Monitoring Approaches for a Pediatric Chronic Kidney Disease Machine Learning Model

Keith E. Morse[1] Conner Brown[2] Scott Fleming[3] Irene Todd[2] Austin Powell[2] Alton Russell[4] David Scheinker[2] Scott M. Sutherland[5] Jonathan Lu[3] Brendan Watkins[2] Nigam H. Shah[3] Natalie M. Pageler[6,7] Jonathan P. Palma[8]

[1] Division of Pediatric Hospital Medicine, Department of Pediatrics, Stanford University School of Medicine, Stanford, California, United States
[2] Information Services Department, Lucile Packard Children's Hospital, Stanford, Palo Alto, California, United States
[3] Department of Biomedical Data Science, Stanford University, Palo Alto, California, United States
[4] Harvard Medical School, Boston, Massachusetts, United States
[5] Division of Nephrology, Department of Pediatrics, Stanford University, Stanford, California, United States
[6] Division of Pediatric Critical Care Medicine, Department of Pediatrics, Stanford University School of Medicine, Stanford, California, United States
[7] Division of Systems Medicine, Department of Pediatrics, Stanford University School of Medicine, Stanford, California, United States
[8] Division of Neonatology, Department of Pediatrics, Orlando Health, Orlando, Florida, United States

Address for correspondence Keith E. Morse, MD, MBA, 780 Welch Road, Palo Alto, CA 94304, United States (e-mail: KMorse@stanfordchildrens.org).

Appl Clin Inform 2022;13:431–438.

Abstract

Objective The purpose of this study is to evaluate the ability of three metrics to monitor for a reduction in performance of a chronic kidney disease (CKD) model deployed at a pediatric hospital.

Methods The CKD risk model estimates a patient's risk of developing CKD 3 to 12 months following an inpatient admission. The model was developed on a retrospective dataset of 4,879 admissions from 2014 to 2018, then run silently on 1,270 admissions from April to October, 2019. Three metrics were used to monitor its performance during the silent phase: (1) standardized mean differences (SMDs); (2) performance of a "membership model"; and (3) response distribution analysis. Observed patient outcomes for the 1,270 admissions were used to calculate prospective model performance and the ability of the three metrics to detect performance changes.

Results The deployed model had an area under the receiver-operator curve (AUROC) of 0.63 in the prospective evaluation, which was a significant decrease from an AUROC of 0.76 on retrospective data ($p = 0.033$). Among the three metrics, SMDs were significantly different for 66/75 (88%) of the model's input variables ($p < 0.05$) between retrospective and deployment data. The membership model was able to discriminate between the two settings (AUROC = 0.71, $p < 0.0001$) and the response distributions were significantly different ($p < 0.0001$) for the two settings.

Conclusion This study suggests that the three metrics examined could provide early indication of performance deterioration in deployed models' performance.

Keywords
- learning algorithm
- electronic health record
- monitoring
- machine learning
- safety

received
September 13, 2021
accepted after revision
March 1, 2022

Georg Thieme Verlag KG,
Rüdigerstraße 14,
70469 Stuttgart, Germany

DOI https://doi.org/10.1055/s-0042-1746168.
ISSN 1869-0327.

Background and Significance

The potential for machine learning to meaningfully improve health care is well recognized,[1] yet few models have been deployed and validated.[2] There is growing recognition of the need to validate deployed models to measure their performance and, more importantly, ensure patient safety.[3–5]

Key to validation is model transportability, or the ability to make accurate predictions on patients who are different from, but plausibly related to, those on whom the model was trained.[6] Specifically, models deployed at the same site with their training data originated are evaluated for temporal transportability, or the ability to make accurate predictions on patients from the same population during a different time period.[7] Previous work in temporal transportability has been done in acute kidney injury (AKI),[8] sepsis,[9] and intensive care unit mortality.[10]

Transportability is often measured by comparing baseline model performance to observed performance, as defined by an observed ground-truth label.[9,11] However, obtaining ground-truth labels is often a slow process in health care because the pace of care anywhere outside the inpatient setting involves a lag time of weeks, months, or years. Models with prediction tasks based on such delayed end points, therefore, face constrained validation options when initially deployed or redeployed after retraining. Models can either be deployed into patient care without validation, introducing a known safety risk,[3] or be run "silently" until the lag period has elapsed, thus delaying the potential clinical benefit of the model.

If models are to become reliable parts of day-to-day health care operations, safety monitoring needs to be closer to real time. There is a need for metrics that are available prior to the observation of ground-truth labels, here termed "label-blind" metrics. As a safety tool, label-blind metrics could serve as an early warning of performance deterioration due to poor model transportability, potentially triggering in-depth examination or suspension of use in clinical care.[12] Note that methods to restore model performance after deterioration is identified are beyond the scope this study.

Several label-blind metrics have been proposed.[13,14] Debray et al outlines the use of standardized mean differences (SMDs), performance of a "membership model," and response distribution analysis for identifying potential transportability failures. These metrics have been demonstrated for a model that predicts deep vein thrombosis using a retrospective dataset.[15] The utility of these metrics in a live clinical environment has not been studied.

This study's deployed model identifies pediatric patients at high risk of developing chronic kidney disease (CKD) following an episode of AKI sustained during a hospital admission. AKI is a common occurrence during admissions, affecting over 25% of critically ill pediatric patients.[16] There is a growing body of evidence linking the renal insult from AKI to subsequent development of CKD, which itself can lead to growth impairments, anemia, and in extreme cases, the need for kidney transplant or life-long dialysis.[17–19] With proper monitoring and treatment under the direction of a pediatric nephrologist, onset and progression of CKD can be mitigated.[20,21] Unfortunately, the current capacity of pediatric nephrologists in the United States is strained and cannot evaluate all pediatric patients who survive an episode of AKI.[22] The clinical problem the algorithm addresses is to identify which pediatric patients are at highest risk of developing CKD and would benefit most from a referral for nephrology follow-up. Similar programs that use rule-based referral methods have been implemented elsewhere.[18]

Objective

The purpose of this study is to evaluate the ability of three metrics to monitor for a reduction in performance of a CKD model deployed at a pediatric hospital.

Methods

Lucile Packard Children's Hospital (LPCH) is a 360-bed academic, freestanding, tertiary care children's hospital affiliated with Stanford University. The CKD risk model was developed by a data science research team at Stanford University using retrospective clinical data extracted from the LPCH electronic health record (EHR; Epic Systems, Verona, Wisconsin, United States). The model outputs the risk of developing CKD in the 3 to 12 months following an episode of AKI sustained during inpatient hospital admission. AKI was defined based on Kidney Disease Improving Global Outcomes (KDIGO)[23] criteria: Stage 1—1.5 to 1.9 times increase from baseline; Stage 2—2.0 to 2.9 times increase; Stage 3 —$\geq$ 3.0 times increase. The lowest recorded serum creatinine prior to or during the index admission was used to define baseline creatinine. Note that due to inconsistent data availability, AKI diagnosis based on decreased urine output was not included. CKD is defined as an estimated glomerular filtration rate of less than 60 mL/min/1.73 m^2 (as calculated by the Schwartz equation[24]) observed in at least two creatinine measurements during the follow-up window of 3 to 12 months after hospital discharge.

The model was trained on a historical dataset of 4,879 LPCH admissions from May 3, 2014 to August 31, 2018, here termed as the "retrospective dataset." Inclusion criteria for the retrospective dataset was: patients aged $\geq$3 months, <18 years, without an existing diagnosis of CKD; admission length of stay $\geq$2 nights; $\geq$1 creatinine measurements during the admission and $\geq$2 creatinine measurements during follow-up window. The final version of the model is a logistic regression model that uses 11 data inputs related to patient characteristics and laboratory results obtained during the admission, which are used to create a total of 75 model input features (see **Supplementary Material** for complete input feature list, available in the online version). In training, the model demonstrated an area under the receiver-operator curve (AUROC) of 0.76 on a held-out evaluation set.

The model was operationalized via a machine learning platform within the EHR that allows custom machine learning models to take input data directly from the EHR in near

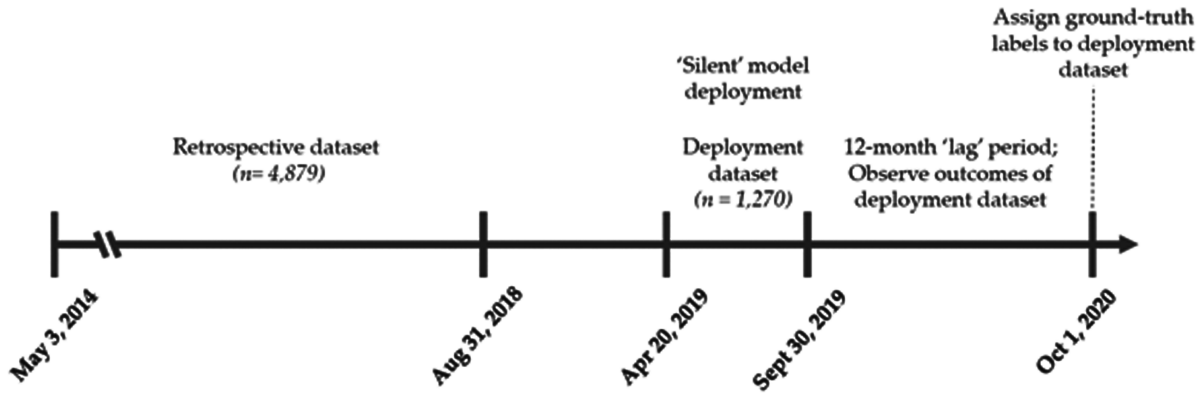

Fig. 1 Study timeline. Indicates date ranges for hospitalization discharge dates in retrospective and deployment datasets, as well as 12-month "lag" period.

real-time. This real-time model was run in a "silent" mode (i.e., the output was not shown to any end-user) on 1,270 admissions from April 20, 2019 to September 30, 2019. The data from these 1,270 admissions are here termed the "deployment dataset." Inclusion criteria for the deployment dataset was the same as for the retrospective dataset.

Following the completion of a 12-month lag period (completed October 1, 2020), available inpatient and outpatient creatinine values were collected and one of three ground-truth labels were assigned to each observation: CKD present, CKD absent, or insufficient data available. The same CKD definition was applied to the deployment dataset as to the retrospective dataset. The performance of the deployed model was summarized by calculating an AUROC value, excluding observations with insufficient follow-up data. Significance testing was done via bootstrapping means and calculating Mann-Whitney U statistics. See ►Fig. 1 for a study timeline.

The three label-blind metrics are the SMDs[25,26] and the performance of a "membership model"[15] on the input features, as well as plotting the response distributions for the model outputs.[14,15] These three metrics attempt to quantify the consistency (i.e., detect differences in "case-mix"[8] or the presence of dataset shift[27]) between the retrospective and deployment input datasets.

SMDs reflect differences in variable means between two groups in units of each variable's pooled standard deviation. This approach allows one to identify which input variables differ the least and which differ the most between the two datasets, after accounting for variable- and dataset-specific variances. Individual input variable significance testing was done with Mann-Whitney U tests for continuous variables and two-sided Z tests for proportional variables. Aggregate significance testing was done using the Bonferroni correction.[28]

Membership models are models trained to predict dataset source (in this case "retrospective" vs. "deployment") for each observation using just the observation's input data. Intuitively, the ability to learn a membership model that distinguishes between observations in the two datasets with high accuracy suggests that there are substantial (potentially complex and multivariate) differences between the two datasets. Conversely, the inability to learn such a model, even when given substantial model flexibility, can provide evidence against the hypothesis that the two datasets differ somehow in their input data distributions. Significance testing was done using bootstrapped confidence intervals against the null hypothesis of AUROC = 0.5.

Response distribution analysis is a method to evaluate model transportability by directly analyzing the output of risk prediction models. A simple histogram of the model output is constructed for retrospective and deployment datasets, and a heuristics-driven evaluation of the resulting distributions is performed. Difference in the two distributions can indicate feature drift that impacts model predictions.[14] The Kolmogorov–Smirnov test was used to assess whether response distributions for the retrospective and deployment datasets differed significantly.

See ►Fig. 2 for methods summary. This study was approved by the Stanford Institutional Review Board.

Results

Patient characteristics and laboratory results for the retrospective and deployment datasets are summarized in ►Table 1. Frequency of AKI was significantly higher in the retrospective dataset, occurring almost twice as often (37.7 vs. 19.1%).

SMD analysis showed 66 of 75 input features (88%) to be significantly different between the retrospective and deployment datasets ($p < 0.05$). For interpretability, features were binned into five categories based on clinical significance ("Patient or Admission Characteristics," "Baseline Kidney Function," "Change in Kidney Function," "AKI—Occurrence," and "AKI—Stage"). See ►Table 2 for summary results of the five categories and **Supplementary Material** (available in the online version) for results of each input feature.

A logistic regression membership model showed significant discriminatory capacity between retrospective and deployment datasets. Training on all retrospective and deployment data (class ratio 79%) and using fivefold cross-

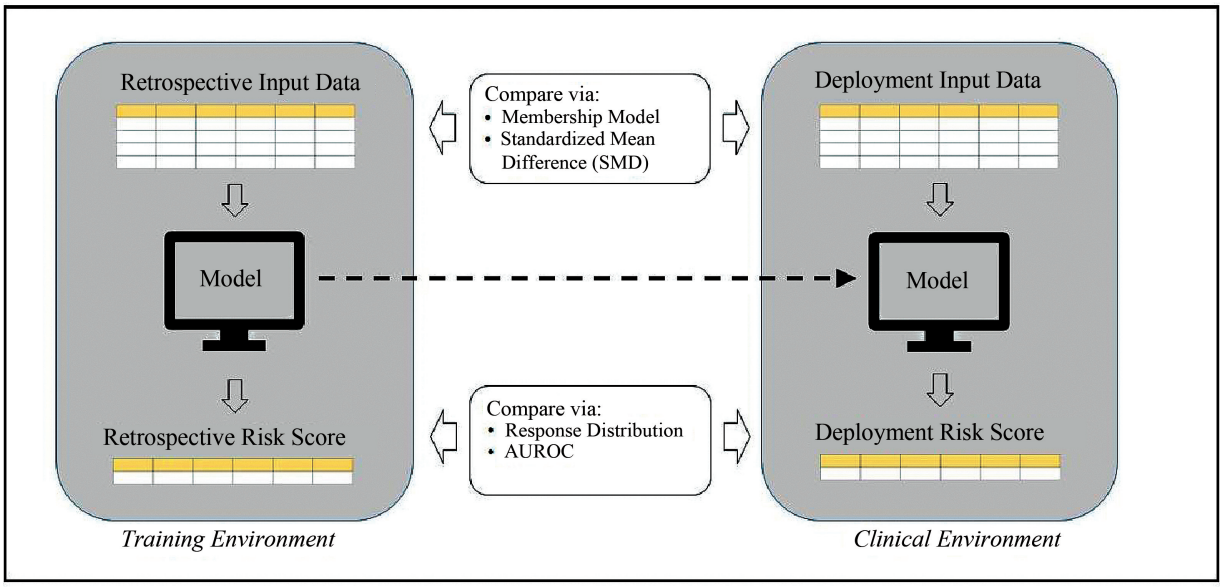

Fig. 2 Research methods summary. In the training environment, retrospective input data was used to train a CKD risk model, which then generated CKD risk scores for a hold-out set of retrospective observations. The model is then deployed in the clinical environment and run "silently" on a deployment input dataset, generating associated deployment risk scores for each observation. Following a 12-month lag period, the deployment risk scores were then paired with observed patient outcomes. Input datasets were compared via standardized mean difference (SMD), performance of a membership model and response distributions. Model performances compared via AUROC. AUROC, area under the receiver-operator curve; CKD, chronic kidney disease.

Table 1 Summary of patient, admission, and laboratory characteristics for the retrospective and deployment datasets

Characteristic	Retrospective ($n = 4,879$)	Deployment ($n = 1,270$)
Age, mean, years (SD)	9.0 (5.6)	9.0 (5.8)
Sex, *N* (%)		
Male	2,506 (51.4)	641 (50.5)
Female	2,373 (48.6)	629 (49.5)
Race, *N* (%)		
White	1,817 (37.2)	435 (34.3)
Asian	921 (18.9)	205 (16.1)
African American	182 (3.7)	26 (2.0)
Native Hawaiian	70 (1.4)	14 (1.1)
American Indian	7 (0.1)	5 (0.4)
Other or Unknown	1,882 (38.6)	585 (46.0)
Length of Stay, mean, days (SD)	10.5 (16.6)	11.1 (22.0)
Count Creatinine Labs, Inpatient or Outpatient, mean (SD)	9.9 (18.5)	8.8 (19.8)
Creatinine Baseline, mean (SD)	0.5 (0.3)	0.4 (0.3)
Max AKI Stage, *N* (%)		
0	3,041 (62.3)	1,027 (80.9)
1	979 (20.1)	140 (11.0)
2	665 (13.6)	70 (5.5)
3	194 (4.0)	33 (2.6)

Abbreviation: AKI, acute kidney injury.

validation with L1 regularization penalty, the model showed an AUROC of 0.71 ($p < 0.0001$).

Response distribution analysis of the deployment risk scores showed a smooth, unimodal distribution with a mean of 0.3 and standard deviation of 0.04 (►Fig. 3). This distribution was significantly different from the response distribution of the retrospective risk scores (Kolmogorov–Smirnov = 0.13, $p < 0.0001$).

Table 2 Results for standardized mean differences (SMD) analysis of input features, broken down by clinically relevant feature groups

Feature Group	Example	Number of features	*N* (%) of significantly different features ($p < 0.05$)
Patient or admission characteristics	Patient age, race	10	6 (60)
Baseline kidney function	Baseline creatinine by age, baseline eGFR	8	5 (63)
Changes in Kidney Function	Max increase in creatinine from baseline	9	9 (100)
AKI—Occurrence	AKI by age baseline	24	23 (96)
AKI—Stage	Max AKI stage by age baseline	24	23 (96)
Total		75	66 (88)

Abbreviations: AKI, acute kidney injury; eGFR, estimated glomerular filtration rate.

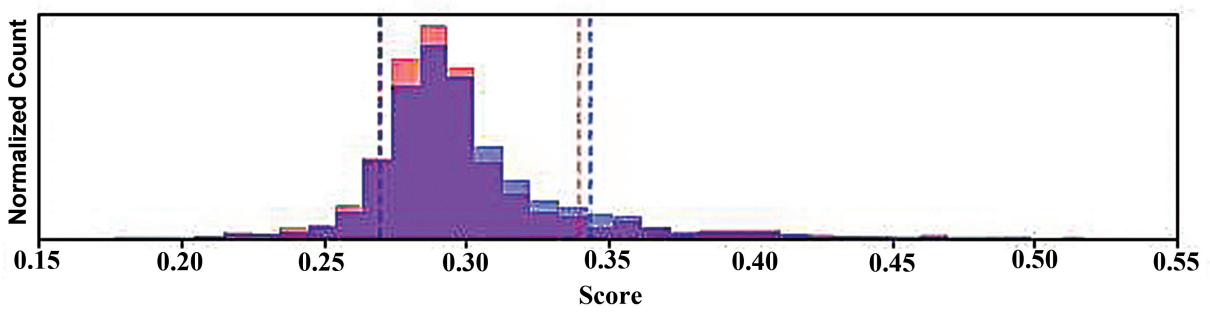

Fig. 3 Response Distribution plots for deployment (*red*) and retrospective (*blue*) chronic kidney disease risk score. Vertical lines indicate the 10th and 90th percentiles.

The observed outcomes of the 1,270 admission follow-ups in the deployment dataset were: CKD present in 20/1,270 (1.6%) cases, CKD absent in 360/1270 (28.3%) cases, and insufficient data available in 890/1,270 (70.1%) cases. The model demonstrated an AUROC of 0.63. This performance is a significant decrease from the performance obtained on retrospective data (retrospective AUROC = 0.76; $p = 0.033$). See ►**Table 3** for study results summary.

Discussion

This study evaluates three label-blind metrics as early warning indicators of performance deterioration and shows that all three correspond to an observed decrease in model performance after deployment into a pediatric hospital. These results suggest that these metrics could contribute to a near real-time assessment of model transportability and improve the safety of deployed machine learning algorithms.

This study is an early evaluation of the correspondence between dataset consistency and model performance in a live clinical environment, and the first to do so with CKD risk prediction. In this study, the retrospective and deployment datasets were significantly different from one another, as indicated by the significant SMDs, "membership model" performance and response distribution analysis. The model performed significantly worse in the deployment environment than the training environment (AUROC decreased from 0.76 to 0.63, $p = 0.033$). Taken together, these results suggest that this CKD model has poor transportability (i.e., it is not generalizable[27]) to the LPCH patient population seen during the deployment period.

This work builds on the framework proposed by Debray et al by applying significance testing to the evaluation of label-blind metrics, which are critical for interpretation and subsequent actionability. For example, while the Debray framework suggests calculation of AUROC for a membership model, explicit guidance on how that AUROC is evaluated and incorporated into a broader assessment of model interpretability is not provided. The application of significance testing offers a reasonable quantitative evaluation to inform downstream action that can be incorporated into a model monitoring program. However, significance testing is only a starting point, as model performance may be susceptible to dataset shifts that do not meet the threshold for significance outlined here, and likewise may be resilient to such extreme shifts in data.[27] Thus the results presented here are insufficient, in isolation, to provide specific thresholds to trigger model re-evaluation.

These label-blind metrics offer specific monitoring tools that are currently lacking in the deployed ML literature. A recent analysis of 15 model reporting guidelines found that only 10 included any reference to monitoring of model performance.[29] Of these, the most in-depth guideline offers a set of "monitoring tests" to consider, but details of how to

Table 3 Results summary of label-blind metrics and model performance

	Result	Interpretation
Label-blind metric		
Standardized mean differences (SMD)	66/75 (88%) input variables significantly different	Datasets are significantly different ($p < 0.05$)
Membership model	AUROC = 0.71	Datasets are significantly different ($p < 0.0001$)
Response distribution analysis	Kolmogorov–Smirnov = 0.13	Datasets are significantly different ($p < 0.0001$)
Model performances	AUROC (retrospective) = 0.76	Deployed model performance is significantly lower than retrospective model performance ($p = 0.033$)
	AUROC (deployment) = 0.63	

Abbreviation: AUROC, area under the receiver-operator curve.

execute these tests are limited or absent.[30] This ambiguity leaves critical decisions about performance monitoring to health systems deployment teams, which may lack necessary data science or statistical expertise to operationalize such guidelines.

The utility of these label-blind metrics lies in their early availability. In this case, the data necessary for their calculations was available within the EHR on October 1, 2019, whereas model performance could not be measured until October 1, 2020. With the results of the label-blind metrics, a model monitoring team would have known that the deployment data was different from the retrospective data, and in the absence of outside evidence suggesting the model was highly generalizable to this patient population, they would have an increased suspicion about poor model performance and could increase model oversight soon after deployment.

It is important to recognize that the observed decrease in the deployed model's AUROC performance does not necessarily equate to clinically significant changes in performance.[31] As a triage tool for nephrologist referral for a low prevalence condition (the prevalence of CKD in pediatric patients is estimated to be 82 cases per million per year[32]), the positive predictive value would be impractically low for all but the highest risk scores, thus the threshold value to refer the patient would likely be set near its upper limit. This threshold would also be informed by the clinical constraints—namely the available capacity of local outpatient nephrologists to care for the additionally referred patients. For these two reasons, the model's true task is to identify only the highest risk patients and its performance on the rest of the patients is, in some sense, irrelevant. While evaluation techniques such as net reclassification improvement[33] or simulation-based evaluations[34] could potentially capture this nuance, AUROC was used here because it remains the gold-standard for evaluating predictive performance.

►**Table 1** shows that patients in the retrospective dataset developed AKI almost twice as frequently as patients in the deployment dataset (37.7 vs. 19.1%). This may partially reflect greater provider awareness of AKI risk factors, such as nephrotoxic medications[35] and AKI morbidity,[23] leading to the development of automated AKI surveillance tools.[36,37] Such tools, when used as part of quality improvement efforts focused on AKI in pediatric hospitals, have decreased AKI rates by as much as 64%.[38] LPCH institution initiated a similar quality improvement effort in 2016 as part of a nine-site collaborative study.[39]

This study is not without limitation. The criteria used to label CKD-positive cases in our dataset is an approximation of the commonly used clinical criteria to diagnose CKD[40] and likely overestimates the prevalence of the disease. While the criteria used here required two measurements of eGFR <60 mL/min/1.73 m^2 during the follow-up window, the KDIGO criteria for CKD diagnosis requires eGFR <60 mL/min/1.73 m^2 for greater than 3 months with implications for the patient's health. Without the 3-month constraint, our methodology potentially labels recurrent AKI episodes or a single AKI episode with multiple eGFR measurements as CKD. While the patient population at a tertiary care center with specialized nephrology care (including renal transplantation) likely has higher rates of renal pathology, the observed prevalence of CKD in our deployment population is 1.6%, which is orders of magnitude higher than estimates of CKD prevalence in pediatrics. It should be noted that we define CKD-positive patients as those with moderate to severe CKD, defined as stage 3 or greater, which excludes patients with preserved glomerular filtration rate as seen in earlier disease states (i.e., Stages 1 or 2).

Further limitations are caused by the substantial proportion of deployment dataset patients who had insufficient data available during the follow-up period (890/1,270 [70.1%]). As a tertiary care center with a wide referral footprint, many patients receive post-hospital care outside of our institution and lost to our follow-up. While this attrition rate itself would not be expected to impact the model's performance on our local population, it would require additional consideration if the model were to be deployed at an outside institution.

Rigorous validation of deployed machine learning models in medicine is necessary to ensure their safe and effective use. This study contributes to this effort, however, more research is needed to identify optimum methods and metrics before these models can become part of routine care.

Conclusion

This study suggests that the three label-blind metrics considered do correlate with a decrease in deployed model performance and could be useful as early indicators of degradation in model performance. Rigorous validation of deployed machine learning algorithms is critical to ensure their efficacy and safety.

Supplementary Material

Supplementary Material. Complete input feature list of CKD algorithm. For each feature, includes SMD value, *p*-value, and value summary from retrospective and deployment datasets (available in the online version).

Clinical Relevance Statement

Health systems hoping to utilize machine learning to improve patient care and outcomes must also ensure patient safety by monitoring the performance of the models once deployed. While numerous monitoring areas have been proposed, appropriate metrics are not well established. This paper outlines three label-blind metrics that could be used to identify deterioration of model performance once deployed in clinical environments.

Multiple Choice Questions

1. A "membership model" is an algorithm designed to:
   a. Evaluate the accuracy of another algorithm's predictions.
   b. Identify the dataset source of a set of input features.
   c. Identify areas of bias within a model's training data.
   d. Suggest additional members, or classes, of data to include in model training.

   Correct Answer: The correct answer is option b. A membership model takes as input features datasets from multiple sources (i.e., retrospective data and prospective data) and is trained to predict which dataset source the input features come from. A membership model that is able to distinguish between observations in the two datasets with high accuracy suggests that there are substantial (potentially complex and multivariate) differences between the two datasets.

2. Referral algorithms for low prevalence conditions, like the CKD algorithm described here, are affected by what two operational constraints?
   a. Low positive predictive value; limited downstream specialist capacity.
   b. Poor inter-rater reliability; limited downstream specialist capacity.
   c. Low positive predictive value; provider alert fatigue.
   d. Poor inter-rater reliability; provider alert fatigue.

   Correct Answer: The correct answer is option a. Referral algorithms for low prevalent conditions are operationally constrained by low PPV and limited downstream specialist capacity. Even with excellent sensitivity and specificity, screening algorithms suffer from low PPV and high false positive rates when true disease prevalence is low. Algorithm utility is depended on available specialist capacity to see additional patients.

Projection of Human and Animal Subjects

This project was reviewed and approved by the Stanford University Institutional Review Board.

Funding

None.

Conflict of Interest

None declared.

Acknowledgments

The authors would like to thank the Lucile Packard Children's Hospital Information Services (IS) Department for their support of this research effort.

References

1 Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. N Engl J Med 2019;380(14):1347–1358
2 Bates DW, Auerbach A, Schulam P, Wright A, Saria S. Reporting and implementing interventions involving machine learning and artificial intelligence. Ann Intern Med 2020;172(11):S137–S144
3 Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. BMJ Qual Saf 2019;28(03):231–237
4 Wong A, Otles E, Donnelly JP, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Intern Med 2021;181(08):1065–1070
5 Sendak MP, Balu S, Schulman KA. Barriers to achieving economies of scale in analysis of EHR data. a cautionary tale. Appl Clin Inform 2017;8(03):826–831
6 Pencina MJ, Goldstein BA, D'Agostino RB. Prediction models—development, evaluation, and clinical application. N Engl J Med 2020;382(17):1583–1586
7 Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med 1999;130(06):515–524
8 Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. J Am Med Inform Assoc 2017;24(06):1052–1061
9 Bedoya AD, Futoma J, Clement ME, et al. Machine learning for early detection of sepsis: an internal and temporal validation study. JAMIA Open 2020;3(02):252–260
10 Nestor B, McDermott MBA, Boag W, et al. Feature Robustness in Non-stationary Health Records: Caveats to Deployable Model Performance in Common Clinical Machine Learning Tasks. MLHC. Published online 2019. Accessed June 7, 2021 at: https://www.semanticscholar.org/paper/dcbf6137fe16b33c2e2d9258bd4a1e3cdabee48f
11 Moons KGM, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart 2012;98(09):691–698
12 Finlayson SG, Subbaswamy A, Singh K, et al. The clinician and dataset shift in artificial intelligence. N Engl J Med 2021;385(03):283–286
13 Rabanser S, Günnemann S, Lipton ZC. Failing Loudly: An Empirical Study of Methods for Detecting Dataset Shift. arXiv. Published

online October 29, 2018. Accessed June 1, 2021 at: http://arxiv.org/abs/1810.11953
14 Bernardi L, Mavridis T, Estevez P. 150 Successful Machine Learning Models: 6 Lessons Learned at Booking.com. Paper presented at: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. KDD '19. Association for Computing Machinery; 2019:1743–1751
15 Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol 2015;68(03):279–289
16 Kaddourah A, Basu RK, Bagshaw SM, Goldstein SLAWARE Investigators. Epidemiology of acute kidney injury in critically ill children and young adults. N Engl J Med 2017;376(01):11–20
17 Coca SG, Singanamala S, Parikh CR. Chronic kidney disease after acute kidney injury: a systematic review and meta-analysis. Kidney Int 2012;81(05):442–448
18 Silver SA, Goldstein SL, Harel Z, et al. Ambulatory care after acute kidney injury: an opportunity to improve patient outcomes. Can J Kidney Health Dis 2015;2:36
19 Kaspar CDW, Bholah R, Bunchman TE. A review of pediatric chronic kidney disease. Blood Purif 2016;41(1-3):211–217
20 Hogg RJ, Furth S, Lemley KV, et al; National Kidney Foundation's Kidney Disease Outcomes Quality Initiative. National Kidney Foundation's Kidney Disease Outcomes Quality Initiative clinical practice guidelines for chronic kidney disease in children and adolescents: evaluation, classification, and stratification. Pediatrics 2003;111(6 Pt 1):1416–1421
21 Goldstein SL, Jaber BL, Faubel S, Chawla LSAcute Kidney Injury Advisory Group of American Society of Nephrology. AKI transition of care: a potential opportunity to detect and prevent CKD. Clin J Am Soc Nephrol 2013;8(03):476–483
22 Glenn D, Ocegueda S, Nazareth M, et al. The global pediatric nephrology workforce: a survey of the International Pediatric Nephrology Association. BMC Nephrol 2016;17(01):83
23 Khwaja A. KDIGO clinical practice guidelines for acute kidney injury. Nephron Clin Pract 2012;120(04):c179–c184
24 Schwartz GJ, Haycock GB, Edelmann CM Jr, Spitzer A. A simple estimate of glomerular filtration rate in children derived from body length and plasma creatinine. Pediatrics 1976;58(02):259–263
25 Faraone SV. Interpreting estimates of treatment effects: implications for managed care. P&T 2008;33(12):700–711
26 Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav Res 2011;46(03):399–424
27 Subbaswamy A, Saria S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. Biostatistics 2020;21(02):345–352
28 Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. BMJ 1995;310(6973):170
29 Lu JH, Callahan A, Patel BS, Morse KE, Dash D, Shah NH. Low adherence to existing model reporting guidelines by commonly used clinical prediction models. bioRxiv. Doi: 10.1101/2021.07.21.21260282
30 Breck E, Cai S, Nielsen E, Salib M, Sculley D. The ML test score: A rubric for ML production readiness and technical debt reduction. Paper presented at: 2017 IEEE International Conference on Big Data (Big Data); 2017:1123–1132
31 Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med 2000;19(04):453–473
32 Massengill SF, Ferris M. Chronic kidney disease in children and adolescents. Pediatr Rev 2014;35(01):16–29
33 Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 2008;27(02):157–172
34 Mišić VV, Rajaram K, Gabel E. A simulation-based evaluation of machine learning models for clinical decision support: application and analysis using hospital readmission. NPJ Digit Med 2021;4(01):98
35 Sethi SK, Bunchman T, Chakraborty R, Raina R. Pediatric acute kidney injury: new advances in the last decade. Kidney Res Clin Pract 2021;40(01):40–51
36 Goldstein SL, Kirkendall E, Nguyen H, et al. Electronic health record identification of nephrotoxin exposure and associated acute kidney injury. Pediatrics 2013;132(03):e756–e767
37 Wang L, McGregor TL, Jones DP, et al. Electronic health record-based predictive models for acute kidney injury screening in pediatric inpatients. Pediatr Res 2017;82(03):465–473
38 Goldstein SL, Mottes T, Simpson K, et al. A sustained quality improvement program reduces nephrotoxic medication-associated acute kidney injury. Kidney Int 2016;90(01):212–221
39 Goldstein SL, Dahale D, Kirkendall ES, et al. A prospective multicenter quality improvement initiative (NINJA) indicates a reduction in nephrotoxic acute kidney injury in hospitalized children. Kidney Int 2020;97(03):580–588
40 Kidney Disease: Improving Global Outcomes (KDIGO) CKD Work Group. KDIGO 2012 Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease. Kidney Int Suppl 2013;3:1–150