Journal of Biomedical Informatics 112 (2020) 103611

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Detection of calibration drift in clinical prediction models to inform model updating

Sharon E. Davis [a,*], Robert A. Greevy Jr. [b], Thomas A. Lasko [a], Colin G. Walsh [a,c,d], Michael E. Matheny [a,b,c,e]

[a] Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA
[b] Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, USA
[c] Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA
[d] Department of Psychiatry, Vanderbilt University Medical Center, Nashville, TN, USA
[e] Geriatrics Research, Education, and Clinical Care, Tennessee Valley Healthcare System VA, Nashville, TN, USA

ARTICLE INFO

Keywords:
Predictive analytics
Calibration
Drift detection
Model updating

ABSTRACT

Model calibration, critical to the success and safety of clinical prediction models, deteriorates over time in response to the dynamic nature of clinical environments. To support informed, data-driven model updating strategies, we present and evaluate a calibration drift detection system. Methods are developed for maintaining dynamic calibration curves with optimized online stochastic gradient descent and for detecting increasing miscalibration with adaptive sliding windows. These methods are generalizable to support diverse prediction models developed using a variety of learning algorithms and customizable to address the unique needs of clinical use cases. In both simulation and case studies, our system accurately detected calibration drift. When drift is detected, our system further provides actionable alerts by including information on a window of recent data that may be appropriate for model updating. Simulations showed these windows were primarily composed of data accruing after drift onset, supporting the potential utility of the windows for model updating. By promoting model updating as calibration deteriorates rather than on pre-determined schedules, implementations of our drift detection system may minimize interim periods of insufficient model accuracy and focus analytic resources on those models most in need of attention.

1. Introduction

Electronic health record-embedded predictive analytics promise to improve health outcomes by supporting clinical care, patient and provider decision-making, and population management [1–3]. Realizing this promise requires methods to address key challenges to widespread, effective implementation. Much of the work establishing best practices for clinical prediction applications focuses on the crucial phases of model development and validation [4–6]. However, the utility of prediction models is critically dependent on successful implementation, maintenance, and de-implementation strategies [3]. Best practices addressing these later phases of the clinical predictive analytics cycle are yet to be fully developed and further research is needed to address the unique challenges of clinical environments [3,7].

One such challenge results from model calibration, increasingly recognized as critical to the success and safety of clinical deployment of prediction models [1,8–10], deteriorating over time [11–17]. This calibration drift is a consequence of deploying models in non-stationary clinical environments where differences arise over time between the population on which a model was developed and the population to which that model is applied [5,18–23]. Abrupt changes result when models are transported across clinical settings, new clinical guidelines are implemented, or information systems are updated. More gradual changes in clinical environments may occur as demographics shifts, new practice patterns emerge, or workflows evolve [5,18,20,21,23]. Most recently, the sweeping changes in care delivery during the COVID-19 pandemic highlight how quickly and significantly the environment in which models function can shift under simultaneous changes in data collection, patient case mix, and clinical decision-making [24–27].

There are existing strategies for mitigating these impacts, most

* Corresponding author at: 2525 West End Ave, Suite 1475, Nashville, TN 37203, USA.
E-mail addresses: sharon.e.davis@vanderbilt.edu (S.E. Davis), robert.greevy@vanderbilt.edu (R.A. Greevy), tom.lasko@vanderbilt.edu (T.A. Lasko), colin.walsh@vanderbilt.edu (C.G. Walsh), michael.matheny@vanderbilt.edu (M.E. Matheny).

https://doi.org/10.1016/j.jbi.2020.103611
Received 14 August 2020; Received in revised form 21 October 2020; Accepted 29 October 2020
Available online 4 November 2020

commonly predefining the interval and method for model updating or refitting [19,28,29]. However, this approach pays little or no attention to model performance between scheduled maintenance points; overlooks guidance recommending recalibration over model refitting [5,21,22,30,31]; and fails to account for variations in the response of different learning algorithms to changes in clinical environments [11,12,32]. As a result, these types of fixed updating or refitting methods are inefficient and sometimes even detrimental [5,7,19,20,33].

Data-driven updating strategies can address the limitations of scheduled refitting by tailoring updates around the timing, extent, and form of observed performance drift. Such strategies require methods to determine both how and when models should be updated. Testing procedures addressing the former question can provide data-driven guidance on selecting between refitting and recalibration [31,33]. Scheduled updates applying such test recommendations improved calibration over time compared to a predefined refitting approach [31,33,34]. However, these testing procedures do not address when updating should be considered nor what window of recent observations may be appropriate for updating. During periods of rapid performance drift, waiting for scheduled updating points may allow for unacceptably long durations of reduced accuracy in the interim. On the other hand, during periods of relatively stable performance, scheduled updates may result in unnecessary efforts updating well-performing models.

We developed a calibration drift detection system to alert users to deteriorating model performance. Our detector monitors a detailed, stringent calibration measure through the implementation of dynamic

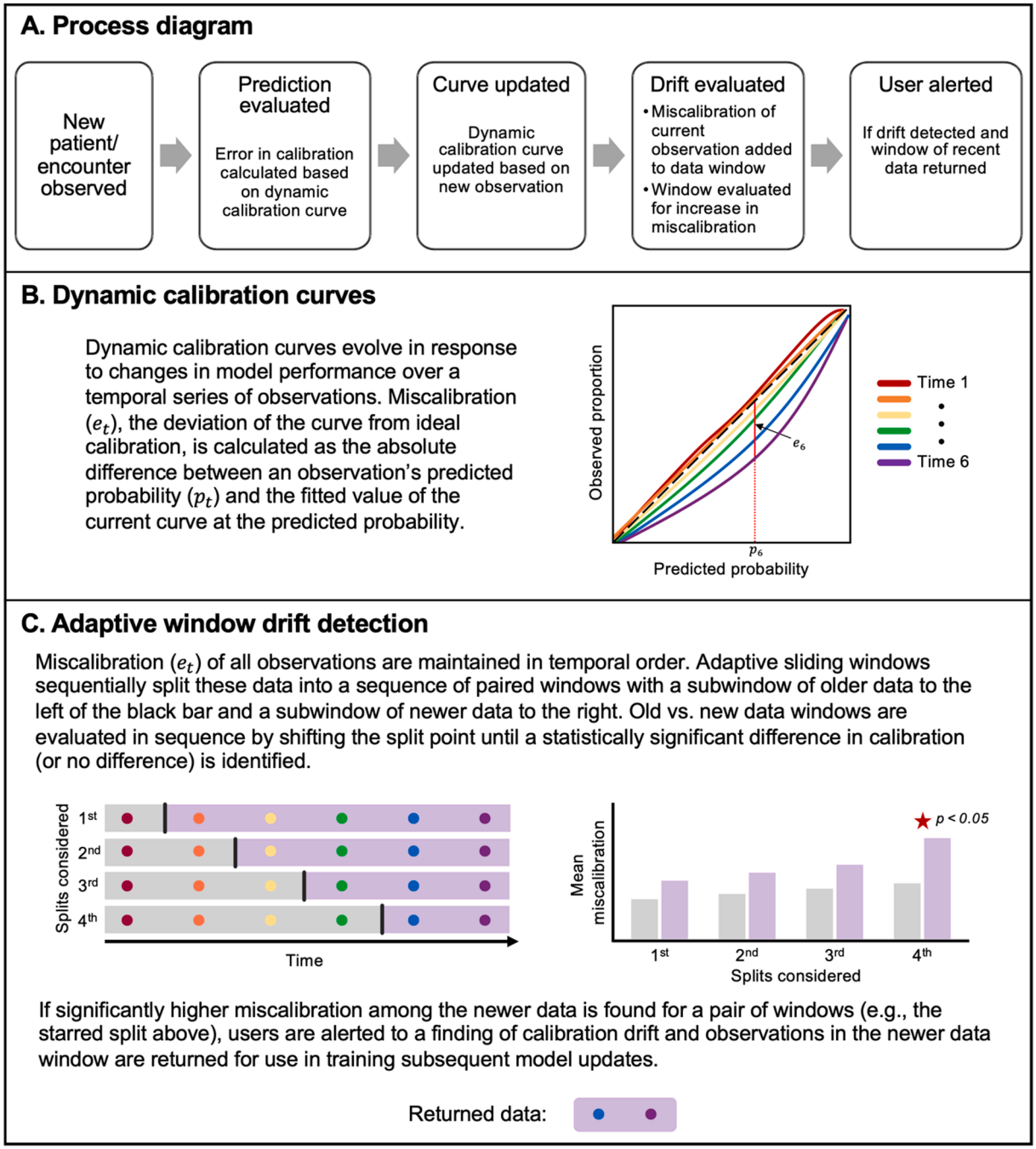

Fig. 1. Overview of the calibration drift detection process. (A) Flow of the process for a single observation. (B) Illustration of evolving dynamic calibration curves and calculation of the error metric as deviation of the curve from ideal calibration line. (C) Illustration of the drift detection process highlighting the sequence paired subwindows evaluated for differences in error.

calibration curves, a new method to maintain an up-to-date representation of model performance as it evolves over time. In accumulating observations, this calibration measure is monitored for drift by a customized adaptive windowing (Adwin) [35] implementation that provides a one-sided test for increasing miscalibration. In addition to alerting users to the presence of calibration drift, our system further supports model managers by providing actionable alerts with information on a window of recent data that may be appropriate for model updating. In this paper, we describe our dynamic calibration curves and Adwin implementation for calibration drift detection and evaluate the properties of this approach in both simulated and real-world data.

2. A system for calibration monitoring and drift detection

2.1. Design overview

We designed a scalable system for use in prospective, production environments that monitors a set of risk prediction models by providing data-driven guidance on the timing of and data to be used for model updating. This design required adapting and extending multiple methods in order to (1) prospectively and iteratively update the assessment of model calibration as data accumulates over time; (2) evaluate sequential calibration assessments for significant drift; and (3) determine the interval of data that should be used to update the existing model. Throughout the design, we emphasize practical, efficient methods to avoid computational or analytical resource burdens; generalizable methods to support diverse prediction models developed using a variety of learning algorithms; and customizable methods to address the unique needs of clinical use cases.

Fig. 1 provides an overview of our calibration drift detection system and illustrates the process at two unique timepoints. A prediction is generated as a new patient's data becomes available. The error of this prediction is estimated from the current dynamic calibration curve. The dynamic calibration curve is subsequently updated once the patient's outcome becomes available. The calibration error for the observation is submitted to an adaptive windowing [35] monitor which triggers an alert when a significant increase in error is discovered. In addition to alerting the user to calibration drift, the system returns a window of recent data in which there are no further statistically significant increases in error [35]. We propose this returned window as a candidate for use in any subsequent updating process. Although our system is designed to support model updating, for the purposes of this study, we focus on detecting calibration drift and leave the updating response to other ongoing research.

2.2. Dynamic calibration curves

2.2.1. Motivation

Calibration curves are graphical representations of model performance across the range of predicted probability. These curves are constructed by regressing observed outcomes on predicted probabilities with loess smoothing or logistic regression. For perfectly calibrated models, such curves fall along the bisector of a plot of observed outcome rates against predicted probabilities [36–38]. The fitted value of a calibration curve at a given predicted probability thus provides an estimate of the observed probability of the outcome among patients with similar predicted risk [37]. Calibration curves are typically constructed once on a prespecified cohort of validation data. However, when applying prediction models prospectively, validation data arrives in a streaming fashion as new patient encounters are recorded. Building new calibration curves on recent data as each new observation arrives could become computationally burdensome and would require assumptions regarding the appropriate batch of recent observations to consider when building each new curve. Given these challenges, implementing calibration curves in the streaming context and with an interest in how these curves change over time requires a new approach to curve construction.

2.2.2. Approach

Our method for dynamic calibration curves maintains evolving logistic calibration curves using online stochastic gradient descent with Adam optimization [39]. Gradient descent estimates the coefficients of a logistic regression by incrementally adjusting coefficients toward those values that minimize the logistic loss function [40]. As a model's calibration changes over time, the true coefficients of the logistic calibration curve change to reflect the new, current association between predictions and observed outcomes. Online gradient descent can respond to such changes by processing observations in temporal order and stepping coefficient estimates toward newly optimal values that reflect the current the loss observed among recent data [41,42]. The learning rate of the gradient descent process is critical to determining how quickly coefficient estimates would step toward new optimal values. In non-stationary environments, a constant learning rate may not be appropriate given preferences to learn more quickly during periods of change and more slowly during periods of stability [41]. Adam addresses this concern by scaling each coefficient's learning rate by the exponentially weighted moving averages of the gradient and squared gradient [39]. The step size, or initial learning rate, of an Adam implementation allows users to balance speed of learning and variability of the fitted curve. Small step sizes reduce the speed at which parameters can move toward ideal values, whereas large step sizes may lead to variability in parameter values from iteration to iteration. Adam optimization is fast, computationally efficient, and widely implemented in machine learning applications.

The logistic model underlying dynamic calibration curves may be parameterized with any number of linear, spline, or polynomial expansions of model predictions [37,38,43,44]. Here we define a default parameterization with fractional polynomials. This parameterization avoids concern that the knots of splines may require repositioning over time and better captures complex nonlinear associations than traditional polynomials [45]. In simulations (see Appendix A), we found a 5-degree fractional polynomial of the form {0.5, 0.5, 0.5, 0.5, 0.5} was able to represent both simple and complex forms of miscalibration. This fractional polynomial curve takes the following form:

$$logit(y) = \beta_0 + \beta_1\sqrt{p} + \beta_2\sqrt{p}log(p) + \beta_3\sqrt{p}log(p)^2 + \beta_4\sqrt{p}log(p)^3 + \beta_5\sqrt{p}log(p)^4$$

where p is the predicted probability and y is the observed dichotomous outcome. We note, however, our method is generalizable to custom, user-specified parametrizations.

Adam requires initial values for each curve coefficient. Randomly generated values may suffice for some use cases. However, we can provide more informative starting points. Prediction models would have been validated prior to implementation and before any subsequent ongoing assessment with dynamic calibration curves. We recommend leveraging information from such validation sets to initialize dynamic calibration curve coefficients. Initial coefficient estimates can be determined by fitting a curve with the preferred parameterization on the validation data using general linear modeling methods.

2.3. Adaptive windowing drift detection

2.3.1. Motivation

Concept drift detection is an established area of research providing methods to identify changes over time in the performance of prediction models [46]. However, the development and validation of these methods have focused on identifying changes in misclassification rates [46–49]. This focus on discrimination rather than calibration does not provide a sufficiently nuanced assessment of model performance for many clinical use cases [8,14,43,50,51].

Several drift detection methods may be extensible to detect calibration drift. We identified several key requirements for calibration drift detection in the context of model updating. First, the algorithm must be applicable to streaming data in which observations may arrive and be

processed individually rather than in batches that would require users to make assumptions regarding the speed of drift. Second, the algorithm should be flexible in terms of the calibration metric evaluated to support multiple modeling use cases. Third, upon detecting calibration drift, the method should inform a response to drift by providing insight into a window of recent observations that may be appropriate for updating the model.

Given these requirements, we implemented a novel variation of the Adwin drift detection algorithm. Adwin uses an adaptive sliding window method to detect drift by comparing error distributions among sliding pairs of subwindows [35]. Adwin adheres to all our requirements, being designed for streaming data, evaluating any bounded error metric, and inherently reporting a window of recent data that may be suitable for updating in response to any detected performance drift [35]. Further, Adwin provides intuitive parameterization, does not require users to prespecify expected performance during periods of stability, and provides statistical bounds on detection accuracy [35]. Extensions to the original Adwin algorithm also support parallel processing, minimize computational requirements, and account for delays between prediction generation and outcome observation [35,52]. Other methods, such as those used for statistical process control [47,53,54] could also be adapted to these requirements.

2.3.2. Approach

Adwin aims to maintain a window (W) of recent data which appears to be produced by a stable generating process. When a new observation arrives, the error in the observation's prediction is appended to the current window. Sliding divisions of W into pairs of temporal subwindows (i.e., W_1 containing newer data and W_0 containing older data; see Fig. 1C) allow for a sequence of comparison between a growing set of older data and a shrinking set of newer data. If a significant difference in the distribution of error between a pair of subwindows is discovered, Adwin shrinks W by dropping the older observations (W_0). Any time W shrinks, the process has identified drift and the retained data (W_1), in which no statistically significant shift in performance is observed, may be appropriate for updating the model to restore performance. Further details of the Adwin algorithm are provided in Appendix B.

We customized our Adwin implementation in two ways. In this work, we are interested in identifying increases in miscalibration that may require model updating, and are less concerned with potential improvements the calibration of predictions. Thus, we adjusted Adwin's test comparing pairs of subwindows to perform a one-sided test. Second, although Adwin was originally described using classification error [35], to focus our detector on calibration, we implemented Adwin with a stringent measure of calibration based on calibration curves maintained with the approach described in Section 2.2.2.

Our Adwin implementation monitors and evaluates a curve-based calibration error metric defined as the absolute difference between the predicted probability and the fitted value of the calibration curve (see Fig. B.1, details in Appendix B). The values evaluated by the Adwin algorithm are defined as follows for an observation at time t:

$$e_t = \left| \hat{p}_t - p_t \right|$$

where e_t is the calibration error, $p_t$the predicted probability generated from the prediction model, and $\hat{p}_t$ is the fitted value of the calibration curve estimated using the most recent coefficients of the dynamic calibration curve. This calibration error metric is bounded on the [0, 1] interval as required by the Adwin algorithm and is interpretable as the magnitude of deviation of the calibration curve from the ideal (similar to the integrated calibration index [37]). We selected a stringent curve-based metric to align our detector with the clinical decision-making context. However, we note any bounded calibration or accuracy metric could be monitored.

Adwin requires users to specify a significance level (δ) for detecting a difference in mean error between pairs of subwindows. Adwin adjusts for multiple comparisons to bound the error rate at the specified δ. As a default, we specify $\delta = 0.05$, which sets the theoretical upper bound of the false positive rate (i.e., detecting drift during periods of stable model performance) at the common Type I error threshold of 5%. Simulations providing some guidance on reasonable ranges of δ are available in Appendix B.

3. System evaluation

3.1. Simulation study

We conducted simulation studies to illustrate the performance properties of both dynamic calibration curves and our calibration drift detection approach. We simulated timeseries transitioning from calibrated predictions to predictions generated from one of four miscalibrated models. To reflect the notion that many clinical applications have predictions clustered in low risk regions, and that risk models operating with clustered high risk observations would present similar challenges, we generated true probabilities from a skewed $Beta(1.25, 5)$ distribution, which enriched for low probability predictions. Simulated outcomes were drawn from $Bernoulli(p)$, where p is the true probability. Miscalibrated probabilities were constructed by transforming the true probabilities with calibration curves that deviated from ideal calibration (see Fig. 2). These transformations created systematic overprediction, overfitting, miscalibration that fluctuated between over and underprediction over the range of probability, or miscalibration resulting from a subset of low risk observations being systematically overpredicted. The miscalibrated subset scenario was designed to reflect the possibility of changes in clinical guidelines or data capture methods that impact a model's ability to be accurately applied in particular subset of patients.

3.1.1. Dynamic calibration curves

To illustrate the evolution of dynamic calibration curves in response to performance changes, we documented the location of calibration curves after an abrupt change in model performance. Simulated timeseries included 1000 observations with calibrated predicted probabilities and a subsequent 5000 observations with miscalibrated predicted probabilities. For each of 1000 timeseries, we recorded values of the coefficients of the dynamic calibration curves after each observation.

In addition to visualizing the progression of curves over the timeseries, we calculated the proportion of the true, known calibration curve represented by the dynamic curve. After processing each observation, we estimated fitted values from the dynamic calibration curves on an evaluation set (n = 5000) from the current performance context. Across

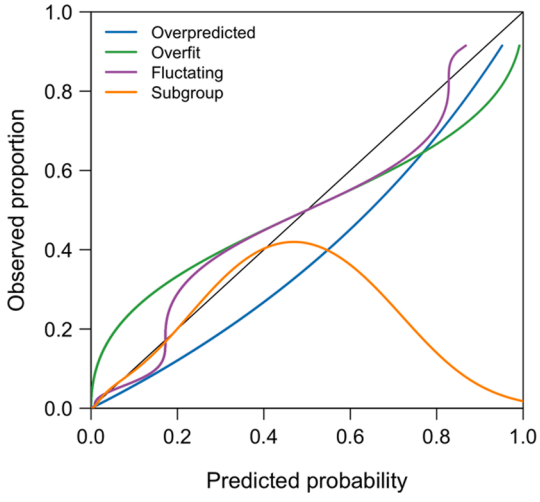

Fig. 2. Simulated forms of miscalibration.

repeated simulations for each scenario, we determined whether the 95% sampling intervals of these fitted values captured the true location of the calibration curve. We then calculated the proportion of the true calibration curve represented by the dynamic curve, weighted by the distribution of predicted probabilities.

During the pre-drift period, curves did not diverge from initial values. After drift onset, curves shifted in response to changes in calibration. For the shift to overprediction, curves illustrated the new post-drift form of calibration except at the highest range of probability (see Fig. 3; see Appendix C for corresponding plots for other forms of miscalibration) and captured 95% of the true calibration curve within approximately 600 observations after the abrupt change in performance. For the overfitted model, dynamic curves represented the calibration relationship in data-dense regions within 150 observation of the shift. Dynamic calibration curves were least responsive to the transition to miscalibration that fluctuated between over and underprediction across the range of probability. For this miscalibration scenario, the proportion of the true curve represented by dynamic curves remained above 80% after drift onset and slowly increased to 95% over the 5000 post-drift observations. This finding is likely due to this form of miscalibration not deviating far from the ideal calibration line in the high density, low probability regions.

3.1.2. Calibration drift detection

To evaluate the behavior of our Adwin-based calibration drift detection approach, we simulated timeseries with 5000 observations over which calibration deteriorated at varying speeds. The speed of calibration drift took four forms – an abrupt transition, a rapid transition over a short period (1000 observations), a gradual transition over an extended period (4000 observations), and a recurrent/seasonal transition in which observations transitioned back and forth between two calibration settings (every 1000 observations). With the exception of the recurrent case, the first 1000 observations in each series were generated from calibrated predictions and temporal transitions began immediately following this stable period. For each combination of temporal transition pattern and post-drift miscalibration, we applied our calibration drift detection system to 1000 timeseries with $\delta = 0.05$.

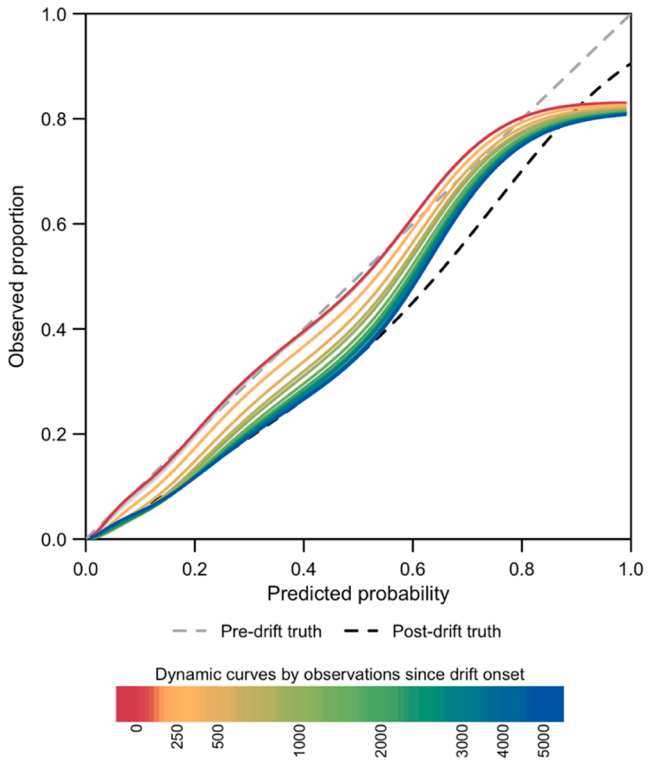

Fig. 3. Dynamic calibration curves for timeseries abruptly transitioning from a calibrated to an overpredicted context.

We evaluated our calibration drift detection approach for false positives, false negatives, and detection delays. False positives are detections occurring prior to drift onset. False negatives, missed opportunities to detect calibration drift, are any timeseries for which no drift was detected. Since varying transition speeds lead to differences in the speed of accumulated change in calibration, we evaluated detection delays as both time to detection (number of observations from drift onset to drift detection) and lag to detection (number of observations from expected detection to observed detection). To specify expected detection points, we compared the mean calibration error from a pre-drift calibrated population to populations with varying mixtures of pre and post-drift observations. The minimum mixture rate at which a statistically significant difference in calibration error was detected ($p < 0.05$) determined the point along each temporal transition at which detection would be expected (see Appendix C for additional detail). We further documented properties of the data window returned by the Adwin algorithm when drift was detected, including the size of the window and whether the window included observations occurring prior to drift onset. Sensitivity analyses using Adwin's standard two-sided test for any change in calibration are presented in Appendix D.

Each simulation provided a variety of details about the drift detection process. Fig. 4 illustrates these detailed results for transitions from a calibrated model toward one that systematically overpredicts. These plots illustrate the impact of speed of transition on the temporal error distribution evaluated by the Adwin algorithm. Fig. 4 further highlights the two critical aspects of the drift detection process: timing of detection and relevance of returned data windows. Box plots below the error distributions indicate the variable timing of detections across iterations, with the majority of detections occurring after drift onset and after expected detection points. In addition to the median window of observations returned at the time of detection, the figure also includes box plots presenting variability in the earliest observation included in returned windows. These results reveal the returned windows typically captured observations occurring after drift onset, with faster temporal transitions more likely to return some pre-drift observations. We discuss more details of each aspect of the simulation results for all drift scenarios below. Corresponding plots for other forms of miscalibration are also presented in Appendix C.

Both false positives and false negatives were rare in most scenarios (see Table 1). False positive rates were well below the 5% threshold our δ might suggest, a finding consistent with prior studies [35]. False negatives were rare and only observed under recurrent transitions in most cases. As an exception, drift toward miscalibration that fluctuated between over and underprediction across the range of probability was often missed, with false negatives rates between 35 and 50%. This form of miscalibration did not deviate far from the ideal in the more densely populated low risk range. As a result, the magnitude of change in calibration over time was small and more likely to be missed.

Delays between drift onset and detection varied by speed of transition and form of post-drift miscalibration (see Fig. 5). Time to detection increased as the speed of transition slowed from abrupt to rapid to gradual. Recurrent transitions experienced the most variability in detection timing. Time to detection was longest for recurrent transitions toward miscalibration that fluctuated between over and underprediction across the range of probability (median time to detection: 2575). Drifts toward this form of miscalibration consistently delayed detection, with a median of more than 1800 post-drift observations required for detections under all temporal transitions.

Lags to detection were generally more consistent than time to detection across temporal transition speeds (see Fig. 5). This is highlighted by transitions toward overfitting in which the median lags ranged from 101 to 173 observations for abrupt through gradual transitions, while corresponding median times to detection ranged from 173 to 640 observations. In contrast to this pattern, transitions toward

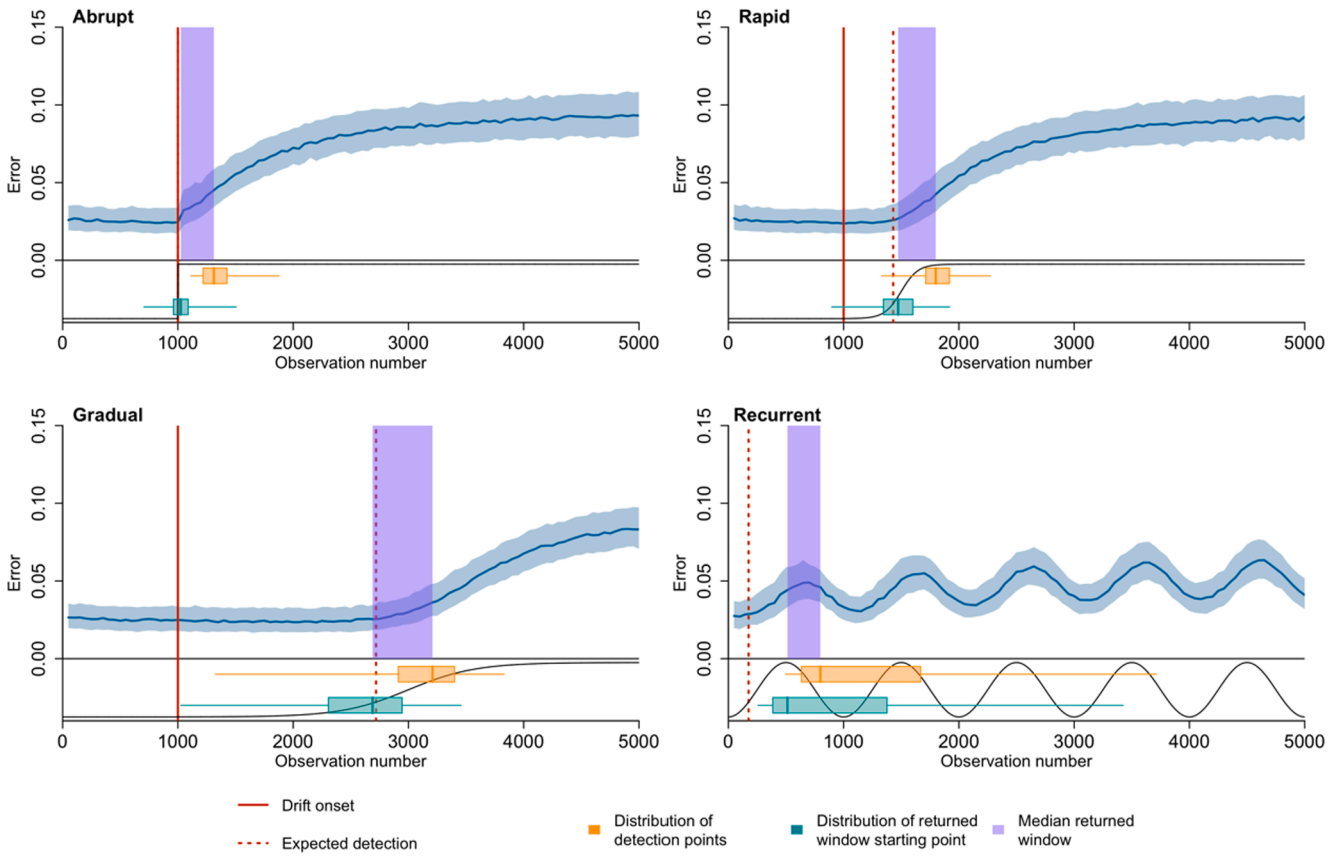

Fig. 4. Calibration error distribution (median and IQR) with detection characteristics for timeseries transitioning from calibrated to overpredicted at varying speeds. Distributions of detection points and the oldest observation included in returned windows, as well as a diagram of the speed of transition between calibration settings, are provided in box plots below the error distributions. The median window of recent data returned when drift was detected overlays the error distribution.

Table 1
False positive (FP) and false negative (FN) rates by temporal transition speed and post-drift calibration setting.

Post-drift calibration setting	Abrupt		Rapid		Gradual		Recurrent	
	% FP	% FN	% FP	% FN	% FP	% FN	% FP	% FN
Overpredicted	1.9	0	2.9	0	1.7	0	–	1.9
Overfit	2.1	0	1.8	0	1.7	0	–	0.9
Fluctuating	2.4	36.5	3.1	41	2.5	50.2	–	42.4
Subgroup	2.8	0	2.3	0	1.3	0	–	0.6

miscalibration that fluctuated between over and underprediction across the range of probability exhibited longer lags to detection for faster transitions. For all forms of miscalibration, gradual transitions were frequently detected prior to the identified expected point, which may reflect an accumulation of performance change over the extended period of transition.

Characteristics of the data windows returned by the calibration drift detector are reported in Table 2. Window size increased as the speed of transition slowed and as the corresponding time to detection increased. For rapid and gradual transitions, less than 6% of detections included pre-drift observations in the returned window. Abrupt transitions more frequently included pre-drift observations in the returned data window. However, in all simulations, windows that extended into the pre-drift period were primarily composed of post-drift observations. We note that there is no pre-drift period in the recurrent case, and thus no possibility of the returned data window containing pre-drift observations.

3.2. Case study

We conducted a case study to illustrate these methods on clinical data. In real-world data, we do not, and cannot, establish a ground truth for when calibration drift began. However, we can compare the timing of detections with observed performance over time. As a result, this case study is not intended to validate the methods in terms of accuracy, but rather serves to demonstrate the methods in the a setting where they may be deployed.

We applied our drift detection system to models for 30-day all-cause mortality among inpatient admissions to Department of Veterans Affairs (VA) facilities between 2006 and 2011 [11,12]. Detailed cohort eligibility criteria and predictor definitions have been previously described [11,12,55]. We trained models to predict 30-day all-cause mortality on 2006 admissions (n = 235,548) and evaluated calibration over the subsequent 5 years (n = 1,205,457). This study was approved by the Institutional Review Board and the Research and Development committee of the Tennessee Valley Healthcare System VA.

Variations in the response of different learning algorithms to changes in clinical environments impact the timing, extent, and form of calibration drift [11,12,32]. Such variability in calibration drift among learning algorithms may impact the timing of triggered updating using drift detection. To explore this in our case study, we applied our calibration drift detection system to three version of the mortality model—a logistic regression, a random forest, and a neural network. For each model, we ran the calibration drift detector over the full 5 years following model development, restarting the detector after each alert and without any intervening updates. We documented the timing and frequency of drift detections; the window returned with each detection; and the magnitude of difference in performance that triggered each

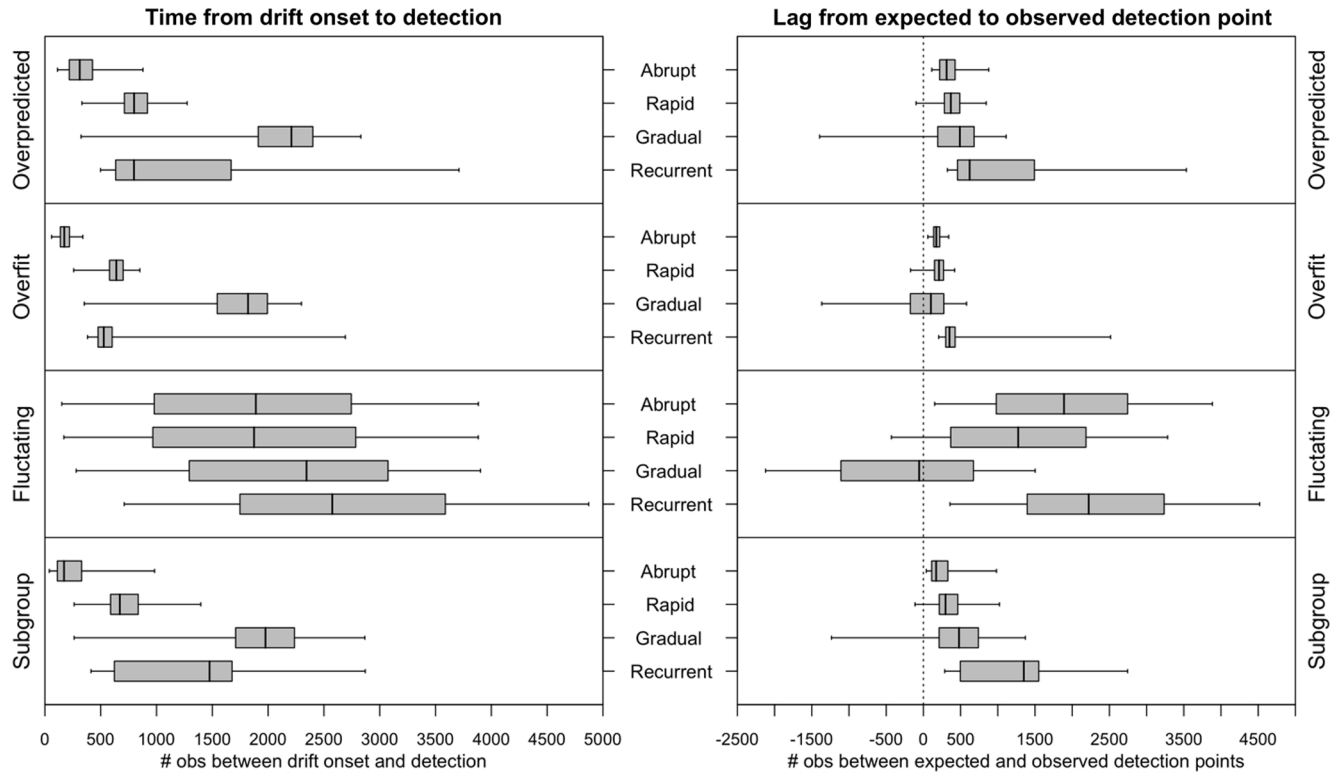

Fig. 5. Time to detection (left) and lag to detection (right) by speed and form of change.

Table 2
Properties of returned windows after drift detection.

Post-drift calibration setting	Transition pattern	Size (median & IQR)	% windows with any pre-drift obs*	% obs from pre-drift period
Overpredicted	Abrupt	266 [217, 356]	44.3	8.8
	Rapid	302 [246, 415]	4.5	1.7
	Gradual	420 [284, 585]	2.2	1.0
	Recurrent	261 [223, 331]	–	–
Overfit	Abrupt	137 [118, 176]	27.8	9.1
	Rapid	189 [140, 256]	4.5	1.8
	Gradual	274 [200, 430]	2.4	1.2
	Recurrent	168 [133, 213]	–	–
Fluctuating	Abrupt	280 [216, 422]	7.2	3.5
	Rapid	287 [226, 442]	5.4	2.2
	Gradual	317 [229, 481]	5.7	2.3
	Recurrent	302 [225, 462]	–	–
Subgroup	Abrupt	210 [146, 313]	69.2	17.9
	Rapid	282 [229, 397]	5.3	2.1
	Gradual	471 [330, 621]	3.3	1.7
	Recurrent	282 [228, 394]	–	–

* Pre-drift obs: observations before drift onset at $t = 1000$.

detection. These detection characteristics were compared to calibration over the study period.

Increases in miscalibration were detected at 6 points for the random forest model and 4 points for both the logistic regression and neural network models (see Fig. 6). Across all detection points, the mean increase in curve-based calibration error that triggered the detection was 0.0008. Detections aligned with observed increases in calibration error and were not noted during periods of stable calibration. The drift detector returned smaller windows during periods of more rapid change in calibration, which is likely related to reduced time to detection during such periods (as was observed in our simulation study).

4. Discussion

To support timely, data-driven identification of performance drift in clinical prediction models, we designed a calibration drift detection system to continuously monitor calibration and inform the model updating process. Our customizable and model-agnostic approach is designed for use with streaming data, making it well-suited for clinical environments continuously managing new patients and care encounters. Dynamic calibration curves support monitoring of detailed calibration metrics while incorporating information from new observations and evolving in response to changes in model performance. The drift detector alerts users to significant increases in miscalibration as information accrues and provides insight into a window of recent data that may be appropriate for model updating.

In evaluations, our system accurately detected drift in timeseries experiencing changes in calibration. Accurate detection is key to avoiding both missed opportunities to address model deterioration and alert fatigue from false alarms during periods of model stability. After drift onset, time to detection was associated with the speed and magnitude of calibration drift. This observation is to be expected, as slower transitions from a calibrated to a miscalibrated model require more observations before change can be distinguished from noise. However, lags to detection, a fairer comparison of any detection delay

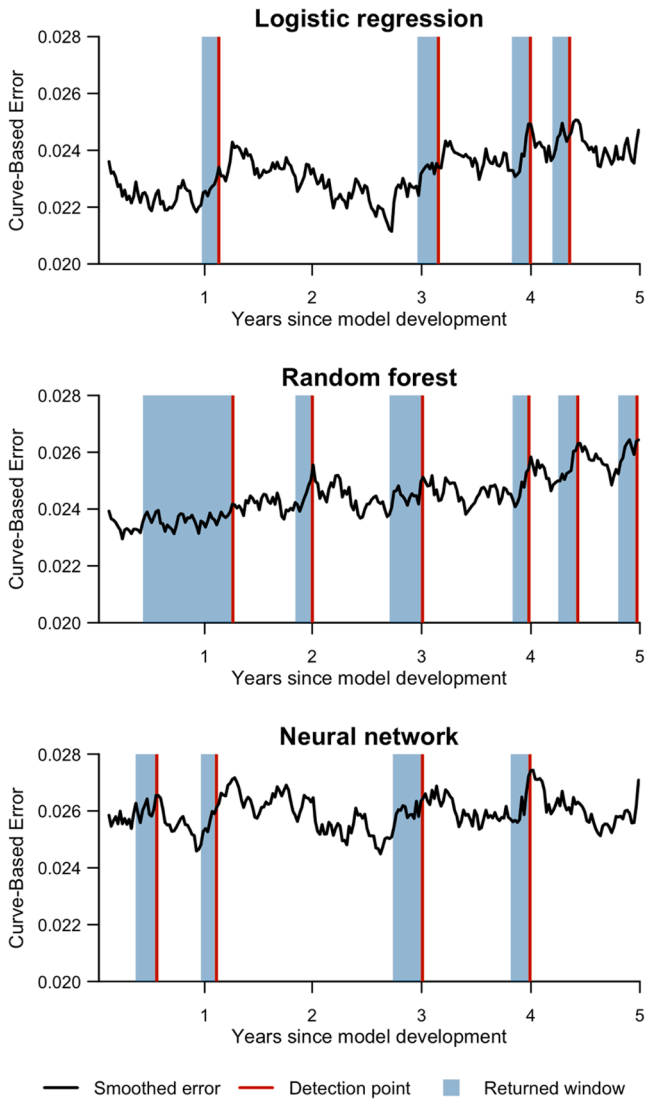

Fig. 6. Calibration over time and drift detections for models of 30-day mortality after hospital admission.

across differing speeds of drift, were generally consistent between abrupt, rapid, and gradual transitions.

The form of miscalibration after drift onset was also associated with the accuracy of performance monitoring and drift detection. Miscalibration in low density, high probability ranges was not easily learned by the dynamic calibration curves due to a lack of information in the impacted probability range. However, we note that drift was still detected in those simulated scenarios in which the curves did not evolve to fully represent the true calibration form. These findings highlight that to detect calibration drift and the potential need for model updating, dynamic calibration curves need to reflect that a deterioration in performance is occurring, but do not necessarily need to accurately reflect the specific form of that deterioration. Another challenging scenario was subtle miscalibration in high density probability ranges, as illustrated by our simulations involving miscalibration that fluctuated between over and underprediction across the range of probability. This scenario was most prone to missed detections and variable timing of detections due to the small magnitude of difference in calibration error before and after drift onset. Whether missed detections of small magnitude changes in performance are acceptable would depend on the accuracy requirements of specific clinical use case. Users may tune the system to increase the power to detect small changes as needed.

In response to a detection of calibration drift, ideally we would update models based on observations occurring after drift onset. In our evaluations, the majority of windows returned by the drift detector did not include pre-drift observations, and those windows that did extend into the pre-drift period were primarily composed of post-drift observations. We note that in both the simulation and case studies, calibration drift was often detected during transitional periods and returned windows represented a transitional state rather than data from a new, stably miscalibrated setting. While updating with such data may improve model performance, it may also require subsequent or even periodic updating as performance continues to evolve. This scenario is likely representative of how model updating would proceed in ever-evolving clinical environments where model performance may never experience extended periods of stability.

Users may tune several parameters to tailor the calibration drift detection system to the needs of specific use cases. The step size, or initial learning rate, of the Adam algorithm allows users to balance the variability of dynamic calibration curves with how quickly these curves evolve in response to changes in model performance. In addition, the drift detector includes two parameters that can be adjusted to promote the utility of alerts. First, in response to detections of very small changes in calibration that may not warrant model updating, users can require a more strongly significant change in performance by reducing the test's δ. Second, the minimum window size can be adjusted to avoid returning windows of insufficient sample size for updating based on model complexity. In future work, we will explore the impact of these parameters on both drift detection and resulting model updating to provide additional guidance.

We selected the Adwin algorithm as the basis for our drift detection method. However, we note that statistical process control (SPC) [53] may also be applicable. SPC methods have been implemented for a variety of healthcare applications—including tracking outcome rates [56,57], device safety [56,58,59], quality improvement [56,60–63], and model performance [32,64]. Prior work specifically evaluating calibration drift with SPC focused on forensic evaluations of model deterioration rather than using these methods to trigger model updating [32,64]. Among SPC methods, exponentially weighted moving averages (EWMA) [47], and particularly risk-adjusted EWMA [54], most closely aligned with the requirements listed in Section 2.2.1. However, compared to Adwin, parameterization for EWMA may be less intuitive and establishing appropriate parameter values may require pre-implementation simulation studies for each use case [47]. Future work will explore the relative performance of risk-adjusted EWMA or EMWA with our dynamic calibration curve method as alternatives to Adwin.

In designing this system, we focus the monitoring effort on calibration of the predictions generated by a model as a first-line indicator that the model may need attention. By emphasizing calibration signals, which are both susceptible to drift and linked to model utility, our approach supports efficient data-driven model updating when such updates may be most necessary. Alternatively, one could choose to monitor features of the dataset (e.g., case mix) and predictor-outcome associations. However, unless changes in these aspects of the data affect the accuracy of model predictions, such changes alone may not warrant model updating. On the other hand, if updating in response to detected calibration drift does not sufficiently correct performance, users would need to further investigate what may be driving the model's failure. In such cases, visualizations of temporal performance and summaries of key data distributions may provide model managers with insights into structural issues in the input data stream (e.g., changes in data capture/coding practices) that warrant examination. Further, maintaining open communication with clinical users can provide insight into critical clinical practice changes that may require substantive model adjustment be undertaken regardless of a drift detector's status.

Finally, we note that this work further highlights the need to better define clinically relevant performance metrics and methods for determining use case specific acceptable performance levels. Our case study showed some detections may result from very small increases in

miscalibration. Such detections may be too sensitive if they do not represent clinically significant calibration drift or identify performance changes that are too small to be corrected through model updating. With our current system, users may want to consider decreasing δ to require larger increases in miscalibration to trigger detections. However, as methods develop for defining clinically acceptable performance and/or clinically significant drift, our system could be extended to provide warnings when statistically significant drift is detected and to alert only when the drift in calibration exceeds a minimum magnitude of concern.

5. Conclusion

We developed and evaluated a calibration drift detection system to monitor detailed calibration metrics and provide data-driven guidance on when clinical prediction models may require updating. Our system, applicable to clinical prediction models built on the diverse and growing suite of learning algorithms, is designed to support alignment of model updating with the timing of performance drift. By updating models as performance deteriorates rather than on pre-determined schedules, model managers may minimize interim periods of insufficient model accuracy and focus analytic resources on those models most in need of attention. Our calibration drift detection system also provides insight into a candidate updating set by returning a window of recent observations occurring after the point at which a change in performance was identified. This system can be used to initiate predefined model updating strategies or in conjunction with data-driven methods to select updating methods.

Funding

This work was supported by funding from the Veterans Health Administration grant numbers VA HSR&D IIR 11-292 and 13-052; and the National Institutes of Health grant number BCHI-R01-130828. Funding agencies were not involved in the study design; the collection, analysis and interpretation of data; the writing of this manuscript; nor in the decision to submit the article for publication.

CRediT authorship contribution statement

Sharon E. Davis: Conceptualization, Methodology, Software, Investigation, Writing - original draft, Visualization. **Robert A. Greevy:** Methodology, Writing - review & editing. **Thomas A. Lasko:** Methodology, Writing - review & editing. **Colin G. Walsh:** Methodology, Writing - review & editing. **Michael E. Matheny:** Conceptualization, Methodology, Resources, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2020.103611.

References

[1] R. Amarasingham, R.E. Patzer, M. Huesch, N.Q. Nguyen, B. Xie, Implementing electronic health care predictive analytics: considerations and challenges, Health Aff. 33 (7) (2014) 1148–1154.
[2] M.E. Matheny, S. Thadaney Israni, M. Ahmed, D. Whicher, Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril, National Academy of Medicine, Washington, DC, 2019.
[3] M.E. Matheny, D. Whicher, Israni S. Thadaney, Artificial Intelligence in health care: a report from the national academy of medicine, JAMA (2019).
[4] K.G. Moons, A.P. Kengne, M. Woodward, et al., Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker, Heart 98 (9) (2012) 683–690.
[5] K.G. Moons, A.P. Kengne, D.E. Grobbee, et al., Risk prediction models: II. External validation, model updating, and impact assessment, Heart 98 (9) (2012) 691–698.
[6] D.G. Altman, Y. Vergouwe, P. Royston, K.G. Moons, Prognosis and prognostic research: validating a prognostic model, BMJ 338 (2009) b605.
[7] K.G. Moons, D.G. Altman, Y. Vergouwe, P. Royston, Prognosis and prognostic research: application and impact of prognostic models in clinical practice, BMJ 338 (2009) b606.
[8] N.D. Shah, E.W. Steyerberg, D.M. Kent, Big data and predictive analytics: recalibrating expectations, JAMA 320 (1) (2018) 27–28.
[9] G.A. Diamond, What price perfection? Calibration and discrimination of clinical prediction models, J. Clin. Epidemiol. 45 (1) (1992) 85–89.
[10] B. Van Calster, D.J. McLernon, M. van Smeden, et al., Calibration: the Achilles heel of predictive analytics, BMC Med. 17 (1) (2019) 230.
[11] S.E. Davis, T.A. Lasko, G. Chen, M.E. Matheny, Calibration drift among regression and machine learning models for hospital mortality. Proceedings of the AMIA Annual Symposium, 2017.
[12] S.E. Davis, T.A. Lasko, G. Chen, E.D. Siew, M.E. Matheny, Calibration drift in regression and machine learning models for acute kidney injury, J. Am. Med. Inform. Assoc. 24 (6) (2017) 1052–1061.
[13] G.L. Hickey, S.W. Grant, G.J. Murphy, et al., Dynamic trends in cardiac surgery: Why the logistic euroscore is no longer suitable for contemporary cardiac surgery and implications for future risk models, Eur. J. Cardiothorac. Surg. 43 (6) (2013) 1146–1152.
[14] L. Minne, S. Eslami, N. De Keizer, E. De Jonge, S.E. De Rooij, A. Abu-Hanna, Effect of changes over time in the performance of a customized SAPS-II model on the quality of care assessment, Intensive Care Med. 38 (1) (2012) 40–46.
[15] D.A. Cook, C.J. Joyce, R.J. Barnett, et al., Prospective independent validation of APACHE III models in an Australian tertiary adult intensive care unit, Anaesth. Intensive Care 30 (3) (2002) 308–315.
[16] E. Paul, M. Bailey, A. Van Lint, V. Pilcher, Performance of APACHE III over time in Australia and New Zealand: a retrospective cohort study, Anaesth. Intensive Care 40 (6) (2012) 980–994.
[17] M.M. Mikkelsen, S.P. Johnsen, P.H. Nielsen, C.J. Jakobsen, The EuroSCORE in western Denmark: a population-based study, J. Cardiothorac. Vasc. Anesth. 26 (2) (2012) 258–264.
[18] D.A. Jenkins, M. Sperrin, G.P. Martin, N. Peek, Dynamic models to predict health outcomes: current status and methodological challenges, Diagn. Progn. Res. 2 (23) (2018).
[19] S. Siregar, D. Nieboer, Y. Vergouwe, et al., Improved prediction by dynamic modelling: An exploratory study in the adult cardiac surgery database of the netherlands association for cardio-thoracic surgery, Circ. Cardiovasc. Qual. Outcomes 9 (2) (2016) 171–181.
[20] D.B. Toll, K.J. Janssen, Y. Vergouwe, K.G. Moons, Validation, updating and impact of clinical prediction rules: a review, J. Clin. Epidemiol. 61 (11) (2008) 1085–1094.
[21] T.H. Kappen, Y. Vergouwe, W.A. van Klei, L. van Wolfswinkel, C.J. Kalkman, K. G. Moons, Adaptation of clinical prediction models for application in local settings, Med. Decis. Making 32 (3) (2012) E1–E10.
[22] K.J. Janssen, K.G. Moons, C.J. Kalkman, D.E. Grobbee, Y. Vergouwe, Updating methods improved the performance of a clinical prediction model in new patients, J. Clin. Epidemiol. 61 (1) (2008) 76–86.
[23] T.P. Debray, Y. Vergouwe, H. Koffijberg, D. Nieboer, E.W. Steyerberg, K.G. Moons, A new framework to enhance the interpretation of external validation studies of clinical prediction models, J. Clin. Epidemiol. 68 (3) (2015) 279–289.
[24] P.M. Chen, T.M. Hemmen, Evolving healthcare delivery in neurology during the coronavirus disease 2019 (COVID-19) pandemic, Front Neurol. 11 (2020) 578.
[25] D.M. Mann, J. Chen, R. Chunara, P.A. Testa, O. Nov, COVID-19 transforms health care through telemedicine: evidence from the field, J. Am. Med. Inf. Assoc.: JAMIA (2020).
[26] U.N. Khot, A.P. Reimer, A. Brown, et al., Impact of COVID-19 pandemic on critical care transfers for ST-elevation myocardial infarction, stroke, and aortic emergencies, Circ. Cardiovasc. Qual. Outcomes (2020).
[27] V.M. Castro, R.H. Perlis, Electronic health record documentation of psychiatric assessments in Massachusetts emergency department and outpatient settings during the coronavirus disease 2019 (COVID-19) pandemic, JAMA Network Open 3 (6) (2020) e2011346.
[28] E.L. Hannan, K. Cozzens, S.B. King 3rd, G. Walford, N.R. Shah, The New York State cardiac registries: history, contributions, limitations, and lessons for future efforts to assess and publicly report healthcare outcomes, J. Am. Coll. Cardiol. 59 (25) (2012) 2309–2316.
[29] R. Jin, A.P. Furnary, S.C. Fine, E.H. Blackstone, G.L. Grunkemeier, Using Society of Thoracic Surgeons risk models for risk-adjusting cardiac surgery results, Ann. Thorac. Surg. 89 (3) (2010) 677–682.
[30] E.W. Steyerberg, G.J. Borsboom, H.C. van Houwelingen, M.J. Eijkemans, J. D. Habbema, Validation and updating of predictive logistic regression models: a study on sample size and shrinkage, Stat. Med. 23 (16) (2004) 2567–2586.
[31] Y. Vergouwe, D. Nieboer, R. Oostenbrink, et al., A closed testing procedure to select an appropriate method for updating prediction models, Stat. Med. 36 (28) (2017) 4529–4539.
[32] L. Minne, S. Eslami, N. de Keizer, E. de Jonge, S.E. de Rooij, A. Abu-Hanna, Statistical process control for monitoring standardized mortality ratios of a classification tree model, Methods Inf. Med. 51 (4) (2012) 353–358.

[33] S.E. Davis, R.A. Greevy, C. Fonnesbeck, T.A. Lasko, C.G. Walsh, M.E. Matheny, A nonparametric updating method to correct clinical prediction model drift, J. Am. Med. Inform. Assoc. 26 (12) (2019) 1448–1457.
[34] S.E. Davis, R.A. Greevy, T.A. Lasko, C.G. Walsh, M.E. Matheny, Comparison of prediction model performance updating protocols: using a data-driven testing procedure to guide updating, in: Proceedings of the AMIA Annual Symposium, 2019, pp. 1002–1010.
[35] A. Bifet, R. Gavalda, Learning from time-changing data with adaptive windowing. Paper Presented at: Proceedings of the 2007 SIAM International Conference on Data Mining, 2007.
[36] P.C. Austin, E.W. Steyerberg, Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers, Stat. Med. 33 (3) (2014) 517–535.
[37] P.C. Austin, E.W. Steyerberg, The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models, Stat. Med. 38 (21) (2019) 4051–4065.
[38] K. Van Hoorde, S. Van Huffel, D. Timmerman, T. Bourne, B. Van Calster, A spline-based tool to assess and visualize the calibration of multiclass risk predictions, J. Biomed. Inform. 54 (2015) 283–293.
[39] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization. Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), 2015.
[40] S. Ruder, An overview of gradient descent optimization algorithms, arXiv preprint arXiv:160904747, 2016.
[41] K. Miyaguchi, H. Kajino, Cogra: Concept-drift-aware stochastic gradient descent for time-series forecasting. Paper Presented at: Proceedings of the AAAI Conference on Artificial Intelligence, 2019.
[42] V. Losing, B. Hammer, H. Wersing, Incremental on-line learning: A review and comparison of state of the art algorithms, Neurocomputing 275 (2018) 1261–1274.
[43] B. Van Calster, D. Nieboer, Y. Vergouwe, B. De Cock, M.J. Pencina, E. W. Steyerberg, A calibration hierarchy for risk models was defined: from utopia to empirical data, J. Clin. Epidemiol. 74 (2016) 167–176.
[44] G. Nattino, S. Finazzi, G. Bertolini, A new test and graphical tool to assess the goodness of fit of logistic regression models, Stat. Med. 35 (5) (2016) 709–720.
[45] P. Royston, G. Ambler, W. Sauerbrei, The use of fractional polynomials to model continuous risk variables in epidemiology, Int. J. Epidemiol. 28 (5) (1999) 964–974.
[46] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, A. Bouchachia, A survey on concept drift adaptation, ACM Comput. Surv. (CSUR) 46 (4) (2014) 44.
[47] G.J. Ross, N.M. Adams, D.K. Tasoulis, D.J. Hand, Exponentially weighted moving average charts for detecting concept drift, Pattern Recogn. Lett. 33 (2) (2012) 191–198.
[48] K. Chen, Y.S. Koh, P. Riddle, Tracking drift severity in data streams. Australasian Joint Conference on Artificial Intelligence, 2015.
[49] M. Baena-Garcıa, J. del Campo-Ávila, R. Fidalgo, A. Bifet, R. Gavalda, R. Morales-Bueno, Early drift detection method. Fourth International Workshop on Knowledge Discovery From Data Streams, 2006.
[50] B. Van Calster, A.J. Vickers, Calibration of risk prediction models: impact on decision-analytic performance, Med. Decis. Making 35 (2) (2015) 162–169.
[51] X. Jiang, M. Osl, J. Kim, L. Ohno-Machado, Calibrating predictive model estimates to support personalized medicine, J. Am. Med. Inform. Assoc. 19 (2) (2012) 263–274.
[52] P.M. Grulich, R. Saitenmacher, J. Traub, S. Breß, T. Rabl, V. Markl, Scalable Detection of concept drifts on data streams with parallel adaptive windowing. Paper Presented at: 21st International Conference on Extending Database Technology (EDBT), 2018.
[53] J.C. Benneyan, R.C. Lloyd, P.E. Plsek, Statistical process control as a tool for research and healthcare improvement, BMJ Qual. Saf. 12 (6) (2003) 458–464.
[54] D.A. Cook, M. Coory, R.A. Webster, Exponentially weighted moving average charts to compare observed and expected values for monitoring risk-adjusted hospital indicators, BMJ Qual. Saf. 20 (6) (2011) 469–474.
[55] R.M. Cronin, J.P. VanHouten, E.D. Siew, et al., National veterans health administration inpatient risk stratification models for hospital-acquired acute kidney injury, J. Am. Med. Inform. Assoc. 22 (5) (2015) 1054–1071.
[56] J. Thor, J. Lundberg, J. Ask, et al., Application of statistical process control in healthcare improvement: systematic review, BMJ Qual. Saf. 16 (5) (2007) 387–399.
[57] J.C. Benneyan, A. Villapiano, N. Katz, M. Duffy, S.H. Budman, S.F. Butler, Illustration of a statistical process control approach to regional prescription opioid abuse surveillance, J. Addict. Med. 5 (2) (2011) 99–109.
[58] M.E. Matheny, L. Ohno-Machado, F.S. Resnic, Risk-adjusted sequential probability ratio test control chart methods for monitoring operator and institutional mortality rates in interventional cardiology, Am. Heart J. 155 (1) (2008) 114–120.
[59] M.E. Matheny, S.L. Normand, T.P. Gross, et al., Evaluation of an automated safety surveillance system using risk adjusted sequential probability ratio testing, BMC Med. Inf. Decis. Making 11 (2011) 75.
[60] A.P. Morton, M. Whitby, M.L. McLaws, et al., The application of statistical process control charts to the detection and monitoring of hospital-acquired infections, J. Qual. Clin. Pract. 21 (4) (2001) 112–117.
[61] A.W. Baker, S. Haridy, J. Salem, et al., Performance of statistical process control methods for regional surgical site infection surveillance: a 10-year multicentre pilot study, BMJ Qual. Saf. 27 (8) (2018) 600–610.
[62] A. Seim, B. Andersen, W.S. Sandberg, Statistical process control as a tool for monitoring nonoperative time, Anesthesiology 105 (2) (2006) 370–380.
[63] L. Pimentel, F. Barrueto Jr., Statistical process control: separating signal from noise in emergency department operations, J. Emerg. Med. 48 (5) (2015) 628–638.
[64] L. Minne, S. Eslami, N. de Keizer, E. de Jonge, S.E. de Rooij, A. Abu-Hanna, Statistical process control for validating a classification tree model for predicting mortality–a novel approach towards temporal validation, J. Biomed. Inform. 45 (1) (2012) 37–44.