


Research and Applications

Development and external validation of multimodal postoperative acute kidney injury risk machine learning models

George K. Karway , PhD¹, Jay L. Koyner, MD², John Caskey, PhD¹, Alexandra B. Spicer, MS¹, Kyle A. Carey, MPH², Emily R. Gilbert, MD³, Dmitriy Dligach, PhD⁴, Anoop Mayampurath, PhD^{1,5}, Majid Afshar, MD, MS^{1,5}, Matthew M. Churpek, MD, MPH, PhD^{*,1,5}

¹Department of Medicine, University of Wisconsin-Madison, Madison, WI 53792, United States, ²Section of Nephrology, Department of Medicine, University of Chicago, Chicago, IL 60637, United States, ³Department of Medicine, Loyola University Chicago, Chicago, IL 60153, United States, ⁴Department of Computer Science, Loyola University Chicago, Chicago, IL 60626, United States, ⁵Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI 53726, United States

*Corresponding author: Matthew M. Churpek, MD, MPH, PHD, Department of Medicine and Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, 600 Highland Ave, Madison, WI 53792 (mchurpek@medicine.wisc.edu)

Abstract

Objectives: To develop and externally validate machine learning models using structured and unstructured electronic health record data to predict postoperative acute kidney injury (AKI) across inpatient settings.

Materials and Methods: Data for adult postoperative admissions to the Loyola University Medical Center (2009-2017) were used for model development and admissions to the University of Wisconsin-Madison (2009-2020) were used for validation. Structured features included demographics, vital signs, laboratory results, and nurse-documented scores. Unstructured text from clinical notes were converted into concept unique identifiers (CUIs) using the clinical Text Analysis and Knowledge Extraction System. The primary outcome was the development of Kidney Disease Improvement Global Outcomes stage 2 AKI within 7 days after leaving the operating room. We derived unimodal extreme gradient boosting machines (XGBoost) and elastic net logistic regression (GLMNET) models using structured-only data and multimodal models combining structured data with CUI features. Model comparison was performed using the receiver operating characteristic curve (AUROC), with Delong's test for statistical differences.

Results: The study cohort included 138 389 adult patient admissions (mean [SD] age 58 [16] years; 11 506 [8%] African-American; and 70 826 [51%] female) across the 2 sites. Of those, 2959 (2.1%) developed stage 2 AKI or higher. Across all data types, XGBoost outperformed GLMNET (mean AUROC 0.81 [95% confidence interval (CI), 0.80-0.82] vs 0.78 [95% CI, 0.77-0.79]). The multimodal XGBoost model incorporating CUIs parameterized as term frequency-inverse document frequency (TF-IDF) showed the highest discrimination performance (AUROC 0.82 [95% CI, 0.81-0.83]) over unimodal models (AUROC 0.79 [95% CI, 0.78-0.80]).

Discussion: A multimodality approach with structured data and TF-IDF weighting of CUIs increased model performance over structured data-only models.

Conclusion: These findings highlight the predictive power of CUIs when merged with structured data for clinical prediction models, which may improve the detection of postoperative AKI.

Lay Summary

Acute kidney injury (AKI) after an operation, called postoperative AKI, is common in hospitalized patients and associated with increased morbidity and mortality. Early detection of high-risk patients could facilitate timely treatment and improve outcomes. Although a few studies have developed machine learning (ML) models to identify patients with postoperative AKI, these are primarily limited to structured data (eg, laboratory values) and ignore predictors from clinical notes. Further, models built from clinical notes are often not externally validated because doing so risks leaking protected health information.

Given these limitations in the field, we developed and externally validated ML models to predict postoperative AKI using structured data and information from clinical notes. To preserve patient privacy, we used concept unique identifiers (CUIs), which are de-identified medical terms from clinical notes. We compared unimodal models with structured data to multimodal models with CUIs plus structured data, as well as different approaches to modeling the CUI data. We found that multimodal models significantly improved model performance compared to unimodal models. We also found that normalizing CUI data based on term frequency had the highest performance. In conclusion, using CUIs to account for information in clinical notes adds significant value for predicting postoperative AKI.

Key words: multimodal models; artificial intelligence; intensive care unit; machine learning; acute kidney injury; natural language processing.

Background and significance

The development of acute kidney injury (AKI) in the postoperative period is common in hospitalized patients and is associated with a significant increase in morbidity and mortality, as well as prolonged duration and increased costs of hospitalization.^{1–6} AKI is defined by either an increase in serum creatinine (Scr) or a decrease in urine output according to the Kidney Disease Improving Global Outcomes (KDIGO) consensus definition established over the last decade.⁷ The incidence of AKI varies from 5.0% to 7.5% in hospitalized patients receiving acute care and can reach up to 20% in patients admitted to the intensive care unit (ICU).⁸ While the incidence of postoperative AKI depends on the type of surgery, 30%–40% of all in-hospital AKI cases are related to surgical procedures.⁹ Patients who develop postoperative AKI are more likely to develop sepsis and coagulopathy,⁸ have a higher risk of receiving prolonged mechanical ventilation,⁸ and their mortality rate at 30 days can be 15-fold higher than those who do not develop postoperative AKI.^{10,11} Early detection of individuals at high risk for postoperative AKI could facilitate timely therapies, resulting in improved quality of care and more efficient use of hospital resources.

While machine learning (ML) models have been developed to detect patients with AKI,^{12–18} many algorithms rely on data from a single modality, such as structured data variables like vital signs and laboratory values.^{13,19,20} However, electronic health records (EHRs) contain a wealth of information stored in various modalities, including clinical notes and radiology reports. Nearly 80% of EHR data reside in an unstructured free text format.^{21,22} The application of natural language processing (NLP) to extract clinically meaningful information from free text is effective in research and clinical practice,^{23–28} and has improved performance of risk stratification models.^{27,29} Leveraging NLP combined with structured data for multimodal learning to predict postoperative AKI risk could improve the performance of these models, resulting in more accurate detection of at risk patients.

External validation is essential to determine a prediction of model's reproducibility and generalizability to a new and different patient population,^{30,31} but few post-operative AKI models have been externally validated.^{32–35} This is compounded by the challenge of externally validating NLP models, which risk leaking protected health information (PHI) when trained on clinical notes.^{36–40}

Objectives

In this study, we aim to develop and externally validate ML models to predict postoperative AKI using structured data and medical concepts extracted from raw clinical notes as concept unique identifiers (CUIs). We further aim to compare different approaches to modeling CUI data to determine which method demonstrates the highest discrimination and calibration. We hypothesize that multimodal models that combine CUIs and structured data will achieve higher performance than unimodal models with structured data only for predicting postoperative AKI.

Methods

Study population

We conducted a retrospective cohort study of adult (≥ 18 years) postoperative patients at Loyola University Medical

Center (LUMC) between 2009 and 2017 and at the University of Wisconsin-Madison Hospital (UW) between 2009 and 2020. Patients were excluded if they had any of the following: (1) no operating room (OR) location; (2) chronic end-stage renal failure billing diagnosis prior to hospital admission; (3) no documented SCr measurement before discharge; (4) had an initial admitting SCr value greater than or equal to 3.0 mg/dL; (5) required kidney replacement therapy (KRT) before or within 48 h of their first documented SCr measurement; (6) developed KDIGO stage 1 before first OR discharge, or (7) had no clinical documentation (eg, notes, reports) before developing stage 2 AKI. Additionally, all admissions that did not have data in the following time intervals were excluded: (1) between first SCr measurement and last vitals; (2) in the emergency department (ED), ward, or ICU; or (3) in the 7-day window following first OR. These exclusions were made to ensure that patients did not have pre-existing renal failure and ensured data were available for the model to predict AKI. The exclusions of those with an initial SCr greater than or equal to 3.0 mg/dL and KRT within 48 h of first SCr measurement were done to minimize the misclassification of patients with pre-existing AKI prior to admission, which is consistent with other AKI studies.^{13,14,18} The study was approved by the LUMC and UW institutional review boards with a waiver of informed consent.

Data collection

Structured data and raw clinical notes were extracted from the Clinical Research Data Warehouses at LUMC and UW. Structured data included demographic characteristics, vital signs, laboratory results, trends of vital signs and laboratory values (eg, highest heart rate in previous 24 h), patient location data (eg, ward, ED, ICU, OR), interventions, and nursing documentation (eg, Braden score). Structured data occurring closest to (at or before) the first ward, ICU, or ED observation following the patient's operation were included as model predictor variables. See [Table S1](#) for the full list of structured variables.

The Apache clinical Text Analysis and Knowledge Extraction System⁴¹ was utilized to map the raw text from the clinical notes and reports to CUIs from the National Library of Medicine Unified Medical Language Systems (UMLS). The CUIs are codes that were derived from the Systematized Nomenclature of Medicine—Clinical Terms medical vocabulary from UMLS Metathesaurus.⁴² The CUIs were obtained from clinical notes collected across the entire encounter consisting of operative notes, progress notes, ancillary notes (eg, physical therapy, occupational therapy, and dietitian notes), procedural notes, and radiology reports. Utilization of CUIs minimizes the number of variables by mapping similar terms to a single code (eg, hematoma, hematomas, and blood clots all map to CUI C0018944). CUIs from all notes occurring prior to and including the time a patient arrived at the wards, ICU, or ED following their OR stay were used as input features to our ML models.

Study outcome

The primary outcome of the study was the development of Stage 2 AKI within 7 days of the first ward, ICU, or ED observation following the patient's operation. If a patient had multiple operations during an admission, only the first operation was included. The SCr-based criteria from the KDIGO consensus definition⁷ was used to define AKI. As described

previously, baseline SCr measurement was defined as the admission SCr value and was updated on a rolling basis for 48-hour and 7-day criteria, as per the KDIGO guidelines.^{7,13,14}

Statistical analysis

A comparison of characteristics, such as patient demographics, patient location data (eg, ward, ED, ICU, OR), lab results, and outcomes between the sites groups (LUMC and UW), was performed using *t*-tests and analyses of variance for normally distributed data, Wilcoxon rank sum tests, and Kruskal–Wallis tests for nonnormally distributed data, and chi-square for categorical data. Missing data were handled separately in each cohort by using the median by location within the hospital (for continuous data) or 0 (for categorical, eg, interventions data).¹⁴ See Table S2 for the summary of missing data.

All analyses were performed using R version 3.6.3 (The R Project for Statistical Computing). Model development and validation along with data visualization were performed using several packages including glmnet, glmnetUtils, pROC, XGBoost, ParBayesianOptimization, caret, ggplot2, and tidyverse. Statistical significance was set at $P < .05$, and all tests were 2-tailed.

Machine learning model development

The extreme gradient boosting machine (XGBoost) algorithm was used to predict the study outcome of stage 2 AKI or higher within 7 days after leaving the OR. XGBoost is a gradient boosted machine algorithm that combines weak learners (decision trees) to achieve stronger overall discrimination.⁴³ This iterative process builds trees aiming to better predict cases of post-operative AKI missed by earlier trees by weighting the difficult-to-predict cases to a greater degree. This results in a tree ensemble model that is more accurate than any one individual tree. This algorithm is recommended for classification and regression problems with tabular data.^{43,44} Hyperparameters, including the number of trees, depth of trees, learning rate, and the minimum size of the terminal leaves were tuned to maximize the area under the receiver operating characteristic curve (AUROC) using 5-fold cross-validation on the training dataset (LUMC cohort) (see Table S3 for list of tuning grids used). As a more interpretable comparator model to XGBoost, elastic net logistic regression models were also derived, which combine lasso and ridge regularization penalties in a single framework. These penalty terms were tuned using 5-fold cross-validation in the training data.

For the unimodal models, demographics, vital signs, laboratory values, interventions, medications, nursing documentation, and diagnostics (see Table S1 for a full list of predictors) from LUMC were used to train the models, and they were externally validated using the same features from UW structured data. For the CUI dataset, 3 approaches were used to engineer the representation of CUIs: binary (absence versus presence of CUIs), TF (term frequency count of each unique CUI), or TF-IDF (term frequency inverse document frequency, a normalized version of the TF). For model development, only CUIs that appeared in at least 5 admissions from the training dataset were retained and used to train the models (based on exploratory analysis of CUI distributions). Predictors from each CUI dataset defined above were

concatenated with features from the structured data to create a multimodal model (Figure S1 for the model architecture).

Model evaluation

Model performances were validated on the external testing dataset (UW cohort). For multimodal model testing, all CUIs included in the models were retained and CUIs in the training set not appearing in the test set were set to zero (see Figure S2 for details). Delong's test for differences in AUROC using the trapezoidal method was used to compare the performance of unimodal and multimodal models.^{45,46} Sensitivity, specificity, and positive and negative likelihood ratios (LR+ and LR−) were compared across a range of thresholds for the XGBoost unimodal and multimodal models (see Table S4). As a secondary metric for comparison, model calibration, which compares the true probability of the outcome versus a model's prediction, was compared using calibration slope, intercept, Brier score, and unreliability index *P*-value. A well-calibrated model is expected to have an intercept of zero, a slope of 1, an unreliability index *P*-value $> .05$, and a low Brier score. The importance of each model feature was computed to examine the predictive influence of structured vs CUI features. The gain-based variable importance metric, which considers both the quantity and quality of the split associated with each feature was used for XGBoost models. This approach provides insights into which feature contributes the most to the model's predictive performance by taking into account the number of times a feature is used for splitting (frequency) and the magnitude of improvement achieved (gain).⁴³ The coefficient-based variable importance metric was used for the GLMNET model. This method examines the magnitude of the estimated coefficients. Features with larger absolute coefficients are considered more important, as they have a strong influence on models' predictions.⁴⁷

Results

Patient characteristics and comparison

There were 138 389 adult patient encounters included in the final cohort, with 61 257 admissions from LUMC and 77 132 admissions from UW. Among those, 2959 (2.1%) developed stage 2 AKI or higher within a week after their operation, with 1445 (2.4%) from LUMC and 1514 (2.0%) from UW. See Figure 1 for the exclusions leading to this cohort. Table 1 shows patient characteristics and clinical outcomes across the 2 sites. Compared to UW, the LUMC cohort was older and had a higher proportion of African-American and female patients (P -value $< .001$). Additionally, LUMC admissions had a higher admission SCr measurement, higher blood urea nitrogen (BUN) concentration, longer ICU stay, and a higher proportion of prior ICU visits than UW admissions (P -value $< .001$).

Unimodal and multimodal machine learning model performance comparisons

The vocabulary size for CUIs in notes collected prior to patients' first observation following their departure from the OR in the training data (LUMC) was 37 570 and the vocabulary size for the validation data (UW) was 43 494. After eliminating rare CUIs (≥ 5 encounters) in the training data, 22 106 (59%) CUIs were selected as features for the training dataset. The unimodal model with structured-only data had 61 variables as inputs and the multimodal models with both

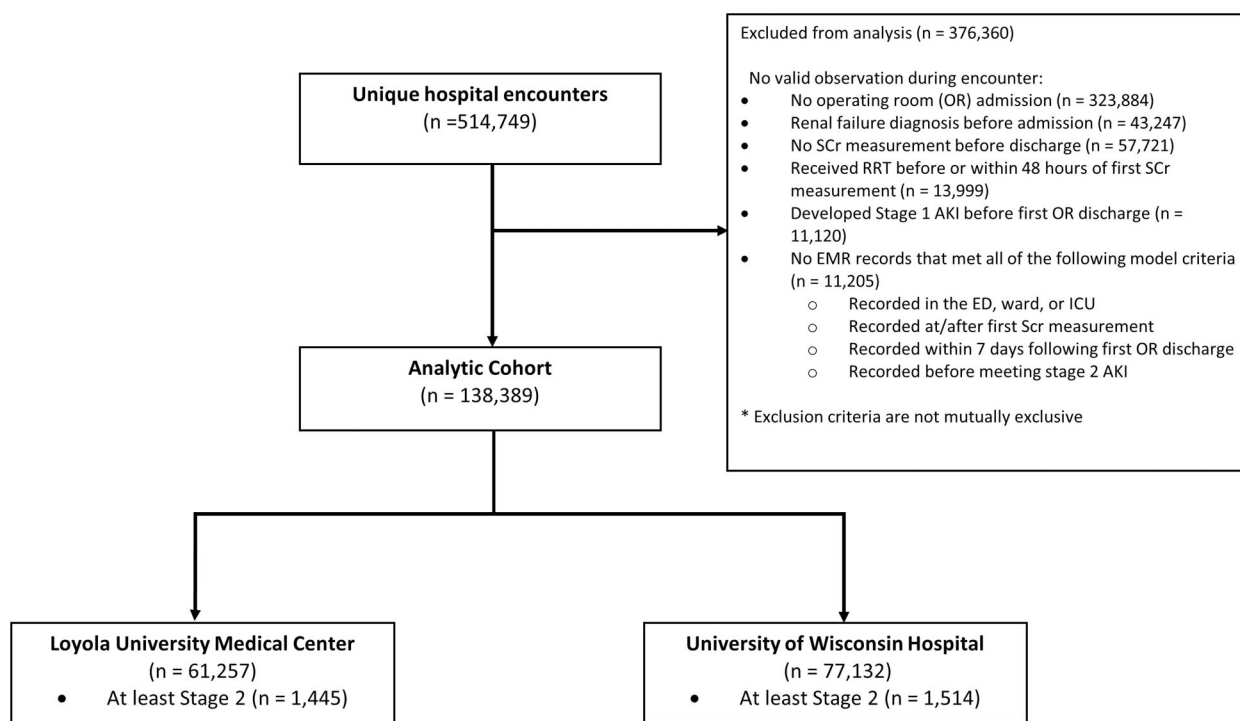


Figure 1. Consort diagram of study cohort included in ML models.

Table 1. Patient characteristics and clinical outcomes comparison between sites (n = 138 389).

Variables	LUMC (n = 61 257)	UW (n = 77 132)	P-value
Demographics			
Age, mean (SD), yr	58 (16)	57 (16)	<.001
Race, n (%)			
African American	8940 (15)	2566 (3)	<.001
Others	52 317 (85)	74 566 (97)	
Sex, n (%)			
Female	32 019 (52)	38 807 (50)	<.001
Lab values			
Admission Scr (mg/dL), mean (SD)	0.99 (0.38)	0.90 (0.33)	<.001
Admission blood urea nitrogen (mg/dL), mean (SD)	14.3 (8.9)	15.0 (8.1)	<.001
Hospital stay details			
Length of hospital stay (hours), median (Q1, Q3)	99 (51, 187)	99 (58, 174)	<.001
Location of AKI, n (%)			
Ward	35 591 (58.1)	59 077 (76.59)	N/A
ICU	25 582 (41.8)	18 037 (23.38)	<.001
Emergency department/other	84 (0.14)	18 (0.02)	<.001

SD, standard deviation; yr, year; ICU, intensive care unit; length of hospital stay, length of patient hospital stay during this admission.

CUIs and structured data combination had 22 167 input features. See Figure S2 for the exclusions leading to the final training and validation CUI dataset.

Table 2 shows the AUROC in the validation dataset (UW) for the models. XGBoost consistently outperformed GLMNET across all data types, with an average AUC of 0.81 (95% CI, 0.80-0.82), whereas GLMNET achieved an average AUC of 0.78 (95% CI, 0.77-0.79). Moreover, the multimodal XGBoost model incorporating TF-IDF showed the highest discrimination performance, achieving an AUC of 0.82 (95% CI, 0.81-0.83). This value was significantly higher than the AUC of the unimodal model, which was 0.79 (Delong test P -value < .05). At a similar specificity cut-off (73%), the multimodal model with CUI TF-IDF had a sensitivity of 75%, CUI TF had

a sensitivity of 75%, and CUI Binary had a sensitivity of 73%. These values were greater than the sensitivity of 68% for the unimodal model at the same specificity cut-off. A detailed summary showing the model performance across a range of thresholds for the XGBoost models is shown in Table S4. A pairwise comparison between the unimodal model with structured-only data and multimodal models for XGBoost using Delong test demonstrated statistical improvement in discrimination for multimodality over single modality (Delong test P value < .05 for all comparisons).

Model calibration comparison

Table 3 shows the calibration plot summary data for the unimodal and multimodal models. Although both unimodal

Table 2. Summary of the unimodal versus multimodal models performance in the external validation dataset (UW).

Data	Number of features	ML algorithm	CUI method	AUROC (95% CI)
Structured data	61	XGBoost	N/A	0.79 (0.78-0.80)
CUI + structured data	22 167		Binary	0.81 (0.80-0.82)
			TF	0.81 (0.80-0.82)
			TF-IDF	0.82 (0.81-0.83)
Structured data	61	GLMNET	N/A	0.79 (0.78-0.80)
CUI + structured data	22 167		Binary	0.77 (0.76-0.78)
			TF	0.77 (0.76-0.78)
			TF-IDF	0.80 (0.77-0.79)

Structured data, unimodal models with structured only data; CUI + structured data, multimodal models with CUI and structured data combined; binary, CUI weighted as 0 if absence and 1 if presence; TF, weighting the term frequency count of each unique CUI; TF-IDF, weighting the term frequency inverse document frequency of each unique CUI, a normalized version of the TF.

Table 3. Calibration summary for the unimodal and multimodal models in the validation cohorts.

Data	CUI method	Model	Intercept	Slope	Unreliability index <i>P</i> -value	Brier score
Structured data	N/A	XGBoost	−0.18	0.91	<.01	0.02
CUI + structured data	Binary		0.48	1.05	<.01	0.02
	TF		0.14	1.00	<.01	0.02
	TF-IDF		0.38	1.04	<.01	0.02
Structured data	N/A	GLMNET	−0.23	0.92	<.01	0.02
CUI + structured data	Binary		−0.27	1.00	<.01	0.02
	TF		−3.19	0.90	<.01	0.02
	TF-IDF		−0.25	0.92	<.01	0.02

Structured data, unimodal models with structured only data; CUI + structured data, multimodal models with CUI and structured data combined; Binary, CUI weighted as 0 if absence and 1 if presence; TF, weighting the term frequency count of each unique CUI; TF-IDF, weighting the term frequency inverse document frequency of each unique CUI, a normalized version of the TF.

and multimodal models were not well-calibrated statistically (unreliability *P*-value<.01), the XGBoost multimodal model using TF had the best overall calibration (calibration slope closer to 1 and intercept closer to 0). This observation is supported by the plots for the multimodal models, which displayed predicted probabilities that closely aligned with the actual probabilities (Figure S3).

Feature importance comparison

Figure 2 illustrates the variable importance plot showcasing the top 20 variables in the highest-performing multimodal model. The analysis revealed that the SF ratio (ratio of oxygen saturation in arterial blood to the percentage of oxygen in inspired air), current Scr level, and current fraction of inspired oxygen (FIO2) were the most influential predictors in the model. Notably, in both the XGBoost and GLMNET models, all of the top 20 predictors in the multimodal models were derived from the structured data. However, when comparing the multimodal models to the structured-only model, it became apparent that certain relevant AKI features such as heart rate, change in Scr, and potassium emerged as top 20 predictors exclusively in the XGBoost multimodal models. Detailed examination of the top 50 variables of the XGBoost multimodal model showed 2 CUIs features (neurosurgical procedures and eye) in the CUI binary + structured data model, 5 CUIs in CUI TF + structured data model (transplantation, pain, patient, neurosurgical procedures, and lisinopril), and 10 CUIs in the CUI + TF-IDF models (transplantation, liver cirrhosis, lisinopril, aortic aneurysm abdominal, pain, tobacco, neurosurgical procedure, vomiting, bowel preparation, and neck). See Table S5 for details.

Discussion

In this multicenter study, we developed and externally validated ML risk models that accurately predicted the development of postoperative AKI using structured and unstructured data. We found that a multimodal approach that combines features from notes and structured data increased model performance over a unimodal approach using structured data only. Specifically, we demonstrate that the integration of structured data with TF-IDF weighting of medical concepts from unstructured clinical notes in an XGBoost algorithm predicted at-risk postoperative AKI with the highest discrimination. These results highlight the predictive power of CUIs when merged with structured data for clinical prediction models for predicting AKI.

Although prior research suggests that multimodality models typically perform better than the traditional single modality models, few studies used external validation.^{48–50} For example, Liao et al. applied NLP to the EHR to build a computable phenotype for multiple diseases and showed that the inclusion of NLP to structured data increased sensitivity while either maintaining or improving the accuracy of the phenotype classification algorithms for Crohn's disease, ulcerative colitis, multiple sclerosis, and rheumatoid arthritis than an algorithm with only structured data.⁴⁹ Halpern et al. also observed that NLP in conjunction with structured data resulted in more precise identification of phenotyping of the patients and generalized better than classifier with only structured data.⁵⁰ The fusion of data from different modalities allows our models to harness the full granularity of the EHR and account for the depth of information hidden in the raw clinical notes to predict at-risk patients with superior performance. Our multimodality models contribute to reducing

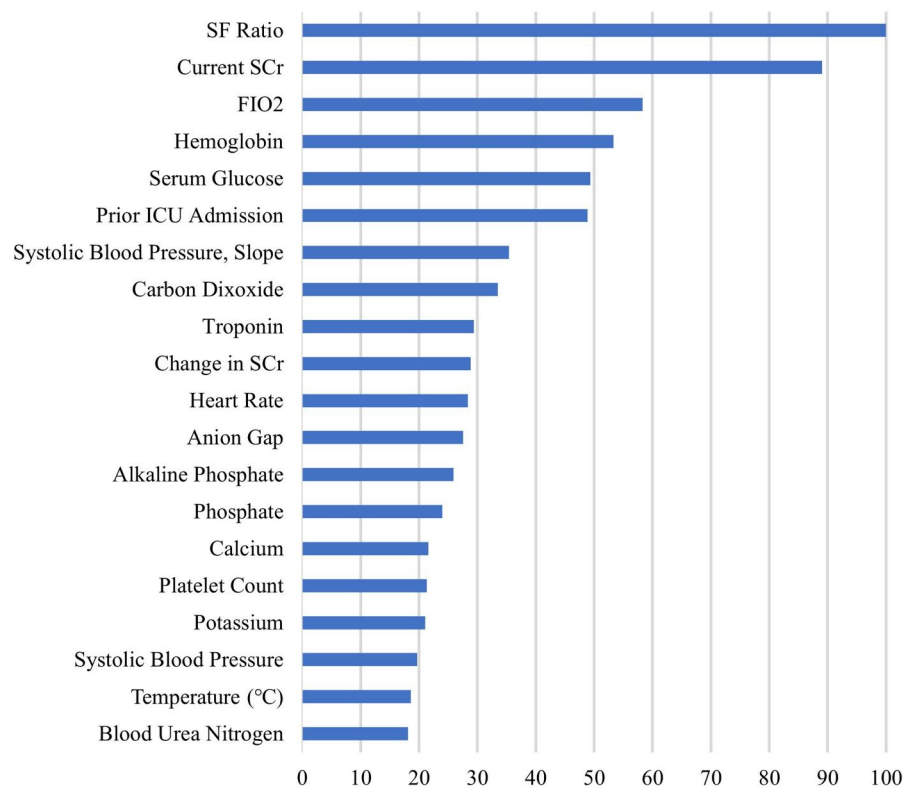


Figure 2. Variable importance plot for the best performing XGBoost model (CUI TF-IDF + structured data) developed in the validation (UW) cohort.

false-positive predictions, thereby alleviating false alarms and the resulting alert fatigue commonly associated with traditional unimodal models that rely on incomplete clinical data.⁵¹ Furthermore, our models are developed in 1 setting and validated in another, enabling investigation of clinical outcomes across different settings and generalization to diverse patient populations. This addresses one of the current limitations observed in existing AKI algorithms.^{13,19,20}

While TF-IDF had better calibration than the other approaches, including an excellent calibration slope (1.00), none of the models studied were well-calibrated overall, which was due to the mis-calibration of the intercept. This mis-calibration is likely due to differences between sites, which could include variations in patient case mix and prevalence, types of surgical procedures, post-operative care, and content of clinical notes. Poor calibration can impact interventional studies that assume the predicted probabilities correlate well to the actual probabilities of the outcome (eg, recommending a screening mammogram if the risk of breast cancer is >10%), so methods to improve calibration of our models, such as Platt scaling and domain adaption,^{52,53} should be considered in these situations. However, our models still order patients from highest to lowest risk well, as shown by their high discrimination, so interventions aimed at the highest risk patients would still be feasible despite their poor overall calibration.^{54,55} These considerations are more important for local adaptation than external validation alone.⁵²

There are several ways to parameterize CUI features in NLP models. In our study, we represented the raw clinical notes as a bag of CUIs and compared 3 approaches to engineer the CUIs features. Previous studies have used different vectorizations of CUIs as ML features. Liao et al. used binary

vectorization of CUIs in their study examining the performance of single modality versus multimodality in phenotype algorithm.⁴⁹ CUIs as TF-IDF and binary vectorizations were used by Kulshrestha et al. in their recent study comparing the performance of bag of CUIs to the bag of words approach.²⁹ However, to the best of our knowledge, ours is the first study that examined 3 different CUI vectorizations as ML features to identify the CUI modeling approach that provides optimal discrimination in the postoperative setting. Our results showed that when modeling CUIs in conjunction with other data types, TF-IDF categorization provided the best discrimination. The high predictive power of TF-IDF can be attributed to the way this modeling approach weights terms within each document across corpus.⁵⁶ TF-IDF not only evaluates how relevant a CUI is to an encounter, but accounts for the significance of that CUI in a collection of encounters. This characteristic enables TF-IDF to discriminate crucial CUIs that occur more frequently but in a limited number of encounters, thereby accounting for variations in CUIs across admissions.

This study has several strengths. First, the representation of raw clinical notes as a bag of CUIs gives us the assurance that our models are compliant with the Health Insurance Portability and Accountability Act (HIPAA), accounts for variations in clinical documentation practices across institutions, and meaningfully reduces the number and complexity of variables needed for modeling. These bag-of-concept features are free of PHI because they effectively de-identify raw text by transforming it into codified data, allowing for the sharing of the trained risk models across institutions without concern for PHI leakage and HIPAA violations.^{22,37} CUIs account for variations in clinical documentation practices between providers and institutions by mapping concepts from the raw

clinical notes to codified data, addressing some of the common problems with semantic analyses of clinical notes including grammatical and spelling errors, lexical variation, and semantic ambiguities of notes.^{57,58} We also minimized the number and complexity of variables used in our ML model significantly by mapping similar terms to a single CUIs. Second, our structured data model uses clinical data that are readily available in the EHR and could be calculated in real time to identify high-risk patients.^{59,60} Third, we externally validated our models in an external health system as opposed to training and validating at the same center. Studies have long established the importance and implications of external validation of multivariable models.^{28,31} Model accuracy often decreases during external validation, and this may be even more impactful when using data from clinical notes given how “smart phrases” and semantics may vary across centers. Finally, our analyses utilized XGBoost as opposed to deep learning models to allow for comparisons and interpretability of the models to better understand the optimal data input as it relates to clinical utility.

Our study also has several limitations. First, we did not use the urine output definitions of AKI due to the inability to accurately measure urine output on an hourly basis in most of our cohorts. We only defined AKI through changes in SCr concentration because of the inability to obtain accurate hourly urine output measurements in all hospitalized patients to comply with KDIGO definitions. However, this is in line with several other previously published AKI risk scores.^{12–14,18} Second, we defined baseline SCr concentration using the admission values as opposed to outpatient values, as we did not have access to outpatient SCr values in our cohort. However, we did define AKI according to international consensus definitions that should allow for replication and validation in future cohorts, and our approach has been used by other AKI modeling studies in the past.^{13,14,34,61} Third, we only utilized 2 of the hundreds of different possible ML algorithms in our study. However, XGBoost is a top-performing algorithm in tabular data,⁶² and logistic regression is a commonly used method in clinical research. Finally, there are other approaches to utilizing PHI to develop ML models while maintaining patient privacy, such as federated learning, which was outside the scope of this manuscript.

Conclusion

We developed and externally validated ML algorithms to identify postoperative patients who develop AKI with high discrimination. A multimodal approach with CUIs weighted as TF-IDF fused with structured data had performance gains over unimodal models with structured-only data. Future work to implement these models may improve the early detection and treatment of patients with postoperative AKI, which could improve patient outcomes.

Acknowledgments

We would like to thank Madeline Oguss, MS, for administrative assistant during this project.

Author contributions

G.K., M.A., and M.M.C. made substantial contributions to the conception or design of the work. G.K., J.C., A.B.S., K.A.C.,

E.R.G., D.D., A.M., M.A., M.M.C. contributed to analysis or interpretation of the data for the work. All authors contributed to drafting of the work or revising it critically for important intellectual content. All authors provided final approval of the version to be published and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are addressed.

Supplementary material

Supplementary material is available at JAMIA Open online.

Funding

This work was supported by grant NIH/NIDDK R01-DK126933 from the National Institute of Diabetes and Digestive and Kidney Diseases (PI: Koyner and Churpek).

Conflict of interest

Dr Churpek is a named inventor on a patent for a risk stratification algorithm for hospitalized patients (U.S. patent # 11410777). The remaining authors have disclosed that they do not have any potential conflicts of interest.

Data availability

The data utilized in this article cannot be shared publicly due to regulatory and legal restrictions. Our data were obtained from 2 hospital systems after our research protocol was reviewed by IRBs from each hospital, and our data use agreements do not permit sharing due to the granular nature of the data. Interested researchers can contact the corresponding author or Madeline Oguss (mkoguss@medicine.wisc.edu) for specific queries related to data sharing.

References

1. Hoste EA, Kellum JA, Selby NM, et al. Global epidemiology and outcomes of acute kidney injury. *Nat Rev Nephrol.* 2018;14(10):607–625.
2. Lok CE, Austin PC, Wang H, et al. Impact of renal insufficiency on short-and long-term outcomes after cardiac surgery. *Am Heart J.* 2004;148(3):430–438.
3. Machado MN, Nakazone MA, Maia LN. Prognostic value of acute kidney injury after cardiac surgery according to kidney disease: improving global outcomes definition and staging (KDIGO) criteria. *PLoS One.* 2014;9(5):e98028.
4. Chertow GM, Burdick E, Honour M, et al. Acute kidney injury, mortality, length of stay, and costs in hospitalized patients. *J Am Soc Nephrol.* 2005;16(11):3365–3370.
5. Hobson C, Ozrazgat-Baslanti T, Kuxhausen A, et al. Cost and mortality associated with postoperative acute kidney injury. *Ann Surg.* 2015;261(6):1207–1214.
6. Chertow GM, Levy EM, Hammermeister KE, et al. Independent association between acute renal failure and mortality following cardiac surgery. *Am J Med.* 1998;104(4):343–348.
7. Kellum JA, Lameire N, Aspelin P, et al. Kidney disease: improving global outcomes (KDIGO) acute kidney injury work group. KDIGO clinical practice guideline for acute kidney injury. *Kidney Int Suppl.* 2012;2(1):1–138.
8. Bihorac A, Yavas S, Subbiah S, et al. Long-term risk of mortality and acute kidney injury during hospitalization after major surgery. *Ann Surg.* 2009;249(5):851–858.

9. Thakar CV. Perioperative acute kidney injury. *Adv Chronic Kidney Dis.* 2013;20(1):67-75.
10. Klionsky DJ, Abdelmohsen K, Abe A, et al. Guidelines for the use and interpretation of assays for monitoring autophagy (3rd edition). *Autophagy.* 2016;12(1):1-222.
11. Park JT. Postoperative acute kidney injury. *Korean J Anesthesiol.* 2017;70(3):258-266.
12. Simonov M, Ugwuowo U, Moreira E, et al. A simple real-time model for predicting acute kidney injury in hospitalized patients in the US: a descriptive modeling study. *PLoS Med.* 2019;16(7):e1002861.
13. Koyner JL, Adhikari R, Edelson DP, et al. Development of a multi-center ward-based AKI prediction model. *Clin J Am Soc Nephrol.* 2016;11(11):1935-1943.
14. Koyner JL, Carey KA, Edelson DP, et al. The development of a machine learning inpatient acute kidney injury prediction model. *Crit Care Med.* 2018;46(7):1070-1077.
15. Hodgson LE, Sarnowski A, Roderick PJ, et al. Systematic review of prognostic prediction models for acute kidney injury (AKI) in general hospital populations. *BMJ Open.* 2017;7(9):e016591.
16. Tomašev N, Glorot X, Rae JW, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature.* 2019;572(7767):116-119.
17. Lei VJ, Luong T, Shan E, et al. Risk stratification for postoperative acute kidney injury in major noncardiac surgery using preoperative and intraoperative data. *JAMA Netw Open.* 2019;2(12):e1916921.
18. Hodgson LE, Roderick PJ, Venn RM, et al. Correction: the ICE-AKI study: impact analysis of a clinical prediction rule and electronic AKI alert in general medical patients. *PLoS One.* 2018;13(8):e0203183.
19. Saly D, Yang A, Triebwasser C, et al. Approaches to predicting outcomes in patients with acute kidney injury. *PLoS One.* 2017;12(1):e0169305.
20. Flechet M, Güiza F, Schetz M, et al. AKIpredictor, an online prognostic calculator for acute kidney injury in adult critically ill patients: development, validation and comparison to serum neutrophil gelatinase-associated lipocalin. *Intensive Care Med.* 2017;43(6):764-773.
21. Meystre SM, Savova GK, Kipper-Schuler KC, et al. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform.* 2008;17(01):128-144.
22. Ford E, Carroll JA, Smith HE, et al. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *J Am Med Inform Assoc.* 2016;23(5):1007-1015.
23. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? *J Biomed Inform.* 2009;42(5):760-772.
24. Afshar M, Phillips A, Karnik N, et al. Natural language processing and machine learning to identify alcohol misuse from the electronic health record in trauma patients: development and internal validation. *J Am Med Inform Assoc.* 2019;26(3):254-261.
25. Patel TA, Puppala M, Ogunti RO, et al. Correlating mammographic and pathologic findings in clinical decision support using natural language processing and data mining methods. *Cancer.* 2017;123(1):114-121.
26. Castro VM, Dligach D, Finan S, et al. Large-scale identification of patients with cerebral aneurysms using natural language processing. *Neurology.* 2017;88(2):164-168.
27. Kreimeyer K, Foster M, Pandey A, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *J Biomed Inform.* 2017;73:14-29.
28. Jones BE, South BR, Shao Y, et al. Development and validation of a natural language processing tool to identify patients treated for pneumonia across VA emergency departments. *Appl Clin Inform.* 2018;9(1):122-128.
29. Kulshrestha S, Dligach D, Joyce C, et al. Prediction of severe chest injury using natural language processing from the electronic health record. *Injury.* 2021;52(2):205-212.
30. Moons KG, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med.* 2015;162(1):W1-W73.
31. Debray TP, Vergouwe Y, Koffijberg H, et al. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol.* 2015;68(3):279-289.
32. Malhotra R, Kashani KB, Macedo E, et al. A risk prediction score for acute kidney injury in the intensive care unit. *Nephrol Dial Transplant.* 2017;32(5):814-822.
33. Bell S, Dekker FW, Vadeloo T, et al. Risk of postoperative acute kidney injury in patients undergoing orthopaedic surgery—development and validation of a risk score and effect of acute kidney injury on survival: observational cohort study. *BMJ.* 2015;351:h5639.
34. Churpek MM, Carey KA, Edelson DP, et al. Internal and external validation of a machine learning risk score for acute kidney injury. *JAMA Netw Open.* 2020;3(8):e2012892.
35. Park S, Cho H, Park S, et al. Simple postoperative AKI risk (SPARK) classification before noncardiac surgery: a prediction index development study with external validation. *J Am Soc Nephrol.* 2019;30(1):170-181.
36. Sharma B, Dligach D, Swope K, et al. Publicly available machine learning models for identifying opioid misuse from the clinical notes of hospitalized patients. *BMC Med Inform Decis Mak.* 2020;20(1):79.
37. Matt Dinerstein versus Google LLC and The University of Chicago. Class action complaint and demand for jury trial. United States District Court. Accessed September 9, 2019. <https://edelson.com/wp-content/uploads/2016/05/Dinerstein-Google-DKT-001-Complaint.pdf>
38. Meystre SM, Ferrández O, Friedlin FJ, et al. Text de-identification for privacy protection: a study of its impact on clinical text information content. *J Biomed Inform.* 2014;50:142-150.
39. Meystre SM, Friedlin FJ, South BR, et al. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Med Res Methodol.* 2010;10(1):70-76.
40. Ferrández Ó, South BR, Shen S, et al. Generalizability and comparison of automatic clinical text de-identification methods and resources. *AMIA;AMIA Annu Symp Proc.* 2012;2012:199-208.
41. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc.* 2010;17(5):507-513.
42. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* 2004;32(Database issue):D267-D270.
43. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining.* 2016;785-794.
44. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat.* 2001;29(5):1189-1232.
45. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics.* 1988;44(3):837-845.
46. Pencina MJ, D'Agostino RB, D'Agostino RB, et al. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med.* 2008;27(2):157-172.
47. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Soft.* 2010;33(1):1-22.
48. Huang SC, Pareek A, Seyyedi S, et al. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ Digit Med.* 2020;3:136.

49. Liao KP, Cai T, Savova GK, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. *BMJ*. 2015;350:h1885.
50. Halpern Y, Horng S, Choi Y, et al. Electronic medical record phenotyping using the anchor and learn framework. *J Am Med Inform Assoc*. 2016;23(4):731-740.
51. Drew BJ, Harris P, Zegre-Hemsey JK, et al. Insights into the problem of alarm fatigue with physiologic monitor devices: a comprehensive observational study of consecutive intensive care unit patients. *PLoS One*. 2014;9(10):e110274.
52. Bashiri FS, Caskey JR, Mayampurath A, et al. Identifying infected patients using semi-supervised and transfer learning. *J Am Med Inform Assoc*. 2022;29(10):1696-1704.
53. Youssef A, Pencina M, Thakur A, et al. External validation of AI models in health should be replaced with recurring local validation. *Nat Med*. 2023;29(11):2686-2687.
54. Alba AC, Agoritsas T, Walsh M, et al. Discrimination and calibration of clinical prediction models: users' guides to the medical literature. *JAMA*. 2017;318(14):1377-1384.
55. Walsh CG, Sharman K, Hripcsak G. Beyond discrimination: a comparison of calibration methods and clinical usefulness of predictive models of readmission risk. *J Biomed Inform*. 2017;76:9-18.
56. Cahyani DE, Patasik I. Performance comparison of TF-IDF and Word2Vec models for emotion text classification. *Bull EEI*. 2021;10(5):2780-2788.
57. Pruitt P, Naidech A, Van Ornam J, et al. A natural language processing algorithm to extract characteristics of subdural hematoma from head CT reports. *Emerg Radiol*. 2019;26(3):301-306.
58. Koopman B, Zuccon G, Waghlikar A, et al. Automated reconciliation of radiology reports and discharge summaries. *AMIA Annu Symp Proc*. 2015;2015:775-784.
59. Churpek MM, Yuen TC, Winslow C, et al. Multicenter development and validation of a risk stratification tool for ward patients. *Am J Respir Crit Care Med*. 2014;190(6):649-655.
60. Churpek MM, Zdravetz FJ, Winslow C, et al. Incidence and prognostic value of the systemic inflammatory response syndrome and organ dysfunctions in ward patients. *Am J Respir Crit Care Med*. 2015;192(8):958-964.
61. Bernier-Jean A, Beaubien-Souligny W, Goupil R, et al. Diagnosis and outcomes of acute kidney injury using surrogate and imputation methods for missing preadmission creatinine values. *BMC Nephrol*. 2017;18(1):141-149.
62. Shwartz-Ziv R, Armon A. Tabular data: deep learning is not all you need. *Inf Fusion*. 2022;81:84-90.