Published in final edited form as:
Circ Cardiovasc Qual Outcomes. 2022 August ; 15(8): e008635. doi:10.1161/CIRCOUTCOMES.121.008635.

Maintaining a National Acute Kidney Injury Risk Prediction Models to Support Local Quality Benchmarking

Sharon E. Davis, PhD[a], **Jeremiah R. Brown, PhD**[b], **Chad Dorn, MS**[a], **Dax Westerman, MS**[a], **Richard J. Solomon, MD**[*,c], **Michael E. Matheny, MD, MS, MPH**[a,d,e,f]

[a]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN

[b]Departments of Epidemiology and Biomedical Data Science, Dartmouth Geisel School of Medicine, Hanover, NH

[c]Department of Medicine, Larner College of Medicine, University of Vermont, Burlington, VT

[d]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, USA

[e]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA

[f]Tennessee Valley Healthcare System VA Medical Center, Veterans Health Administration, Nashville, TN, USA

Abstract

Background.—The utility of quality dashboards to inform decision-making and improve clinical outcomes is tightly linked to the accuracy of the information they provide and, in turn, accuracy of underlying prediction models. Despite recognition of the need to update prediction models to maintain accuracy over time, there is limited guidance on updating strategies. We compare pre-defined and surveillance-based updating strategies applied to a model supporting quality evaluations among US veterans.

Methods.—We evaluated the performance of a VA-specific model for post-cardiac catheterization acute kidney injury (AKI) using routinely collected observational data over the six years following model development (n=90,295 procedures in 2013–2019). Predicted probabilities were generated from the original model, an annually retrained model, and a surveillance-based approach that monitored performance to inform the timing and method of updates. We evaluated how updating the national model impacted regional quality profiles. We compared observed to expected outcome ratios (O:E), where values above and below 1 indicated more and fewer adverse outcomes than expected, respectively.

Results.—The original model overpredicted risk at the national level (O:E=0.75 [0.74–0.77). Annual retraining updated the model five times; surveillance-based updating retrained once and recalibrated twice. While both strategies improved performance, the surveillance-based approach provided superior calibration (O:E=1.01 [0.99, 1.03] vs 0.94 [0.92–0.96]). Overprediction by

Corresponding Author: Sharon E. Davis, 2525 West End Ave, Suite 1475, Nashville, TN 37203, Phone: 615-936-5430, sharon.e.davis@vanderbilt.edu.

[*]second senior

Disclosures None.

the original model led to optimistic quality assessments, incorrectly indicating most of VA's 18 regions observed fewer AKI events than predicted. Both updating strategies revealed 16 regions performed as expected and two regions increasingly underperformed, having more AKI events than predicted.

Conclusions.—Miscalibrated clinical prediction models provide inaccurate pictures of performance across clinical units, and degrading calibration further complicates our understanding of quality. Updating strategies tailored to health system needs and capacity should be incorporated into model implementation plans to promote the utility and longevity of quality reporting tools.

Keywords

Acute kidney injury; predictive analytics; quality benchmarking; calibration; model updating

Introduction

Over two million diagnostic or interventional cardiac catheterization procedures are performed in the United States each year.[1, 2] Acute kidney injury (AKI) occurs in up to 14% of all patients following a procedure,[3] making AKI one of the most prevalent adverse events.[4–6] Among patients with multiple risk factors and/or chronic kidney disease, rates commonly double and have reached as high as 50% in some specific sub-populations.[3, 7, 8] When AKI occurs, patients also experience increased risks for future cardiovascular events, prolonged hospitalization, progression to end-stage renal disease, all-cause mortality, and increased acute care costs of over $7,500 per case.[9–11] Reducing the prevalence AKI has been identified as a key patient safety objective of the National Quality Forum,[12, 13] and it is estimated that over 70,000 cases of AKI could be prevented annually in the US through more consistent implementation of evidence-based preventive strategies.[3, 14]

Outcome surveillance programs can promote prevention efforts and quality improvement by integrating complex healthcare data and predictive analytics to provide concise summaries to decision makers for assessment and planning.[15] Quality improvement dashboards supported by clinical risk prediction models provide near-real-time data on adverse outcome incidence, and temporal trends, and risk-adjusted insight into care variation at an institutional or provider level.[14, 16, 17] Insights from such dashboards can prompt local investigation of quality successes and barriers, increase awareness of safety objectives, promote adoption of preventive interventions, and quantify the success of improvement initiatives.[18, 19]

Model calibration, the alignment of predicted and observed probabilities of adverse outcomes, is crucial to the validity of risk-adjusted outcome surveillance programs. Overprediction or underprediction of risk can lead to overly optimistic or pessimistic risk-adjusted quality measures, respectively.[20] Both scenarios mislead users, possibly preventing or delaying the benefits of such tools. Thus, the longterm utility of risk-adjusted outcome surveillance tools depends on sustained calibration of any underlying risk prediction models. This makes the tendency model calibration to degrade over time a key challenge[21–23] and model updating strategies critical components of risk-adjusted outcome surveillance.[20, 23]

We previously developed a national risk prediction model for post-procedural AKI within the US Department of Veterans Affairs (VA) health system.[8] This model is designed to support prospective AKI surveillance programs across VA sites. Acknowledging the importance of maintaining model calibration to sustain the utility of any such outcome surveillance tools, we sought to explore whether and how updating the national model impacted risk-adjusted quality evaluations nationally and regionally. In this study we compare the common practice of annually retraining clinical risk prediction models[22, 24, 25] and a data-driven surveillance-based updating strategy in which the timing and method of updating is driven by observed changes in calibration.[26] We implement both strategies on a retrospective national cohort of catheterization procedures and evaluated changes in apparent local AKI performance. This work both sets up prospective use of the VA's AKI prediction model for outcome surveillance and provides a structure for building and assessing outcome surveillance programs in other clinical contexts.

Methods

The data that support the findings of this study are available from the corresponding author upon reasonable request from qualified researchers trained in human subject protocols and approval of the appropriate institutional review board. Code supporting the statistical methods are available from the corresponding author upon request

Baseline prediction model

In this study, we explored the temporal performance characteristics of a previously published and validated model for AKI.[8] The original model was trained on a national cohort of diagnostic or interventional cardiac catheterization procedures performed at U.S. Department of Veterans Affairs (VA) facilities between January 01, 2009 to September 30, 2013. Procedures on patients under 18 years of age or with a history of dialysis were excluded. AKI was defined using KDIGO (Kidney Disease Improving Global Outcomes) criteria as ≥0.30 (mg/dL) or ≥50% increase in serum creatinine over baseline within 48-hours of the procedure or within 7-days for in- patients, or onset of dialysis within 7-days.[27] Predictors were constructed by collecting information on patient demographics, clinical presentation, and procedural urgency, as well as comorbidities, renal function, and medication use within 1 year prior to the catheterization procedure. Clinical information was assumed to be negative or not present when it is not found in the coded medical record data. L1-penalized logistic regression was used to construct a reduced model with the set of significant predictors. The following features remained in the reduced model: age at procedure, race, tobacco use, prior percutaneous coronary intervention, prior coronary artery bypass graft surgery, congestive heart failure, diabetes, prior myocardial infarction, peripheral vascular disease, prior stroke, shock, chronic kidney disease, estimated glomerular filtration rate, urgency of presentation, hypertension, unstable angina, ejection fraction, myocardial infarction in week prior to procedure, and anemia.

Temporal validation and updating cohort

We collected data on all diagnostic or interventional cardiac catheterization procedures at VA facilities between October 2013 and September 2019, the period immediately following that of the training cohort. To accurately apply, evaluate, and update the existing model, data on patient and procedure were collected using inclusion criteria and variable definitions established during initial model construction.[8] Procedures on patients under 18 years of age or with a history of dialysis were excluded. We extracted information on all retained predictors and AKI outcomes as noted above. This study was approved by the institutional review board of the Tennessee Valley Healthcare System VA Medical Center and informed consent was waived.

Model updating strategies

Over the 6-year study period, we applied two model updating strategies at the national level. Following common updating protocols, we annually retrained the model with the reduced set of predictors at the end of September using the preceding 12 months of data (i.e., procedure occurring in October of the prior year through September of current year). We also implemented a surveillance-based updating strategy that implemented a framework of data-driven methods for ongoing calibration monitoring, performance drift detection, updating cohort identification,[28] and updating approach selection[29] (see Figure 1).

The methods used in our surveillance-based updating strategy were developed in a setting in which data accrues as a stream of new observations.[28] However, the operational context in which this model would be used prospectively releases data on new procedures on a monthly basis rather than as an observation-level data stream. Given this setting, we customized the calibration monitoring and performance drift detection components of the surveillance framework. Rather than use dynamic calibration curves that update after each observation, we constructed a calibration curve with each monthly batch of data and used this curve to evaluate miscalibration for all procedures in the batch. Similarly, rather than testing for the presence of performance drift after each observation, we revised the drift detector to incorporate miscalibration information from all procedures in each monthly batch and only evaluate whether miscalibration is increasing at the end of each month.

As our study period started immediately after the timeframe of the training cohort used to develop the prediction model and following common practice, for the original model and the annual retraining strategy, we did not perform an initial update prior to model application. Given the potential for temporal patterns within the 5 years training period, however, we included an optional initial recalibration at the start of the surveillance updating strategy.[30] We applied the updating recommender testing procedure (Figure 1) to the last 6 months of training observations and implemented any recommended update prior to beginning surveillance-based updating.

For both updating strategies, updates were fit on a window of training data and then applied prospectively to the next set of observations, avoiding concern for overfitting in subsequent validations. Under the annual retraining approach, at each scheduled updating point, the prior 12 months of data comprised the training set for the updated model. This model was then applied to generate predictions for observations occurring in the 12 months after the scheduled updating point. Under the surveillance-based approach, the window of observations recommended by the drift detector only included observations occurring prior to the triggered updating point. These observations were used to determine and train the updates applied to subsequent observations occurring after the triggered updating point and until the next alert received from the drift detector.

Evaluations

To document the degree of intervention each strategy required, we recorded the timing and method used for each model update under each strategy. We also noted each instance in which the surveillance-based strategy was alerted by the drift detector, even if the updating method recommender component did not recommend any subsequent model adjustments. This may occur if none of the available updating methods are able to significantly improve upon the performance of the current model given the sample size of the available updating cohort.

We compared model performance at the national level under the original model and each updating strategy with discrimination and calibration metrics. We measured discrimination with the area under the receiver operating characteristics curve (AUC), for which a value near 0.5 indicates weak discrimination of cases from controls and a value near 1 indicates strong discrimination.[31] Mean calibration was measured with the observed to expected outcome ratio (O:E), which has an ideal value of 1.0 when mean predicted probability aligns with the observed event rate but increases or decreases as probabilities are systematically underpredicted or overpredicted, respectively.[32] More stringent calibration was measured by the alignment of observed outcome rates and predicted probabilities across the range of probability using the estimated calibration index (ECI; ideal value of 0) and calibration curves.[32, 33] We used bootstrapping (n=1000) to construct confidence intervals for each metric.

We also evaluated the impact of updating strategies on quality profiling applications using the national model to risk-adjust local understanding of relative performance. To effectively use a national prediction model to investigate local variation in performance, the national model would ideally maintain calibration on average, as measured by an O:E ratio near 1. With a well-calibrated national model, local O:Es significantly higher than 1 would indicate more adverse outcomes than expected (i.e., underperforming sites), while local O:Es significantly lower than 1 would indicate fewer adverse outcomes than expected (i.e., overperforming sites). With stable national performance, local centers with trending O:E ratios can validly interpret these trends as indicating improvement or deterioration of local performance over time. We documented performance over time of each Veterans Integrated Service Network (VISN, n=18). VISNs represent regional state grouping that for management and oversight of VA hospitals that generally include 4–8 cardiac catheterization laboratories each. We measured 12-month O:Es in each VISN using the predictions generated by the original model, the annually retrained model, and the surveillance-based updating model. Patterns in and understanding of relative quality over time under each updating strategy were compared.

Results

Characteristics of the study population are provided in Table 1. Between October 2013 and September 2019, 90,295 procedures to 89,244 unique patients met inclusion criteria. AKI occurred after 10.6% of these procedures, with annual AKI rates exhibiting no temporal pattern and ranging from 9.9% to 11.1%.

Updating recommendations

Figure 2 provides an overview of model updates across strategies, including the timing and methods of surveillance-based updating. The updating method recommender testing procedure supported an initial intercept correction based on the 6 months of data prior to the updating and validation period. Over the following six year study period, the batched surveillance-based updating strategy alerted to an increase in miscalibration 14 times. In response to two of these alerts, the surveillance system updated the model. In early 2015, the drift detector alerted to the presence of calibration drift. However, the surveillance system found none of the available updating methods significantly improved model accuracy and recommended no changes after 4 such alerts. Subsequently, in March 2015, the model was recalibrated with intercept correction using the prior 8 months of data. Following this update, the model was retrained in June 2016 using 13 months of data. During the period between Fall 2017 and Spring 2018, the drift detector repeatedly alerted to the possible presence of calibration drift. However, the testing procedure did not find updating to significantly improve performance, recommending not additional changes. Performance stabilized without any updates and the drift detector stopped alerting.

Overall performance

The original model systematically overpredicted the risk of AKI, with observed to expected outcomes ratio (O:E) of 0.75 (95% CI: 0.74–0.77). Updating improved upon the original model’s performance (see Table 2). The batched surveillance approach exhibited the best overall calibration ($p<0.05$ for pairwise comparisons to other approaches).

Temporal performance

The original model consistently overpredicted risk, with to O:Es below 1 in most months (see Figure 3). Updating with either the batched surveillance or annual retraining strategy returned the model to mean calibration (O:E near 1) in most months. The annually retrained model experienced a dip in O:E, indicating an increase in overprediction of risk, in early to mid 2016. During this time, the batched surveillance-based strategy experienced a short-term dip in performance, which resulted in an update resulting in the the approach generally maintaining calibration over this period with O:E confidence intervals including the ideal value of 1 in most months.

Calibration curves by updating strategy constructed for each 12-month period are presented in Figure 4 and provide a more detailed evaluation of calibration across the range of probability. In the first two years, the calibration curves for the surveillance-based updating strategy were more stable and closer to ideal calibration across a wide range of probability than those of the annual retraining strategy and the original model. In years four through six, the annual retraining and surveillance-based updating strategies resulted in similar calibration across the range of probability, and both exhibited better calibration than the original model.

Quality evaluations

None of the strategies achieved annual mean calibration for the entire study period at the national level (see Figure 5). The original model overpredicted nationally in all 6 years, making profiling of VISN or other sub-national level performance with this model difficult to interpret. With both the annually retraining and the batched surveillance-based approaches, the model was calibrated on average (O:E confidence interval captured 1) in 3 of 6. Confidence intervals were narrow given the large sample size and O:Es for both updating strategies were near 1 and fairly stable. Since models under both updating strategies achieved close to mean calibration at the national level, both could serve as a basis for quality profiling.

Updating the national model provided a stable benchmark for evaluating performance at the VISN-level. While updating maintained calibration of the national model (as shown in Figure 5), we do not expect all VISNs to achieve or attain stable calibration across the study period. Rather, we evaluated whether and how updating strategies impact VISN-level understanding of local performance trends. Figure 6 illustrates four patterns observed when profiling VISNs using models using the national model maintained under based each updating strategy. Most VISNs followed the pattern illustrated in Figure 6a. For these regions, the original model indicated the VISN reported better outcomes than expected (i.e., OE<1); however, both updating strategies indicated these regions performed as expected (i.e., OE=1). For two VISNs, the original model indicated initial expected performance (i.e., O:Es=1) with quality improvement such that the VISNs experienced better outcomes than expected after year 2 (i.e., O:Es<1) (see Figure 6b for example). After correcting calibration of the national model, these VISNs were actually initially underperforming and quality improved to expected levels after two years. The original model indicated one VISN performed as expected (see Figure 6c), with the exception of the first year. However, the updated models indicated increasingly worse outcomes than expected as O:Es increased over 1 in the final 3 years—reaching as high as 1.35 (95% CI: 1.20–1.50) in the final year based on the batched surveillance updating strategy. In another VISN (see Figure 6d), while the original model indicated the VISN generally performed as expected, updating revealed this VISN experienced more adverse outcomes than expected most years, with O:Es around 1.3 and reaching as high as 1.55 (95%CI: 1.36–1.76).

Discussion

This study provides evidence of the critical need for model maintenance protocols to be routinely incorporated into risk-adjusted outcome surveillance and quality benchmarking tools in order to provide interpretable and actionable information. Using 6 years of catheterization procedures at VA facilities nationwide, we evaluated the impact of model updating strategies on performance of a national AKI risk prediction model and subsequent outcome surveillance information at the regional (VISN) level. We found a national VA-specific model for post-catheterization AKI significantly overpredicted risk in the years immediately following the model's development. Annually retraining the model, a common updating strategy, corrected overprediction but required a new model be applied each year. A data-driven surveillance-based updating approach, however, achieved the best overall calibration and did so with less aggressive updates, retraining the model once and adjusting the intercept twice.

An outcome surveillance program using the original model would have provided VISNs with overly optimistic quality evaluations, in most cases incorrectly indicating VISNs experienced fewer AKI events than expected. However, both updating strategies corrected these evaluations and revealed most VISNs performed as expected. In two VISNs, using the original model may have given local leaders the impression quality was improving over the study period, when in fact performance was stable. Most concerning is that a quality evaluation based on the original model would have given two other VISNs the impression that they generally performed at acceptable levels, experiencing an expected number of AKI events. Updating revealed these VISNs to be underperforming and recording significantly more AKI events than their patient profiles predicted. In such cases, the original model could have prevented these VISNs from identifying the need to improve adherence to AKI preventive approaches or investigate quality barriers.

Our results highlight a need for updating strategies to be as part of initial model implementation within outcome surveillance and quality reporting tools. Early miscalibration may occur if the outcome rates changed over the course of the development period.[30] For example, if the AKI rate declined during development period, the model would be calibrated to the mean AKI rate throughout the entire development period rather than the lower AKI recent rate. This misalignment may lead to overprediction when the model is applied to subsequent observations, as was the case for our VA-specific AKI model which overpredicted risk even in the year following model development. An immediate intercept correction as recommended by the surveillance-based approach corrected this initial inaccuracy. Monitoring for miscalibration right from the start of implementation and considering temporal recalibration during model development will thus be important to initial utility of prediction-based tools, regardless of subsequent updating strategy.

This study also has important implications for outcome surveillance programs relying on clinical prediction models. Model updating can be integrated into the implementation of prediction-based tools. Our findings indicate both pre-defined and data-driven updating strategies led to accurate and consistent information on predicted AKI events. While the data-driven surveillance-based strategy provided more highly calibrated predictions across the range of probability, both strategies offered similar insights in quality assessments based on mean calibration (i.e., O:E ratios). In this study, we had access to a large dataset with over 15,000 catheterization procedures each year. More complex models or smaller populations would increase the risk of overfitting during the updating process.[34–36] In such cases, differences between updating strategies may be more pronounced. The annual retraining strategy would be susceptible to performance instability.[26, 34, 37] The surveillance-based approach, on the other hand, may produce more stable performance[26] by permitting the accumulation of more data prior to updating and aligning update complexity to sample size.[29]

The utility of outcome surveillance tools and quality dashboards is tightly linked to the accuracy of the information they provide and, in turn, the accuracy of underlying prediction models. Systematic overprediction by the original model provided overly optimistic assessments of the burden of AKI across VISNs. Updating revealed no VISNs performed better than expected during the study period (i.e., had O:Es<1) and, in fact, several VISNs underperformed. This finding is consistent with a temporal calibration study of Dutch intensive care units where the proportion of sites with higher than expected mortality increased from an optimistic 15% under the original model to 35% after correcting model calibration.[20] Model updating is thus critical to the utility of outcome surveillance tools and quality dashboards. Erroneously optimistic assessments, such as observed in this study, may delay interventions to improve patient outcomes by failing to highlight local needs for such programs. Pessimistic assessments, on the other hand, may dissuade adherence to successful quality improvement activities that incorrectly appear to be unsuccessful.

Our updating strategies were limited in the complexity of updating considered. In the study period, we had access to predictors selected by the previously developed regularized regression model rather than the full suite of predictors originally considered.[8] The most complex updating method considered here was thus retraining by re-estimating the coefficients of these selected predictors. As clinical environments and practice shift over time, the most relevant predictors may have changed. Similarly, were this model to be transported and applied in a non-VA population, we may expect substantial changes in predictor availability and feature distributions. In such cases, having the full suite of predictors and allowing both the annual and surveillance-based strategies to select new features by rebuilding the regularized regression model could have further improved model performance.

We note the surveillance-based updating framework is designed for prospective implementation and updating. In this retrospective analysis, we were able to build the entire cohort and harmonize any temporal changes in coding or structure of the input data. This prevents us from commenting on how unanticipated changes in data systems may affect the surveillance-based approach as it continuously monitors model performance. We will address this limitation in ongoing work implementing this updating strategy within a prospective study of an AKI outcome surveillance application.

Conclusion

Risk-adjustment with clinical prediction models provides critical insight into key quality and safety measures. Miscalibrated models, where predicted and observed risk are misaligned, provide an inaccurate picture of relative performance across clinical units, and degrading calibration over time further complicates local understanding of quality. Model updating with a surveillance-based strategy can be tailored to the unique needs of clinical systems and stabilize risk model performance by updating in response to observed changes in calibration. Implementing such an updating strategy within risk-adjusted quality reporting tools promotes the utility and longevity of these tools by sustaining the accuracy of information that can drive decision-making around resource allocation and local initiatives to improve care.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Sources of Funding

This study was funded by the National Institute of Diabetes and Digestive and Kidney Diseases grant number 5R01DK113201.

Non-standard Abbreviations and Acronyms

AKI	acute kidney injury
AUC	Area under the receiver operating characteristics curve
ECI	Estimated calibration index
KDIGO	Kidney Disease Improving Global Outcomes
O:E	Observed to expected outcome ratio
VA	US Department of Veterans Affairs
VISN	Veterans Integrated Service Network

References

1. Benjamin EJ, Blaha MJ, Chiuve SE, Cushman M, Das SR, Deo R, de Ferranti SD, Floyd J, Fornage M, Gillespie C, et al. Heart Disease and Stroke Statistics-2017 Update: A Report From the American Heart Association. Circulation. 2017;135:e146–e603. [PubMed: 28122885]
2. Go AS, Mozaffarian D, Roger VL, Benjamin EJ, Berry JD, Borden WB, Bravata DM, Dai S, Ford ES, Fox CS, et al. Heart disease and stroke statistics--2013 update: a report from the American Heart Association. Circulation. 2013;127:e6–e245. [PubMed: 23239837]
3. Brown JR, McCullough PA, Splaine ME, Davies L, Ross CS, Dauerman HL, Robb JF, Boss R, Goldberg DJ, Fedele FA, et al. How do centres begin the process to prevent contrast-induced acute kidney injury: a report from a new regional collaborative. BMJ Qual Saf. 2012;21:54–62.
4. Brown JR, Solomon RJ, Sarnak MJ, McCullough PA, Splaine ME, Davies L, Ross CS, Dauerman HL, Stender JL, Conley SM, et al. Reducing contrast-induced acute kidney injury using a regional multicenter quality improvement intervention. Circ Cardiovasc Qual Outcomes. 2014;7:693–700. [PubMed: 25074372]
5. Brown JR, Malenka DJ, DeVries JT, Robb JF, Jayne JE, Friedman BJ, Hettleman BD, Niles NW, Kaplan AV, Schoolwerth AC, et al. Transient and persistent renal dysfunction are predictors of survival after percutaneous coronary intervention: insights from the Dartmouth Dynamic Registry. Catheter Cardiovasc Interv. 2008;72:347–354. [PubMed: 18729173]
6. Brown JR and McCullough PA. Contrast Nephropathy and Kidney Injury. In: Thompson CA, ed. Textbook of Cardiovascular Intervention London: Springer London; 2014: 53–63.
7. Tsai TT, Patel UD, Chang TI, Kennedy KF, Masoudi FA, Matheny ME, Kosiborod M, Amin AP, Messenger JC, Rumsfeld JS, et al. Contemporary incidence, predictors, and outcomes of acute kidney injury in patients undergoing percutaneous coronary interventions: insights from the NCDR Cath-PCI registry. JACC Cardiovasc Interv. 2014;7:1–9. [PubMed: 24456715]
8. Brown JR, MacKenzie TA, Maddox TM, Fly J, Tsai TT, Plomondon ME, Nielson CD, Siew ED, Resnic FS, Baker CR, 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:e002136. [PubMed: 26656858]

9. James MT, Samuel SM, Manning MA, Tonelli M, Ghali WA, Faris P, Knudtson ML, Pannu N and Hemmelgarn BR. Contrast-induced acute kidney injury and risk of adverse clinical outcomes after coronary angiography: a systematic review and meta-analysis. Circ Cardiovasc Interv. 2013;6:37–43. [PubMed: 23322741]

10. Brown JR, Robb JF, Block CA, Schoolwerth AC, Kaplan AV, O'Connor GT, Solomon RJ and Malenka DJ. Does safe dosing of iodinated contrast prevent contrast-induced acute kidney injury? Circ Cardiovasc Interv. 2010;3:346–50. [PubMed: 20587788]

11. Jurado-Roman A, Hernandez-Hernandez F, Garcia-Tejada J, Granda-Nistal C, Molina J, Velazquez M, Albarran A and Tascon J. Role of hydration in contrast-induced nephropathy in patients who underwent primary percutaneous coronary intervention. Am J Cardiol. 2015;115:1174–8. [PubMed: 25759106]

12. Chertow GM, Burdick E, Honour M, Bonventre JV and Bates DW. Acute kidney injury, mortality, length of stay, and costs in hospitalized patients. J Am Soc Nephrol. 2005;16:3365–70. [PubMed: 16177006]

13. National Quality Forum (NQF). Safe Practices for Better Healthcare–2006 Update: A Consensus Report. Washington, DC. National Quality Forum2007.

14. Matheny ME, Ohno-Machado L and Resnic FS. Risk-adjusted sequential probability ratio test control chart methods for monitoring operator and institutional mortality rates in interventional cardiology. American Heart Journal. 2008;155:114–20. [PubMed: 18082501]

15. Matheny ME, Morrow DA, Ohno-Machado L, Cannon CP, Sabatine MS and Resnic FS. Validation of an automated safety surveillance system with prospective, randomized trial data. Med Decis Making. 2009;29:247–56. [PubMed: 19015285]

16. Matheny ME, Normand SL, Gross TP, Marinac-Dabic D, Loyo-Berrios N, Vidi VD, Donnelly S and Resnic FS. Evaluation of an automated safety surveillance system using risk adjusted sequential probability ratio testing. BMC Medical Informatics and Decision Making. 2011;11:75. [PubMed: 22168892]

17. Woodall WH, Fogel SL and Steiner SH. The monitoring and improvement of surgical-outcome quality. Journal of Quality Technology. 2015;47:383–399.

18. Shaw SJ, Jacobs B, Stockwell DC, Futterman C and Spaeder MC. Effect of a Real-Time Pediatric ICU Safety Bundle Dashboard on Quality Improvement Measures. Jt Comm J Qual Patient Saf. 2015;41:414–20. [PubMed: 26289236]

19. McLaughlin N, Afsar-Manesh N, Ragland V, Buxey F and Martin NA. Tracking and sustaining improvement initiatives: leveraging quality dashboards to lead change in a neurosurgical department. Neurosurgery. 2014;74:235–43; discussion 243–4. [PubMed: 24335812]

20. Minne L, Eslami S, De Keizer N, De Jonge E, De Rooij SE and 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;38:40–46. [PubMed: 22042520]

21. Davis SE, Lasko TA, Chen G, Siew ED and Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. Journal of the American Medical Informatics Association. 2017;24:1052–1061. [PubMed: 28379439]

22. Siregar S, Nieboer D, Vergouwe Y, Versteegh M, Noyez L, Vonk A, Steyerberg E and Takkenberg JJM. 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. 2016;9:171–181. [PubMed: 26933048]

23. Hickey GL, Grant SW, Murphy GJ, Bhabra M, Pagano D, McAllister K, Buchan I and Bridgewater B. 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;43:1146–1152. [PubMed: 23152436]

24. Hannan EL, Cozzens K, King SB 3rd, Walford G and 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;59:2309–16. [PubMed: 22698487]

25. Jin R, Furnary AP, Fine SC, Blackstone EH and Grunkemeier GL. Using Society of Thoracic Surgeons risk models for risk-adjusting cardiac surgery results. Annals of Thoracic Surgery. 2010;89:677–82. [PubMed: 20172107]
26. Davis SE, Greevy RA, Lasko TA, Walsh CG and Matheny ME. Comparison of prediction model performance updating protocols: using a data-driven testing procedure to guide updating. Proceedings of the AMIA Annual Symposium. 2019:1002–1010.
27. Lameire N, Kellum JA and Group KAGW. Contrast-induced acute kidney injury and renal support for acute kidney injury: a KDIGO summary (Part 2). Critical care. 2013;17:205. [PubMed: 23394215]
28. Davis SE, Greevy RA Jr., Lasko TA, Walsh CG and Matheny ME. Detection of calibration drift in clinical prediction models to inform model updating. J Biomed Inform. 2020;112:103611. [PubMed: 33157313]
29. Davis SE, Greevy RA, Fonnesbeck C, Lasko TA, Walsh CG and Matheny ME. A nonparametric updating method to correct clinical prediction model drift. Journal of the American Medical Informatics Association. 2019;26:1448–1457. [PubMed: 31397478]
30. Booth S, Riley RD, Ensor J, Lambert PC and Rutherford MJ. Temporal recalibration for improving prognostic model development and risk predictions in settings where survival is improving over time. Int J Epidemiol. 2020;49:1316–1325. [PubMed: 32243524]
31. Hanley JA and McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143:29–36. [PubMed: 7063747]
32. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ and Kattan MW. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21:128–38. [PubMed: 20010215]
33. Van Hoorde K, Van Huffel S, Timmerman D, Bourne T and 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. [PubMed: 25579635]
34. Vergouwe Y, Nieboer D, Oostenbrink R, Debray TP, Murray GD, Kattan MW, Koffijberg H, Moons KG and Steyerberg EW. A closed testing procedure to select an appropriate method for updating prediction models. Statistics in medicine. 2017;36:4529–4539. [PubMed: 27891652]
35. Toll DB, Janssen KJ, Vergouwe Y and Moons KG. Validation, updating and impact of clinical prediction rules: a review. Journal of clinical epidemiology. 2008;61:1085–94. [PubMed: 19208371]
36. Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG and Woodward M. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012;98:691–8. [PubMed: 22397946]
37. Van Calster B, Van Hoorde K, Vergouwe Y, Bobdiwala S, Condous G, Kirk E, Bourne T and Steyerberg EW. Validation and updating of risk models based on multinomial logistic regression. Diagnostic and Prognostic Research. 2017;1:1–14. [PubMed: 31093533]

What is Known

- The utility of clinical quality dashboards is tightly linked to the accuracy of the information they provide and, in turn, the accuracy of underlying prediction models.
- As the accuracy of clinical prediction models tends to degrade over time, model updating is necessary to sustain model performance and utility of associated quality metrics.

What the Study Adds

- Despite being trained on the same population using data from the immediately preceding years, a model for post-procedural acute kidney injury provided overly optimistic quality assessments, potentially delaying efforts to improve adherence to AKI preventive approaches or investigate quality barriers.
- Both a pre-defined and a surveillance-based updating strategies corrected the overly optimistic assessments of the original model, with the surveillance-based updating doing so with only limited model adjustments.
- Quality benchmarking applications can and should include model updating strategies in their implementation plans.

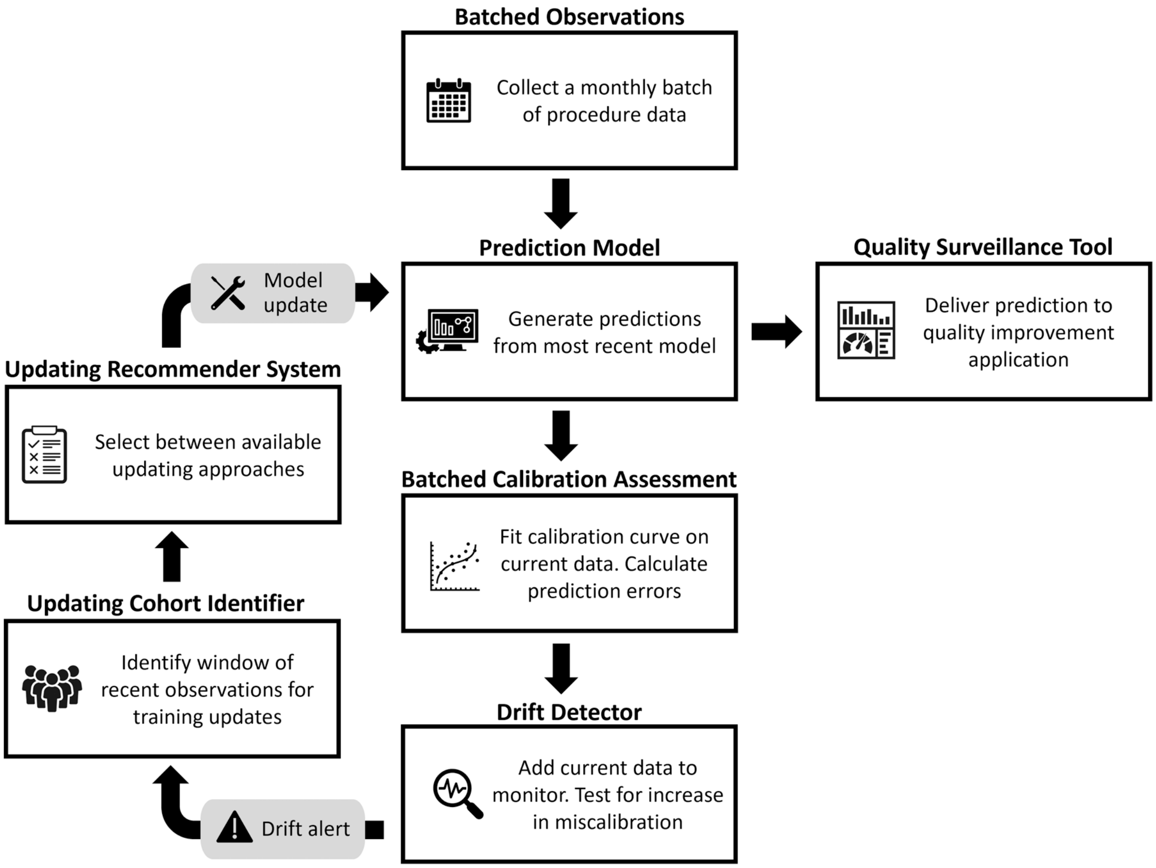
Batched Observations
Collect a monthly batch of procedure data
Prediction Model
Generate predictions from most recent model
Quality Surveillance Tool
Deliver prediction to quality improvement application
Model update
Updating Recommender System
Select between available updating approaches
Batched Calibration Assessment
Fit calibration curve on current data. Calculate prediction errors
Updating Cohort Identifier
Identify window of recent observations for training updates
Drift Detector
Add current data to monitor. Test for increase in miscalibration
Drift alert

Figure 1.
Cyclical surveillance-based updating framework.

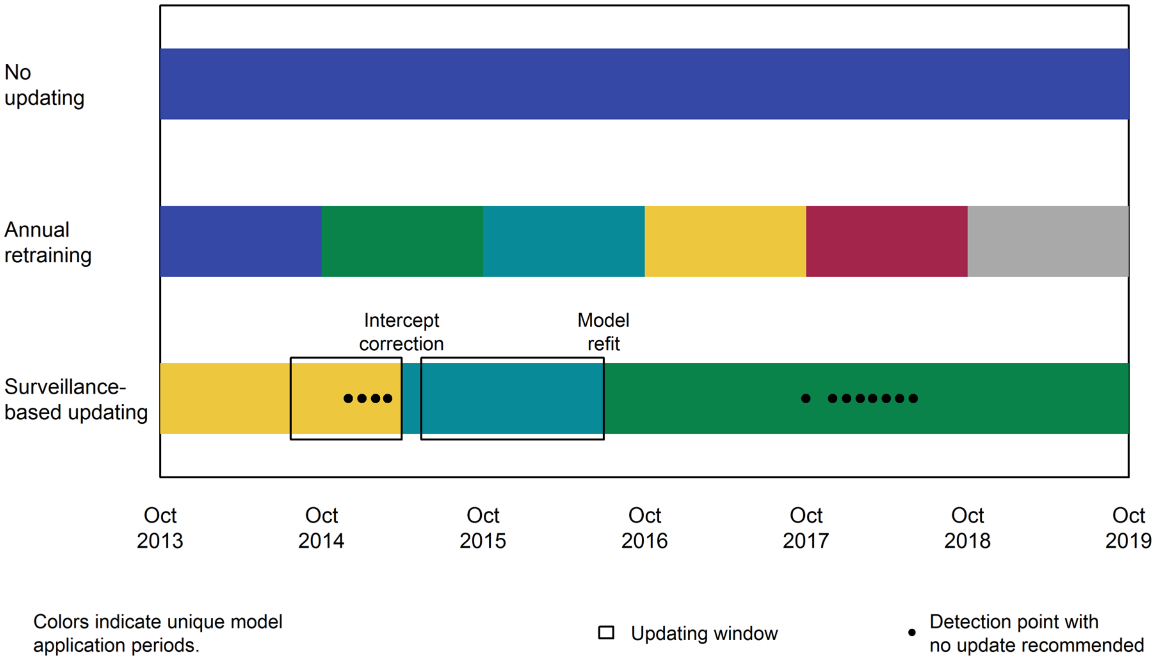

Figure 2.

Summary of updates by strategy. Colors highlight application periods for each version of the model within each updating strategy. For surveillance-based updating, the updating window (black boxes) and updating method are also indicated.

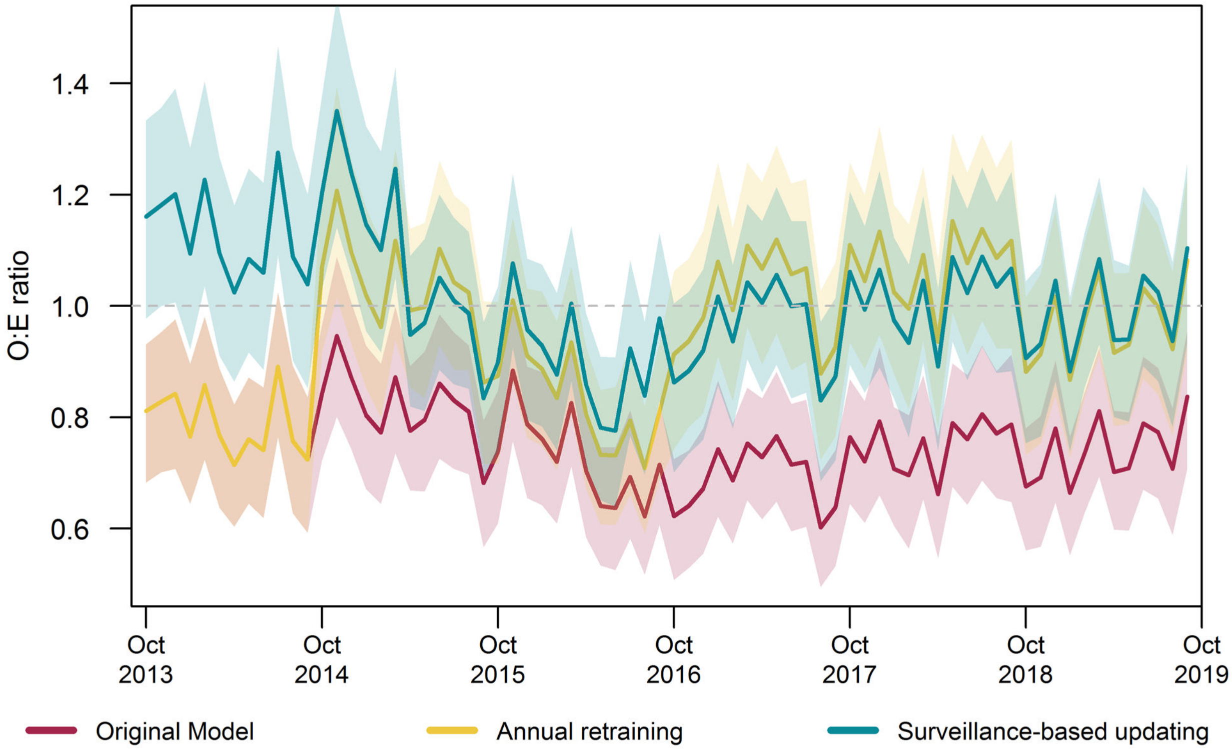

Figure 3.

Month by month model calibration by updating strategy. Mean calibration measured by the observed to expected outcome ratios (O:E) and 95% confidence intervals. Ideal value of 1, with higher values indicating underprediction and lower values indicating overprediction of risk.

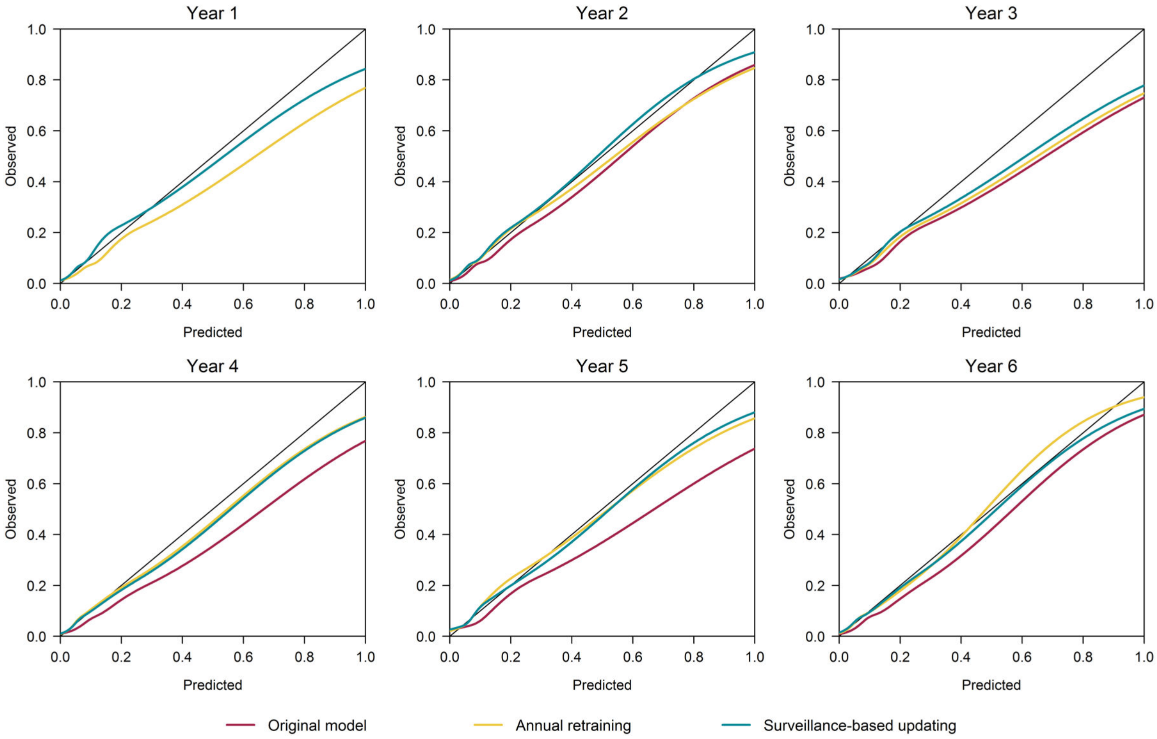

Figure 4.

Yearly calibration curves by updating strategy. For perfectly calibrated models in which predicted probabilities perfectly aligned with observed outcome rates, these curves would follow with 45° line shown in black on each plot.

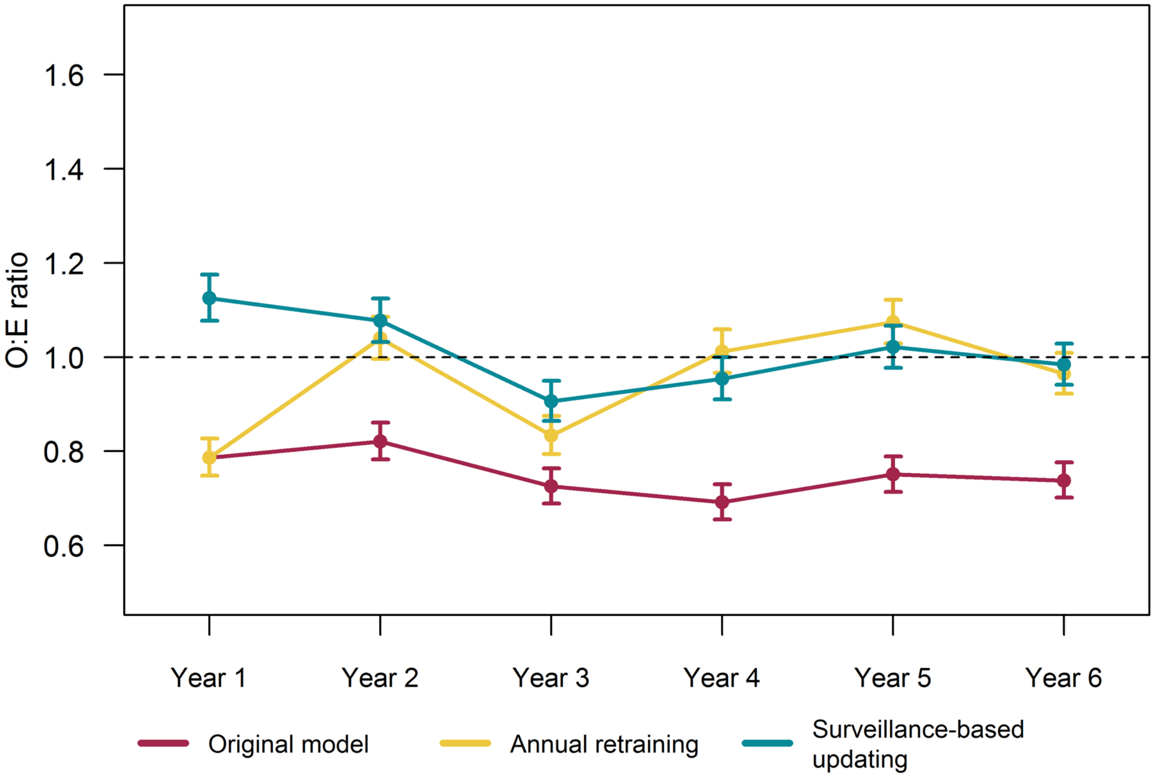

Figure 5.

National 12-month observed to expected outcome ratios (O:E) and 95% confidence intervals by updating strategy.

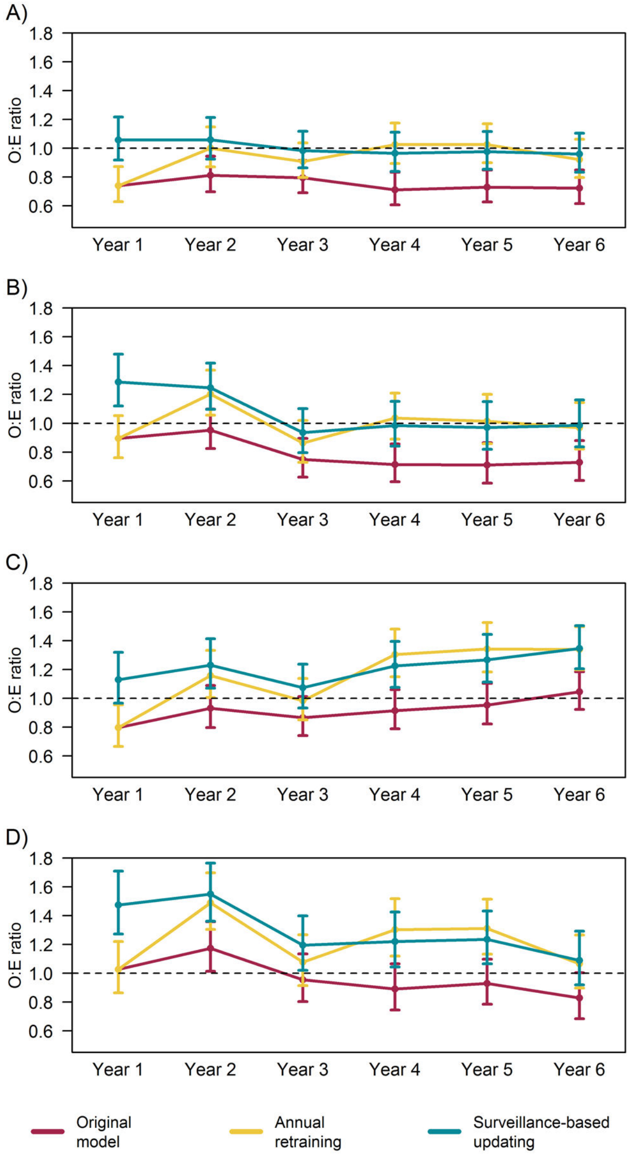

Figure 6.

12-month observed to expected outcome ratios (O:E) and 95% confidence intervals indicating relative performance in four example Veterans Integrated Service Network (VISN) regions using a national model for expected outcomes by national model's updating strategy. Each panel displays results for a VISN illustrating observed patterns: A) Apparent performance was better than expected under original model, but was as expected after updating calibration; B) Apparent performance improved to better than expected levels under original model, but performance improved from underperforming to expected levels after updating calibration; C) Apparent performance was as expected under the original model, but updating indicated increasingly worse outcomes than expected; and D) Apparent performance was as expected under the original model, but updating indicated underperformance that improved over time.

Table 1.

Study population characteristics.

	Overall	**AKI**	**Non-AKI**
N	90,295	9,597	80,698
Age (median, IQR)	68 [62, 73]	69 [64, 75]	68 [62, 73]
Non-white race	23.3%	26.1%	22.9%
Selected Patient Characteristics			
Tobacco use	41.0%	39.3%	41.2%
Chronic kidney disease	19.4%	36.0%	17.4%
Prior PCI	28.3%	27.4%	28.4%
Diabetes	36.0%	46.8%	34.7%
Procedure Urgency			
Routine	60.2%	47.9%	61.6%
Urgent	35.8%	44.1%	34.8%
Emergent	3.9%	7.7%	3.5%
Salvage	0.1%	0.3%	0.1%

AKI: acute kidney injury; IQR: interquartile range; PCI: percutaneous coronary intervention

Table 2.

Model performance (and 95% confidence interval) by updating strategy across the entire study period. Bold values indicate the best performance for a given metric.

Updating strategy	AUC	OE	ECI
Original model	0.684 [0.678, 0.69]	0.751 [0.737, 0.766]	0.152 [0.129, 0.169]
Annually retrained model	**0.690** [0.685, 0.697]	0.940 [0.923, 0.959]	0.014 [0.005, 0.020]
Surveillance-based updating	**0.690** [0.684, 0.695]	**1.007** [0.988, 1.028]	**0.005** [0, 0.008]

AUC: area under the receiver operator characteristic curve (ideal values of 1); OE: observed to expected outcome ratio (ideal value of 1); ECI: estimated calibration index (ideal value of 0).