1)	0.0		0 (0-0.1)	0.0
Temperature (max, 90 days preadmission)		98.6 (97.9-99.4)	23.9		98.6 (97.9-99.4)	21.1		98.6 (98-99.2)	18.6		98.6 (98-99.2)	18.8
Temperature (max, admission)		98.9 (98.2-99.8)	5.7		98.9 (98.3-99.8)	2.5		98.8 (98.3-99.6)	1.5		98.8 (98.4-99.6)	1.2

APPENDIX D

SUMMARY OF 30-DAY ALL-CAUSE MORTALITY AFTER HOSPITAL ADMISSION PREDICTORS OVER TIME

Summary of all predictors in 30-day all-cause hospital mortality models at development (2006) and for three years across the 8-year validation period.

	2003			2006			2009			2012		
Total admissions	235,548			235,734			243,631			214,798		
Variable	**%**	**Median (IQR)**	**% Missing**	**%**	**Median (IQR)**	**% Missing**	**%**	**Median (IQR)**	**% Missing**	**%**	**Median (IQR)**	**% Missing**
Outcome												
30-day mortality	5.0		0.0	4.9		0.0	4.9		0.0	4.7		0.0
Demographics												
Age at admission (years)		61 (54 - 74)	0.0		61 (54 - 74)	0.0		63 (55 - 74)	0.0		64 (56 - 73)	0.0
Female	4.5		0.0	4.7		0.0	4.9		0.0	5.5		0.0
Race												
White	71.7		0.0	71.6		0.0	72.3		0.0	72.1		0.0
Black	19.8		0.0	20.0		0.0	19.6		0.0	19.8		0.0
Asian/Pacific Islander	1.3		0.0	1.4		0.0	1.5		0.0	1.6		0.0
American Indian/Alaskan	1.1		0.0	1.2		0.0	1.2		0.0	1.3		0.0
Unreported	6.0		0.0	5.9		0.0	5.5		0.0	5.3		0.0
Health history												
AIDS/HIV	1.2		0.0	1.3		0.0	1.1		0.0	1.2		0.0
Alcohol abuse	4.5		0.0	5.4		0.0	7.3		0.0	8.5		0.0
Blood loss anemia	1.4		0.0	1.8		0.0	2.3		0.0	2.5		0.0
Cardiac arrhythmias	8.4		0.0	10.6		0.0	15.7		0.0	19.3		0.0
Chronic pulmonary disease	28.4		0.0	32.4		0.0	38.5		0.0	41.2		0.0
Coagulopathy	5.9		0.0	7.4		0.0	9.6		0.0	11.1		0.0
Congestive heart failure	17.3		0.0	19.1		0.0	22.0		0.0	23.7		0.0

	2003			2006			2009			2012		
Total admissions	235,548			235,734			243,631			214,798		
Variable	**%**	**Median (IQR)**	**% Missing**	**%**	**Median (IQR)**	**% Missing**	**%**	**Median (IQR)**	**% Missing**	**%**	**Median (IQR)**	**% Missing**
Deficiency anemia	7.6		0.0	9.4		0.0	13.1		0.0	15.0		0.0
Depression	20.1		0.0	24.5		0.0	32.6		0.0	38.4		0.0
Diabetes, complicated	12.7		0.0	15.1		0.0	18.5		0.0	20.6		0.0
Diabetes, uncomplicated	31.5		0.0	33.3		0.0	36.9		0.0	39.0		0.0
Dialysis	2.8		0.0	3.0		0.0	3.2		0.0	3.3		0.0
Drug abuse	12.4		0.0	14.9		0.0	18.8		0.0	22.0		0.0
Dyslipidemia	41.8		0.0	49.6		0.0	61.5		0.0	66.4		0.0
Fluid and electrolyte disorders	19.0		0.0	23.9		0.0	32.8		0.0	38.1		0.0
Hypertension, complicated	8.4		0.0	11.1		0.0	16.0		0.0	18.2		0.0
Hypertension, uncomplicated	61.7		0.0	67.4		0.0	74.2		0.0	76.4		0.0
Hypothyroidism	7.1		0.0	8.3		0.0	10.2		0.0	11.6		0.0
Liver disease	9.3		0.0	11.3		0.0	14.2		0.0	16.5		0.0
Lymphoma	1.2		0.0	1.3		0.0	1.5		0.0	1.6		0.0
Metastatic cancer	3.5		0.0	3.9		0.0	4.1		0.0	4.5		0.0
Other neurological disorders	4.6		0.0	5.8		0.0	8.6		0.0	10.3		0.0
Paralysis	2.3		0.0	2.7		0.0	3.6		0.0	4.0		0.0
Peptic ulcer disease	2.2		0.0	3.0		0.0	4.1		0.0	4.6		0.0
Peripheral vascular disorder	13.0		0.0	15.3		0.0	18.8		0.0	20.3		0.0
Psychoses	12.3		0.0	13.2		0.0	14.6		0.0	15.9		0.0
Pulmonary circulation disorder	2.4		0.0	3.3		0.0	5.0		0.0	6.0		0.0
Renal failure	12.3		0.0	15.3		0.0	19.5		0.0	21.9		0.0
Rheumatoid arthritis/collagen vascular diseases	2.6		0.0	3.0		0.0	3.9		0.0	4.6		0.0
Solid tumor without metastasis	15.8		0.0	17.4		0.0	19.8		0.0	20.6		0.0

	2003			2006			2009			2012		
Total admissions	235,548			235,734			243,631			214,798		
Variable	**%**	**Median (IQR)**	**% Missing**	**%**	**Median (IQR)**	**% Missing**	**%**	**Median (IQR)**	**% Missing**	**%**	**Median (IQR)**	**% Missing**
Valvular disease	6.7		0.0	8.3		0.0	10.5		0.0	11.6		0.0
Weight loss	3.6		0.0	4.5		0.0	5.9		0.0	7.0		0.0
Admission characteristics												
Admission type												
Planned	13.8		0.0	13.8		0.0	13.8		0.0	13.9		0.0
Unplanned readmission	10.3		0.0	10.5		0.0	10.5		0.0	10.4		0.0
Unplanned, not readmission	75.8		0.0	75.7		0.0	75.7		0.0	75.7		0.0
Body mass index		27.26 (23.61-31.71)	18.5		27.35 (23.64-31.8)	14.8		27.7 (23.92-32.29)	10.0		27.87 (24.02-32.48)	8.8
Health care utilization (prior year)												
# inpatient visits		1 (0 - 2)	0.0		1 (0 - 2)	0.0		1 (0 - 2)	0.0		1 (0 - 2)	0.0
# outpatient visits		25 (12 - 45)	0.0		26 (13 - 46)	0.0		29 (14 - 52)	0.0		30 (15 - 55)	0.0
Any unplanned readmissions	24.6			25.0		0.0	25.6		0.0	25.2		0.0
Vitals (admission window)												
SBP, minimum		109 (99 - 121)	2.5		108 (98 - 120)	2.4		108 (98 - 119)	2.0		109 (99 - 120)	1.4
SBP, maximum		148 (134 - 163)	2.4		148 (134 - 163)	2.4		149 (136 - 164)	1.9		150 (137 - 165)	1.3
DBP, minimum		60 (52 - 68)	2.5		60 (53 - 69)	2.4		60 (53 - 68)	2.0		61 (54 - 69)	1.4
DBP, maximum		85 (77 - 94)	2.5		86 (77 - 94)	2.4		87 (79 - 96)	1.9		88 (80 - 96)	1.4
Pulse, minimum		67 (59 - 76)	2.5		66 (59 - 76)	2.4		66 (58 - 75)	2.0		66 (59 - 75)	1.4
Pulse, maximum		94 (82 - 106)	2.4		94 (82 - 106)	2.3		95 (83 - 107)	1.9		95 (84 - 107)	1.3
Laboratory values (last in admission window)												
Alanine transaminase		24 (17 - 36)	14.2		24 (16 - 35)	12.9		23 (16 - 34)	11.8		24 (17 - 35)	11.9
Albumin		3.8 (3.4 - 4.2)	17.2		3.8 (3.4 - 4.2)	15.9		3.8 (3.3 - 4.1)	14.1		3.8 (3.3 - 4.1)	13.0
Alkaline phosphatase		82 (66 - 104)	14.6		81 (65 - 103)	13.7		81 (65 - 104)	12.3		79 (64 - 101)	12.4

	2003		2006		2009		2012	
Total admissions	235,548		235,734		243,631		214,798	
Variable	**% Median (IQR)**	**% Missing**	**% Median (IQR)**	**% Missing**	**% Median (IQR)**	**% Missing**	**% Median (IQR)**	**% Missing**
Aspartate aminotransferase	24 (19 - 32)	15.4	24 (19 - 32)	14.4	24 (19 - 32)	12.6	23 (18 - 32)	12.8
Blood sugar	106 (93 - 132)	9.6	106 (93 - 132)	9.0	106 (93 - 133)	7.8	106 (93 - 134)	8.0
Blood urea nitrogen	16 (12 - 23)	11.7	16 (12 - 23)	11.4	16 (12 - 23)	13.4	16 (12 - 23)	13.2
Calcium	9.2 (8.8 - 9.5)	15.0	9.2 (8.8 - 9.5)	13.5	9.2 (8.8 - 9.5)	12.0	9.1 (8.8 - 9.5)	11.9
Chloride	103.2 (101 - 106)	36.4	104 (101 - 106)	23.7	103 (100 - 106)	22.4	103 (100 - 106)	23.9
Hematocrit	39.9 (35.3 - 43.6)	10.9	39.6 (35.1 - 43.4)	10.1	39.4 (34.8 - 43.1)	8.7	39.7 (35 - 43.3)	8.6
Hemoglobin	13.5 (11.8 - 14.8)	12.5	13.4 (11.8 - 14.8)	11.8	13.3 (11.6 - 14.6)	10.7	13.2 (11.5 - 14.6)	10.8
Mean corpuscular hemoglobin	30.8 (29.3 - 32.2)	40.9	30.8 (29.3 - 32.3)	26.1	30.8 (29.2 - 32.3)	24.7	30.4 (28.9 - 31.9)	26.1
Mean corpuscular hemoglobin concentration	33.8 (33.2 - 34.4)	40.4	33.9 (33.2 - 34.5)	25.5	33.7 (33 - 34.4)	24.2	33.3 (32.6 - 34)	25.7
Mean corpuscular volume	90.5 (86.8 - 94.2)	11.0	90.9 (87.2 - 94.7)	10.2	91.3 (87.4 - 95.1)	8.8	91.1 (87.2 - 95)	8.7
Platelets	242 (191 - 302)	12.4	235 (185 - 294)	11.6	214 (169 - 266)	10.3	211 (167 - 263)	9.9
Potassium	4.2 (3.9 - 4.5)	8.9	4.2 (3.9 - 4.5)	8.4	4.2 (3.9 - 4.5)	7.2	4.1 (3.9 - 4.5)	7.4
Serum bicarbonate	27 (25 - 29)	9.4	27 (24.4 - 29)	8.9	27 (25 - 29)	7.6	26.9 (24 - 29)	7.5
Serum creatinine	1.1 (0.9 - 1.3)	9.8	1.1 (0.9 - 1.33)	9.3	1.0 (0.84 - 1.3)	8.6	1.0 (0.82 - 1.3)	9.4
Sodium	139 (137 - 141)	8.8	139 (137 - 141)	8.4	139 (136 - 140)	7.1	139 (136 - 141)	7.2
Total bilirubin	0.6 (0.4 - 0.8)	15.2	0.6 (0.4 - 0.8)	14.2	0.6 (0.4 - 0.8)	12.9	0.6 (0.4 - 0.8)	12.3
White blood cell count	7.4 (5.9 - 9.3)	11.6	7.3 (5.8 - 9.2)	10.8	7.3 (5.8 - 9.1)	9.4	7.3 (5.8 - 9.1)	9.3

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. Annual review of biomedical engineering. 2006;8:567-99.
3. Matheny ME, Miller RA, Ikizler TA, Waitman LR, Denny JC, Schildcrout JS, et al. Development of inpatient risk stratification models of acute kidney injury for use in electronic health records. Medical decision making : an international journal of the Society for Medical Decision Making. 2010;30(6):639-50.
4. Kansagara D, Englander H, Salanitro A, Kagen D, Theobald C, Freeman M, et al. Risk prediction models for hospital readmission: a systematic review. Jama. 2011;306(15):1688-98.
5. Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS medicine. 2013;10(2):e1001381.
6. Toll DB, Janssen KJ, Vergouwe Y, Moons KG. Validation, updating and impact of clinical prediction rules: a review. Journal of clinical epidemiology. 2008;61(11):1085-94.
7. Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012;98(9):691-8.
8. Bjornson E, Boren J, Mardinoglu A. Personalized Cardiovascular Disease Prediction and Treatment-A Review of Existing Strategies and Novel Systems Medicine Tools. Front Physiol. 2016;7:2.
9. Moons KG, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart. 2012;98(9):683-90.
10. Muller B, Wilcke A, Boulesteix AL, Brauer J, Passarge E, Boltze J, et al. Improved prediction of complex diseases by common genetic markers: state of the art and further perspectives. Human genetics. 2016;135(3):259-72.
11. Torres FA, Pasarelli I, Cutri A, Ossorio MF, Ferrero F. Impact Assessment of a Decision Rule for Using Antibiotics in Pneumonia: A Randomized Trial. Pediatric Pulmonology. 2013;49(7).
12. Hall LM, Jung RT, Leese GP. Controlled trial of effect of documented cardiovascular risk scores on prescribing. Bmj. 2003;326(7383):251-2.
13. Feldman M, Stanford R, Catcheside A, Stotter A. The use of a prognostic table to aid decision making on adjuvant therapy for women with early breast cancer. European journal of surgical oncology : the journal of the European Society of Surgical Oncology and the British Association of Surgical Oncology. 2002;28(6):615-9.
14. Amarasingham R, Patel PC, Toto K, Nelson LL, Swanson TS, Moore BJ, et al. Allocating scarce resources in real-time to reduce heart failure readmissions: a prospective, controlled study. BMJ Qual Saf. 2013;22(12):998-1005.

15. Jarman B, Pieter D, van der Veen AA, Kool RB, Aylin P, Bottle A, et al. The hospital standardised mortality ratio: a powerful tool for Dutch hospitals to assess their quality of care? Qual Saf Health Care. 2010;19(1):9-13.
16. Sajda P. Machine learning for detection and diagnosis of disease. Annual review of biomedical engineering. 2006;8:537-65.
17. Steyerberg EW, van der Ploeg T, Van Calster B. Risk prediction with machine learning and regression methods. Biometrical journal Biometrische Zeitschrift. 2014;56(4):601-6.
18. Pencina MJ, Peterson ED. Moving From Clinical Trials to Precision Medicine: The Role for Predictive Modeling. Jama. 2016;315(16):1713-4.
19. Parikh RB, Kakad M, Bates DW. Integrating Predictive Analytics Into High-Value Care: The Dawn of Precision Delivery. Jama. 2016;315(7):651-2.
20. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J. 2015;13:8-17.
21. 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.
22. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. New York, NY: Spring; 2009.
23. Hickey GL, Grant SW, Murphy GJ, Bhabra M, Pagano D, McAllister K, 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. European Journal of Cardio-thoracic Surgery. 2013;43(6):1146-52.
24. 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 Medicine. 2012;38(1):40-6.
25. Tangri N, Kitsios GD, Inker LA, Griffith J, Naimark DM, Walker S, et al. Risk prediction models for patients with chronic kidney disease: a systematic review. Annals of internal medicine. 2013;158(8):596-603.
26. Pencina MJ, D'Agostino RB, Sr., Larson MG, Massaro JM, Vasan RS. Predicting the 30-year risk of cardiovascular disease: the framingham heart study. Circulation. 2009;119(24):3078-84.
27. Maguire JL, Kulik DM, Laupacis A, Kuppermann N, Uleryk EM, Parkin PC. Clinical prediction rules for children: a systematic review. Pediatrics. 2011;128(3):e666-77.
28. Kengne AP, Masconi K, Mbanya VN, Lekoubou A, Echouffo-Tcheugui JB, Matsha TE. Risk predictive modelling for diabetes and cardiovascular disease. Critical reviews in clinical laboratory sciences. 2014;51(1):1-12.
29. Meng X, Huang Z, Wang R, Yu J. Prediction of response to preoperative chemoradiotherapy in patients with locally advanced rectal cancer. Bioscience trends. 2014;8(1):11-23.
30. Lambin P, van Stiphout RG, Starmans MH, Rios-Velazquez E, Nalbantov G, Aerts HJ, et al. Predicting outcomes in radiation oncology--multifactorial decision support systems. Nature reviews Clinical oncology. 2013;10(1):27-40.

31. Stevenson JM, Williams JL, Burnham TG, Prevost AT, Schiff R, Erskine SD, et al. Predicting adverse drug reactions in older adults; a systematic review of the risk prediction models. Clinical interventions in aging. 2014;9:1581-93.
32. Mallett S, Royston P, Waters R, Dutton S, Altman DG. Reporting performance of prognostic models in cancer: a review. BMC medicine. 2010;8:21.
33. Ettema RG, Peelen LM, Schuurmans MJ, Nierich AP, Kalkman CJ, Moons KG. Prediction models for prolonged intensive care unit stay after cardiac surgery: systematic review and validation study. Circulation. 2010;122(7):682-9, 7 p following p 9.
34. Walsh C, Hripcsak G. The effects of data sources, cohort selection, and outcome definition on a predictive model of risk of thirty-day hospital readmissions. Journal of biomedical informatics. 2014;52:418-26.
35. Amarasingham R, Audet AJ, Bates DW, Cohen IG, Entwistle M, Escobar GJ, et al. Consensus statement on electronic health predictive analytics: a guiding framework to address challenges. eGEMs (Generating Evidence & Methods to improve patient outcomes). 2016;4(1):1-11.
36. Hendriksen JM, Geersing GJ, Moons KG, de Groot JA. Diagnostic and prognostic prediction models. Journal of thrombosis and haemostasis : JTH. 2013;11 Suppl 1:129-41.
37. van Dieren S, Beulens JW, Kengne AP, Peelen LM, Rutten GE, Woodward M, et al. Prediction models for the risk of cardiovascular disease in patients with type 2 diabetes: a systematic review. Heart. 2012;98(5):360-9.
38. Hill JC, Whitehurst DG, Lewis M, Bryan S, Dunn KM, Foster NE, et al. Comparison of stratified primary care management for low back pain with current best practice (STarT Back): a randomised controlled trial. Lancet. 2011;378(9802):1560-71.
39. Whitehurst DG, Bryan S, Lewis M, Hill J, Hay EM. Exploring the cost-utility of stratified primary care management for low back pain compared with current best practice within risk-defined subgroups. Annals of the rheumatic diseases. 2012;71(11):1796-802.
40. Harrell F. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York: Springer; 2001.
41. Harrell FE, Jr., Lee KL, Matchar DB, Reichert TA. Regression models for prognostic prediction: advantages, problems, and suggested solutions. Cancer Treat Rep. 1985;69(10):1071-77.
42. Steyerberg EW, Eijkemans MJ, Harrell FE, Jr., Habbema JD. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Statistics in medicine. 2000;19(8):1059-79.
43. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. Journal of clinical epidemiology. 1996;49(12):1373-9.
44. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC medical research methodology. 2014;14:137.
45. 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. Statistics in medicine. 2004;23(16):2567-86.

46. Siregar S, Nieboer D, Vergouwe Y, Versteegh M, Noyez L, Vonk A, et al. Improved prediction by dynamic modelling: An exploratory study in the adult cardiac surgery database of the netherlands association for cardio-thoracic surgery. Interactive Cardiovascular and Thoracic Surgery. 2014;19:S8.
47. Breiman L. Statistical modeling: the two cultures. Statistical Science. 2001;16(3):199-231.
48. Binder H. What subject matter questions motivate the use of machine learning approaches compared to statistical models for probability prediction? Biometrical journal Biometrische Zeitschrift. 2014;56(4):584-7.
49. Malley JD, Kruppa J, Dasgupta A, Malley KG, Ziegler A. Probability machines: consistent probability estimation using nonparametric learning machines. Methods of information in medicine. 2012;51(1):74-81.
50. Ganjisaffar Y, Debeauvais T, Javanmardi S, Caruana R, Lopes CV, editors. Distributed Tuning of Machine Learning Algorithms using MapReduce Clusters. Proceedings of the Third Workshop on Large Scale Data Mining: Theory and Applications; 2011 August 21 - 21, 2011; San Diego, CA.
51. Ribeiro MT, Singh S, Guestrin C, editors. Why Should I Trust You?: Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016: ACM.
52. Kruppa J, Liu Y, Diener HC, Holste T, Weimar C, Konig IR, et al. Probability estimation with machine learning methods for dichotomous and multicategory outcome: applications. Biometrical journal Biometrische Zeitschrift. 2014;56(4):564-83.
53. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. International journal of medical informatics. 2008;77(2):81-97.
54. Valdes G, Luna JM, Eaton E, Simone CB. MediBoost: a Patient Stratification Tool for Interpretable Decision Making in the Era of Precision Medicine. Sci Rep. 2016;6.
55. Luo G. Automatically explaining machine learning prediction results: a demonstration on type 2 diabetes risk prediction. Health Inf Sci Syst. 2016;4:2.
56. Ross EG, Shah NH, Dalman RL, Nead KT, Cooke JP, Leeper NJ. The use of machine learning for the identification of peripheral artery disease and future mortality risk. J Vasc Surg. 2016;64(5):1515-22 e3.
57. VanHouten JP, Starmer JM, Lorenzi NM, Maron DJ, Lasko TA. Machine learning for risk prediction of acute coronary syndrome. AMIA Annu Symp Proc. 2014;2014:1940-9.
58. Singal AG, Mukherjee A, Elmunzer BJ, Higgins PD, Lok AS, Zhu J, et al. Machine learning algorithms outperform conventional regression models in predicting development of hepatocellular carcinoma. Am J Gastroenterol. 2013;108(11):1723-30.
59. Kim SK, Yoo TK, Oh E, Kim DW. Osteoporosis risk prediction using machine learning and conventional methods. Conference proceedings : Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Annual Conference. 2013;2013:188-91.
60. Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. Journal of biomedical informatics. 2002;35(5-6):352-9.

61. Austin PC, Lee DS, Steyerberg EW, Tu JV. Regression trees for predicting mortality in patients with cardiovascular disease: what improvement is achieved by using ensemble-based methods? Biometrical journal Biometrische Zeitschrift. 2012;54(5):657-73.
62. Cronin RM, VanHouten JP, Siew ED, Eden SK, Fihn SD, Nielson CD, et al. National Veterans Health Administration Inpatient Risk Stratification Models for Hospital-Acquired Acute Kidney Injury. Journal of the American Medical Informatics Association. 2015;22(5):1054-71.
63. van der Ploeg T, Nieboer D, Steyerberg EW. Modern modeling techniques had limited external validity in predicting mortality from traumatic brain injury. Journal of clinical epidemiology. 2016;78:83-9.
64. Ennis M, Hinton G, Naylor D, Revow M, Tibshirani R. A comparison of statistical learning methods on the Gusto database. Statistics in medicine. 1998;17(21):2501-8.
65. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1):128-38.
66. Matheny ME, Ohno-Machado L, Resnic FS. Discrimination and calibration of mortality risk prediction models in interventional cardiology. Journal of biomedical informatics. 2005;38(5):367-75.
67. Jiang X, Osl M, Kim J, Ohno-Machado L. Calibrating predictive model estimates to support personalized medicine. Journal of the American Medical Informatics Association : JAMIA. 2012;19(2):263-74.
68. Van Calster B, Vickers AJ. Calibration of risk prediction models: impact on decision-analytic performance. Medical decision making : an international journal of the Society for Medical Decision Making. 2015;35(2):162-9.
69. Bouwmeester W, Zuithoff NP, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, et al. Reporting and methods in clinical prediction research: a systematic review. PLoS medicine. 2012;9(5):1-12.
70. Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC medical research methodology. 2014;14:40.
71. 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. Journal of clinical epidemiology. 2016.
72. 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. Journal of biomedical informatics. 2015;54:283-93.
73. 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. Journal of clinical epidemiology. 2015;68(3):279-89.
74. 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 of information in medicine. 2012;51(4):353-8.
75. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29-36.

76. Cook DA, Joyce CJ, Barnett RJ, Birgan SP, Playford H, Cockings JG, 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.
77. Knaus WA, Wagner DP, Draper EA, Zimmerman JE, Bergner M, Bastos PG, et al. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. Chest. 1991;100(6):1619-36.
78. Harrison DA, Brady AR, Parry GJ, Carpenter JR, Rowan K. Recalibration of risk prediction models in a large multicenter cohort of admissions to adult, general critical care units in the United Kingdom. Crit Care Med. 2006;34(5):1378-88.
79. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. Critical care medicine. 1985;13(10):818-29.
80. Le Gall JR, Lemeshow S, Saulnier F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. Jama. 1993;270(24):2957-63.
81. Lemeshow S, Teres D, Klar J, Avrunin JS, Gehlbach SH, Rapoport J. Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients. Jama. 1993;270(20):2478-86.
82. Harrison DA, Lone NI, Haddow C, MacGillivray M, Khan A, Cook B, et al. External validation of the Intensive Care National Audit & Research Centre (ICNARC) risk prediction model in critical care units in Scotland. BMC anesthesiology. 2014;14:116.
83. Hekmat K, Kroener A, Stuetzer H, Schwinger RH, Kampe S, Bennink GB, et al. Daily assessment of organ dysfunction and survival in intensive care unit cardiac surgical patients. The Annals of thoracic surgery. 2005;79(5):1555-62.
84. Marshall JC, Cook DJ, Christou NV, Bernard GR, Sprung CL, Sibbald WJ. Multiple organ dysfunction score: a reliable descriptor of a complex clinical outcome. Critical care medicine. 1995;23(10):1638-52.
85. Hekmat K, Doerr F, Kroener A, Heldwein M, Bossert T, Badreldin AMA, et al. Prediction of mortality in intensive care unit cardiac surgical patients. European Journal of Cardio-thoracic Surgery. 2010;38(1):104-9.
86. Hickey GL, Grant SW, Caiado C, Kendall S, Dunning J, Poullis M, et al. Dynamic prediction modeling approaches for cardiac surgery. Circulation: Cardiovascular Quality and Outcomes. 2013;6(6):649-58.
87. Roques F, Michel P, Goldstone AR, Nashef SA. The logistic EuroSCORE. European heart journal. 2003;24(9):881-2.
88. Madan P, Elayda MA, Lee VV, Wilson JM. Risk-prediction models for mortality after coronary artery bypass surgery: application to individual patients. International journal of cardiology. 2011;149(2):227-31.
89. McCormick TH, Raftery AE, Madigan D, Burd RS. Dynamic Logistic Regression and Dynamic Model Averaging for Binary Classification. Biometrics. 2012;68(1):23-30.
90. Mikkelsen MM, Johnsen SP, Nielsen PH, Jakobsen CJ. The EuroSCORE in western Denmark: a population-based study. J Cardiothorac Vasc Anesth. 2012;26(2):258-64.
91. de Rooij SE, Abu-Hanna A, Levi M, de Jonge E. Identification of high-risk subgroups in very elderly intensive care unit patients. Critical care. 2007;11(2):R33.

92. 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):353-8.
93. Osswald BR, Gegouskov V, Badowski-Zyla D, Tochtermann U, Thomas G, Hagl S, et al. Overestimation of aortic valve replacement risk by EuroSCORE: implications for percutaneous valve replacement. European heart journal. 2009;30(1):74-80.
94. Nashef SA, Roques F, Michel P, Gauducheau E, Lemeshow S, Salamon R. European system for cardiac operative risk evaluation (EuroSCORE). European journal of cardio-thoracic surgery : official journal of the European Association for Cardio-thoracic Surgery. 1999;16(1):9-13.
95. 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 and intensive care. 2012;40(6):980-94.
96. Rogers FB, Osler T, Krasne M, Rogers A, Bradburn EH, Lee JC, et al. Has TRISS become an anachronism? A comparison of mortality between the National Trauma Data Bank and Major Trauma Outcome Study databases. Journal of Trauma and Acute Care Surgery. 2012;73(2):326-31.
97. Champion HR, Sacco WJ, Copes WS. Injury severity scoring again. The Journal of trauma. 1995;38(1):94-5.
98. Hagendorf BA, Liao JG, Price MR, Burd RS. Evaluation of race and insurance status as predictors of undergoing laparoscopic appendectomy in children. Annals of surgery. 2007;245(1):118-25.
99. Barili F, Capo A, Ardemagni E, Rosato F, Grossi C. Trend analysis of euroscore performance: A prospective tenyear experience. Giornale Italiano di Cardiologia. 2012;2):171S.
100. Arvis P, Lehert P, Guivarc HLA. Simple adaptations to the Templeton model for IVF outcome prediction make it current and clinically useful. Human Reproduction. 2012;27(10):2971-8.
101. Bellomo R, Ronco C, Kellum JA, Mehta RL, Palevsky P, Acute Dialysis Quality Initiative w. Acute renal failure - definition, outcome measures, animal models, fluid therapy and information technology needs: the Second International Consensus Conference of the Acute Dialysis Quality Initiative (ADQI) Group. Critical care. 2004;8(4):R204-12.
102. Uchino S, Kellum JA, Bellomo R, Doig GS, Morimatsu H, Morgera S, et al. Acute renal failure in critically ill patients: a multinational, multicenter study. Jama. 2005;294(7):813-8.
103. Hou SH, Bushinsky DA, Wish JB, Cohen JJ, Harrington JT. Hospital-acquired renal insufficiency: a prospective study. The American journal of medicine. 1983;74(2):243-8.
104. Nash K, Hafeez A, Hou S. Hospital-acquired renal insufficiency. American journal of kidney diseases : the official journal of the National Kidney Foundation. 2002;39(5):930-6.
105. 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. American journal of kidney diseases : the official journal of the National Kidney Foundation. 2009;53(6):961-73.

106. Sawhney S, Mitchell M, Marks A, Fluck N, Black C. Long-term prognosis after acute kidney injury (AKI): what is the role of baseline kidney function and recovery? A systematic review. BMJ Open. 2015;5(1):e006497.
107. Shusterman N, Strom BL, Murray TG, Morrison G, West SL, Maislin G. Risk factors and outcome of hospital-acquired acute renal failure. Clinical epidemiologic study. The American journal of medicine. 1987;83(1):65-71.
108. 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 international Supplement. 1998;66:S16-24.
109. 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. Critical care medicine. 1996;24(2):192-8.
110. Breidthardt T, Christ-Crain M, Stolz D, Bingisser R, Drexler B, Klima T, et al. A combined cardiorenal assessment for the prediction of acute kidney injury in lower respiratory tract infections. The American journal of medicine. 2012;125(2):168-75.
111. Kim WH, Lee SM, Choi JW, Kim EH, Lee JH, Jung JW, et al. Simplified clinical risk score to predict acute kidney injury after aortic surgery. J Cardiothorac Vasc Anesth. 2013;27(6):1158-66.
112. Kristovic D, Horvatic I, Husedzinovic I, Sutlic Z, Rudez I, Baric D, 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.
113. McMahon GM, Zeng X, Waikar SS. A risk prediction score for kidney failure or mortality in rhabdomyolysis. JAMA internal medicine. 2013;173(19):1821-8.
114. Ng SY, Sanagou M, Wolfe R, Cochrane A, Smith JA, Reid CM. Prediction of acute kidney injury within 30 days of cardiac surgery. The Journal of thoracic and cardiovascular surgery. 2014;147(6):1875-83, 83 e1.
115. Park MH, Shim HS, Kim WH, Kim HJ, Kim DJ, Lee SH, 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.
116. 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 journal of surgery. 2013;37(11):2618-28.
117. 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.
118. 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.
119. 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. Journal of burn care & research : official publication of the American Burn Association. 2012;33(2):242-51.
120. Legrand M, Pirracchio R, Rosa A, Petersen ML, Van der Laan M, Fabiani JN, et al. Incidence, risk factors and prediction of post-operative acute kidney injury following

cardiac surgery for active infective endocarditis: an observational study. Critical care. 2013;17(5):R220.
121. Brown JR, MacKenzie TA, Maddox TM, Fly J, Tsai TT, Plomondon ME, et al. Acute Kidney Injury Risk Prediction in Patients Undergoing Coronary Angiography in a National Veterans Health Administration Cohort With External Validation. Journal of the American Heart Association. 2015;4(12).
122. 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. Journal of the American College of Cardiology. 2013;61(22):2242-8.
123. Wilson T, Quan S, Cheema K, Zarnke K, Quinn R, de Koning L, et al. Risk prediction models for acute kidney injury following major noncardiac surgery: systematic review. Nephrology, dialysis, transplantation : official publication of the European Dialysis and Transplant Association - European Renal Association. 2016;31(2):231-40.
124. Hall MJ, Levant S, DeFrances CJ. Trends in inpatient hospital deaths: National Hospital Discharge Survey, 2000–2010. NCHS data brief, no 118. Hyattsville, MD: National Center for Health Statistics; 2013.
125. Yale New Haven Health Services Corporation Center for Outcomes Research and Evaluation. Medicare hospital quality chartbook performance report on outcome measures Centers for Medicare and Medicaid Services; 2014.
126. Taylor RA, Pare JR, Venkatesh AK, Mowafi H, Melnick ER, Fleischman W, et al. Prediction of In-hospital Mortality in Emergency Department Patients With Sepsis: A Local Big Data-Driven, Machine Learning Approach. Acad Emerg Med. 2016;23(3):269-78.
127. Pirracchio R, Petersen ML, Carone M, Rigon MR, Chevret S, van der Laan MJ. Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study. Lancet Respir Med. 2015;3(1):42-52.
128. Liu Y, Traskin M, Lorch SA, George EI, Small D. Ensemble of trees approaches to risk adjustment for evaluating a hospital's performance. Health Care Manag Sci. 2015;18(1):58-66.
129. Perlin JB, Kolodner RM, Roswell RH. The Veterans Health Administration: quality, value, accountability, and information as transforming strategies for patient-centered care. The American journal of managed care. 2004;10(11 Pt 2):828-36.
130. Fihn SD, Francis J, Clancy C, Nielson C, Nelson K, Rumsfeld J, et al. Insights from advanced analytics at the Veterans Health Administration. Health affairs. 2014;33(7):1203-11.
131. Khwaja A. KDIGO clinical practice guidelines for acute kidney injury. Nephron Clinical practice. 2012;120(4):c179-84.
132. Yale New Haven Health Services Corporation Center for Outcomes Research and Evaluation. 2015 Condition-Specific Measures Updates and Specifications Report Hospital-Level 30-Day Risk-Standardized Mortality Measures - Version 4.0. 2015.
133. Dalton JE, Kurz A, Turan A, Mascha EJ, Sessler DI, Saager L. Risk quantification for 30-day postoperative mortality and norbidity in non-cardiac surgical patients. Anesthesiology. 2011;114(6):1336-44.
134. Nashef SA, Roques F, Sharples LD, Nilsson J, Smith C, Goldstone AR, et al. EuroSCORE II. European Journal of Cardio-thoracic Surgery. 2012;41(4):734-44.

135. Yale New Haven Health Services Corporation/Center for Outcomes Research & Evaluation. 2015 Measure Updates and Specifications Report. Hospital-Wide All-Cause Unplanned Readmission Measure – Version 4.0 2015.
136. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. Med Care. 1998;36(1):8-27.
137. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 1970;12(1):55-67.
138. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B. 2005;67(2):301-20.
139. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B. 1996;58(1):267–88.
140. Zhang H. The optimality of naive Bayes. FLAIRS Conference: American Association for Artificial Intelligence; 2004.
141. Domingos P, Pazzani M. On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning. 1998;29(2):103-30.
142. Niculescu-Mizil A, Caruana R, editors. Predicting good probabilities with supervised learning. 22nd International Conference on Machine Learning; 2005; Bonn, Germany.
143. Bishop CM. Neural Networks for Pattern Recognition: Oxford University Press; 1995.
144. Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. Journal of clinical epidemiology. 1996;49(11):1225-31.
145. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and Regression Trees. New York: Chapman and Hall/CRC; 1984.
146. Breiman L. Random forests. Machine Learning. 2001;45(1):5-32.
147. Cox DR. Two further applications of a model for binary regression. Biometrika. 1958;45:562–5.
148. Vergouwe Y, Moons KG, Steyerberg EW. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. American journal of epidemiology. 2010;172(8):971-80.
149. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. Statistics in medicine. 1997;16(9):965-80.
150. Paul P, Pennell ML, Lemeshow S. Standardizing the power of the Hosmer-Lemeshow goodness of fit test in large data sets. Statistics in medicine. 2013;32(1):67-80.
151. Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. Statistics in medicine. 2014;33(3):517-35.
152. Meinshausen N, Bühlmann P. Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2010;72(4):417-73.
153. Kappen TH, Vergouwe Y, van Klei WA, van Wolfswinkel L, Kalkman CJ, Moons KG. Adaptation of clinical prediction models for application in local settings. Medical decision making : an international journal of the Society for Medical Decision Making. 2012;32(3):E1-10.
154. Hannan EL, Cozzens K, King SB, 3rd, Walford G, Shah NR. The New York State cardiac registries: history, contributions, limitations, and lessons for future efforts to assess and

publicly report healthcare outcomes. Journal of the American College of Cardiology. 2012;59(25):2309-16.
155. Jin R, Furnary AP, Fine SC, Blackstone EH, Grunkemeier GL. Using Society of Thoracic Surgeons risk models for risk-adjusting cardiac surgery results. The Annals of thoracic surgery. 2010;89(3):677-82.
156. Janssen KJ, Moons KG, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. Journal of clinical epidemiology. 2008;61(1):76-86.