Development and Validation of Transportable, Clinically Applicable and Scalable Machine Learning Models for Acute Kidney Injury

by

Jie Cao

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2024

Doctoral Committee:

Professor Kayvan Najarian, Co-Chair
Associate Professor Karandeep Singh, Co-Chair
Professor Michael Heung
Associate Professor Arvind Rao
Assistant Professor Xu Shi
Professor Ji Zhu

Jie Cao

caojie@umich.edu

ORCID iD: 0000-0001-9803-3836

Dedication

To my husband Dr. Yan-Cheng Chao and our daughter Yihuan Chao

Acknowledgements

I extend my deepest gratitude to all those whose support and encouragement have played a pivot role in the completion of this dissertation.

First and foremost, my sincere appreciation goes to my advisor, Dr. Karandeep Singh, whose guidance, expertise, and continuous encouragement have been the cornerstone of this research journey. Your steadfast support has extended far beyond the realms of academia, encompassing career development and personal life guidance. Your mentorship has been a guiding light, and I am fortunate to have had your support throughout this challenging yet rewarding endeavor.

I would like to express my gratitude to the members of my dissertation committee, each of whom has contributed significantly to my academic and professional growth. Dr. Kayvan Najarian, your provision of opportunities to engage in signal processing and image processing projects has been instrumental in broadening my research horizons. Dr. Xu Shi, your guidance in EHR analysis classes and ongoing support in career development, especially for female scientists, has been invaluable. Dr. Michael Heung, your valuable insights in the AKI field and continuous support in career development have left a lasting impact. Dr. Ji Zhu, your statistical insights have enriched the methodological rigor of this research. Dr. Arvind Rao, your support throughout my entire PhD journey, from rotations to prelims to the dissertation, has been a source of strength and encouragement.

I also extend my sincere appreciation to Dr. Michael Mathis, the co-PI of my federated learning project, for his invaluable contributions. Dr. Mathis has brought valuable clinical insights to the

table, enriching our discussions and contributing to the depth of our research. He has also provided huge support in obtaining data access, coordinating project meetings, and reviewing data and handling requests. His expertise and commitment to fostering a collaborative and supportive research environment have left a lasting impact on my academic journey. I am truly fortunate to have had the opportunity to work alongside Dr. Michael Mathis, and I am grateful for his contributions, mentorship, and the positive impact he has had on my academic and research endeavors.

I extend my heartfelt gratitude to the esteemed faculty members in the Department of Computational Medicine and Bioinformatics, with special appreciation for the guidance provided by directors Dr. Margit Burmeister and Dr. Maureen Sartor. Their academic insights have been invaluable, shaping crucial decisions related to course plans, lab rotations, and offering support for scholarship and fellowship applications, as well as internship pursuits. Their mentorship has been instrumental in my academic journey, contributing significantly to the depth and breadth of my education.

I would also like to express my thanks to the dedicated staff members in the department, particularly Julia Eussen and Kati Ellis. Their swift and responsive communication has been crucial in navigating various administrative tasks and ensuring a seamless experience for students. Their commitment to supporting student success is evident in their diligence and efficiency, and I am grateful for their tireless efforts behind the scenes.

To the faculty members and staff in the Department of Computational Medicine and Bioinformatics, your collective contributions have played a pivotal role in shaping my academic trajectory. Your support has been a cornerstone of my success, and I am truly appreciative of the collaborative and nurturing environment you have fostered within the department.

I extend my appreciation to my colleagues and fellow researchers who have contributed to stimulating discussions, shared valuable insights, and provided a supportive academic environment. Collaborative endeavors have truly enriched the depth and breadth of this dissertation.

I would like to acknowledge the support and resources provided by the University of Michigan. The research facilities, access to data, and the scholarly community have played a pivotal role in the successful completion of this project.

My deepest gratitude extends to my family, especially my husband, Dr. Yan-Cheng Chao. His unconditional love and unwavering support have been my rock throughout this journey. Our numerous scientific conversations spanning biology, public health, statistics, and beyond have been both enriching and grounding. As we welcomed our daughter, Yihuan Chao, in the last year of my PhD, she brought not only mess but immeasurable joy to our lives. I am grateful for the balance they have brought to my life during this intense academic pursuit. For everyone else in my family, thank you for your encouragement, understanding, and belief in my abilities. Your love and support have been my anchor during the challenging phases of this academic pursuit.

I extend heartfelt gratitude to my circle of friends, who have been a source of steadfast support and encouragement throughout my academic journey. Their camaraderie, understanding, and shared moments of joy and challenge have made this endeavor more fulfilling. Whether it was late-night study sessions, moments of celebration, or simply providing a listening ear during times of stress, my friends have been pillars of strength. Their belief in my abilities and their constant motivation have played a significant role in my perseverance. I am fortunate to have such an incredible group of friends who have added warmth, laughter, and a sense of community

to my life, making this academic pursuit not only a personal achievement but also a shared triumph.

Lastly, I would like to express my gratitude to all those unnamed individuals who, in various ways, have contributed to the realization of this dissertation. Their support, whether direct or indirect, has not gone unnoticed and is deeply appreciated.

Completing this dissertation has been a collective effort, and I am humbled by the support and encouragement I have received along the way. Thank you to everyone who has been a part of this academic journey.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ADQI Acute Dialysis Quality Initiative

AI Artificial Intelligence

AKI Acute Kidney Injury

AKIN Acute Kidney Network

AUROC/AUC Area Under the Receiver Operating Characteristic Curve

CI Confidence Interval

CKD Chronic Kidney Disease

CPB Cardiopulmonary Bypass

CSA-AKI Cardiac Surgery-Associated Acute Kidney Injury

DBSCAN Density-Based Spatial Clustering and Application with Noise

ECE Expected Calibration Error

eGFR Estimated Glomerular Filtration Rate

EHR Electronic Health Record

ESRD End-Stage Renal Disease

FL Federated Learning

FSL Federated Stacked Learning

GBDT Gradient-Boosted Decision Tree

GBM Gradient-Boosted Machine

ICD International Classification of Diseases

ICU Intensive Care Unit

KDIGO Kidney Disease Improving Global Outcomes

MIMIC Medical Information Mart for Intensive Care

MPOG Multicenter Perioperative Outcomes Group

RIFLE Risk, Injury, Failure, Loss, End-stage kidney disease

sCr Serum Creatinine

TRIPOD Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis

UM the University of Michigan

VA Veterans Affairs

UO Urine output

Abstract

Acute kidney injury (AKI), a frequent complication in hospitalized patients, poses significant challenges due to its high incidence, short-term mortality, and substantial economic burden. Current AKI models utilizing electronic health records (EHR) and machine learning (ML) confront limitations in external validation, the exclusion of urine output as a predictor, and a predominant reliance on single-center data. In this dissertation, I present a comprehensive exploration of ML applications for AKI, with a focus on crucial dimensions such as transportability, clinical applicability, and scalability.

In Chapter II, I reproduce and evaluate the transportability of a leading AKI model originally developed by DeepMind for the veterans. Despite the model's high performance in predicting AKI, the predominantly male population on which it is trained have led to questions about its generalizability in other cohorts. I reproduce key aspects of their GBDT model and assess its performance in a sex-balanced patient population at the UM, revealing suboptimal discrimination and calibration in females. A continued training approach at UM partially addresses model differential performance in sex. An exploration of potential reasons for this model discrepancy by sex reveals that it is complex and cannot be simply explained by a low sample size or difference in patient characteristics. This study demonstrates that local fine-tuning may be a promising solution for mitigating sex and gender inequalities in healthcare ML models.

In Chapter III, I investigate the urine output (UO) documentation pattern in the EHR and assess the role of UO as an AKI predictor. Analysis of a five-year inpatient cohort at UM reveals

frequent and diverse UO documentation for non-ICU patients. Despite its value, the inclusion of UO as a predictor minimally improves the ability to predict AKI over a comprehensive model without UO. This study emphasizes the ongoing need for refining UO documentation practices to augment its clinical utility.

In Chapter IV, I introduce a novel Federated Stacked Learning (FSL) framework to enhance the scalability of AKI models in multicenter settings where data sharing may not be permitted. Focusing on predicting cardiac surgery-associated AKI within a national perioperative research network, the study compares the performance of single-center models with both a pooled model and the proposed FSL approach. The single-center models perform worse than the multicenter approaches. The FSL approach demonstrates comparable performance with pooled models, suggesting that it is a practical alternative when patient-level data sharing is not an option. The study underscores the significance of collaborative research networks and illustrates how the size of both the hospital and the network can influence the optimal modeling strategy.

Collectively, this dissertation contributes valuable insights into AKI prediction, advocating for a pragmatic model development approach encompassing transportability, clinical applicability, and scalability. The findings pave the way for future advancements in ML applications for AKI, promoting the development of models that are not only accurate but also accessible, generalizable, and adaptable across diverse healthcare settings.

Chapter 1 Introduction

1.1 Acute Kidney Injury (AKI)

Acute kidney injury (AKI) is a common complication that is associated with various etiologies and pathophysiological changes that can lead to a rapid decline in kidney function, often occurring among hospitalized patients. It is characterized by its high incidence, short-term mortality, and heavy economic burden. AKI complicates 10-20% hospitalized admissions in the United States and worldwide[1–3]. Its occurrence is higher in critically ill patients, about one-third to two-thirds, according to several multinational studies[4–6]. In the COVID-19 pandemic, two studies analyzed hospitalized COVID-19 patients in New York City metropolitan and found about 40% COVID-19 patients developed AKI[7,8]. AKI is also associated with worse health outcomes. The in-hospital mortality for AKI patients is estimated to be over 10%[1,9]; among COVID-19 patients in the Mount Sinai health system in New York, the in-hospital mortality was 50% among AKI patients versus 8% among those without AKI[8]. In addition, patients with AKI have increased risk of developing chronic kidney disease (CKD), end-stage renal disease (ESRD) and cardiovascular disease[10,11]. Health-related quality of life among survivors of AKI patients in the intensive care unit (ICU) is also impaired compared to population norms[12]. In addition to its harm on patient health, AKI also imposes heavy economic burden on the society. AKI is associated with an increase in hospital length of stay and healthcare resource utilization[13]. An estimated 5.4 to 24.0 billion dollars increase in hospitalization cost was attributed to AKI in the U.S in 2012[13]. Although a potentially life-threatening condition, a substantial proportion of cases are considered preventable with early identification, intervention and treatment[14,15].

Therefore, AKI risk prediction models are needed to guide healthcare providers on identifying the right patients to prioritize treatments and allocate resources.

1.2 Stages of AKI

Currently, the international consensus on AKI staging uses the definitions published by the Kidney Disease Improving Global Outcomes (KDIGO) group in 2012[16]. The KDIGO criteria defines AKI stages by changes in serum creatinine (sCr) and/or urine output (UO) (**Table 1.1**). Research has shown that use of the KDIGO definition identified more AKI patients and was more predictive for related in-hospital mortality[17,18] when compared to previously used RIFLE (Risk, Injury, Failure, Loss, End-Stage Kidney Disease)- and Acute Kidney Network (AKIN)-criteria[19,20]. From mild (stage 1) to moderate (stage 2) to severe (stage 3), the definition of the staging system is based on the finding that greater rises in serum creatinine are associated with poorer outcomes, including prolonged hospital stay, increased mortality and higher cost[21]. Choice of therapy, optimal timing for intervening and other patient management decisions may differ by AKI stages. Hence, it is important to evaluate the performance of an AKI risk prediction model for each stage to understand its potential impact on clinical use and patient management.

Table 1.1 KDIGO definition of AKI stages.

AKI Stage	Changes in sCr	Changes in UO
Stage 1	1.5-1.9 times baseline within 7 days or ≥0.3 mg/dl increase within 48 h	<0.5/ml/kg/h for 6-12h
Stage 2	2.0-2.9 times baseline	<0.5/ml/kg/h for ≥12h
Stage 3(D)	3 times baseline or ≥4.0 mg/dl increase or initiation of RRT (Stage 3D)	<0.3/ml/kg/h for ≥ 24h or anuria ≥12h

1.3 Electronic Health record (EHR) and Machine Learning (ML) for AKI Risk Prediction

Risk prediction in healthcare is one of the most exciting frontiers in data science because it enables care to be tailored to a patient's risk. The universal adoption of electronic health records (EHRs) in the United States has expanded the availability of digital clinical data and has made it possible to develop prognostic models and early warning systems using routinely collected data. Such models and early warning systems have the potential to serve as accurate, timely and cost-effective alternatives for future AKI event prediction. An international nephrology group, Acute Dialysis Quality Initiative (ADQI), published an official document recognizing the significance of utilization of EHR and modern modeling tools for AKI risk prediction in this "big data" era, and outlined consensus statements on best approaches[22]. A prototype AKI prediction model should predict risk both for KDIGO Stage 2/3 AKI and patient-centered and clinically important AKI-related outcomes; well-established risk factors, together with novel risk factors identified by machine learning techniques should be used in prediction; and the model should be easily and effectively integrated into EHR and present clinical utility.

A considerable amount of literature has shown promising results in predicting AKI in a more and accurate way by leveraging EHR and machine learning tools (**Table 1.2**)[23–33]. These studies share many similarities. Their models were developed based on large sample sizes and numerous predictors through routinely collected EHR data. Such rich information has also demonstrated the value of machine learning techniques, which are usually "data-hungry"[34] and underperform traditional statistical methods in the absence of sufficient data. In prior work, the AUCs generally range from 0.7 to 0.8 across studies when predicting any AKI (i.e., AKI stage 1+), with the exception of the models developed by the Google DeepMind group, which reported AUCs from 0.863 to 0.921 across the tested algorithms (see Supplementary Table 4)[31]. Although not every

study evaluated model performance for each AKI stage, when examined, the AUCs increase as AKI cases become more severe. Models developed by Koyner et al. demonstrated AUCs of 0.87 for AKI Stage 2+, 0.93 for AKI Stage 3 and 0.96 for AKI requiring dialysis[28]. Demirjian et al. were able to predict AKI following cardiac surgery within 72 hours at AUCs of 0.860 for AKI Stage 2+ and 0.879 for AKI requiring dialysis, when validated externally.

Despite the common characteristics, these studies differ in their choice of modeling strategies (**Figure 1.1**). One group of studies (Cheng et al.[26], Mohamadlou et al.[29], etc.) made the AKI risk prediction completely retrospectively, where AKI event time was first anchored and data within a certain period of time (e.g. 1 day) before AKI onset was excluded from prediction use (**Figure 1.1a**). For example, Mohamadlou et al. trained and tested their GBM model on patient data from Stanford Medical Center and showed AUCs of 0.800, 0.795, 0.761 and 0.728 for predictions made at 0, 12, 24, 48, and 72 hours before onset of AKI stage 2+; when externally validated on the Medical Information Mart for Intensive Care III (MIMIC-III) dataset, the model predicted AKI stage 2+ at AUCs of 0.844, 0.826 and 0.760 for 0, 12 and 24 hours before disease onset[29]. The model performance decreases when data closer to the AKI onset was excluded from the data collection window. Although this strategy is easy to understand in the modeling perspective, it is not clinically applicable as it is a one-time prediction but the prediction time varies for different patients. Additionally, ADQI recommends: "forecasting AKI within a horizon of 48 to 72 hours as it gives providers adequate time to modify practice optimize hemodynamics, and mitigate potential injury without sacrificing predictive power"[22], hence, models that predict AKI at onset, 12 or 24 hours prior do not provide sufficient time for providers to intervene.

Another popular modeling strategy is to make a one-time prediction at a pre-determined time (e.g. admission, after procedure, etc.) for future AKI risk in a given period (**Figure 1.1b**). Cronin

et al. made the prediction at 48 h after admission to identify patient risk for AKI at VA hospitals in the next 7 days[23]. Demirjian et al. at Cleveland Clinic carried out the prediction when first postoperative metabolic panel results are available for cardiac surgery patients for AKI risk within 72 hours and 14 days after the procedure[33]. This is an approach that can be applied clinically because providers can use the model to evaluate patients' AKI risk around admission or procedure time and make subsequent decisions accordingly. However, it may not offer enough granularity of the prediction window and only information at admission or the pre-defined time is used, which may not reveal patients' most recent physiological changes near AKI onset. This may lead to patients' AKI being missed at the defined time when it potentially could have been identified at a later time when the information would still be actionable.

A more clinically applicable approach is to make the prediction dynamically (**Figure 1.1c**). The model should aim to predict the AKI onset in a fixed prediction window (e.g. next 48 hours) and the data collection window should move forward to include most recent data as new predictions are to be made. Koyner et al. at the University of Chicago follows this rationale and developed an GBM model that was able to make the prediction every 12 hours for AKI risk in the next 48 hours[28]. The most successful example of such a modeling strategy was presented in the Google DeepMind research where Tomašev and colleagues designed the model to run every 6 hours to evaluate whether patients will develop AKI in the next 48 hours. Developed and validated using VA data, their models achieved AUCs from 0.863 to 0.921 for all algorithms tested, which outperformed previous studies[31]. This suggests that the dynamic prediction strategy should be the preferred method for future AKI risk prediction model development, both for clinical applicability and optimal model performance.

Table 1.2 Selected publications using EHR and ML for AKI risk prediction.

Publication	Study population	N	AKI definition	Algorithms	AUC	Time of prediction
Cronin et al. 2015[23]	116 VA hospitals Adult inpatients	1,620,898	KDIGO (sCr only) between d2 and d9 of admission Dialysis procedure codes	LR (best) LASSO RF	Stage 1+: 0.746-0.758 Stage 2+: 0.714-0.720 Dialysis: 0.823-0.825	48 h after admission
Koyner et al. 2016[24]	5 hospitals Adult inpatients	202,961	KDIGO (sCr only) within next 24 h	LR	Stage 1+: 0.74 Stage 2+: 0.76 Stage 3: 0.83	Every 12 h
Davis et al. 2017[25]	All VA hospitals	170,675	KDIGO (sCr only) between 48h and d9 of admission	LR LASSO Ridge Elastic-Net RF ANN NB	Stage 1+: 0.69-0.76	48 h after admission
Cheng et al. 2018[26]	1 hospital Adult inpatients	48,955	KDIGO (sCr only) within next 24 h	LR RF AdaBoostM1	Stage 1+: 0.751-0.765 (1-day prior) 0.727-0.733 (2-day prior) 0.691-0.709 (3-day prior)	1-5 days before AKI onset
Huang et al. 2018[27]	NCDR CathPCI registry, 1000+ hospitals PCI patients	947,091	AKIN (sCr only)	LR LASSO XGBoost	Stage 1+: 0.711-0.759	Pre-procedure
Koyner et al. 2018[28]	1 hospital Adult inpatients	121,158	KDIGO (sCr only) within 48 h	GBM	Stage 1+: 0.73 Stage 2+: 0.87 Stage 3+: 0.93 Dialysis: 0.96	Every 12 h
Mohamadlou et al. 2018[29]	1 hospital Adult inpatients	19,737	NHS England AKI algorithm/ KDIGO (sCr only)	GBM	Stage 2+: - Internal validation 0.872 (onset) 0.800 (12 h prior) 0.795 (24 h prior) 0.761 (48 h prior) 0.728 (72 h prior) - External validation 0.844 (onset) 0.826 (12 h prior) 0.760 (24 h prior)	0/12/24/48/72 h before AKI onset
He et al. 2019[30]	1 hospital	76,957	KDIGO (sCr only)	LR NB Bayes Net RF Emsemble (LR & RF, voting)	Stage 1+: 0.687-0.744 (1 day prior) 0.676-0.734 (at admission, any time AKI) 0.720-0.764(at admission, AKI within N days, N = 1, 2, 3, 7, 15, 30) 0.600-0.764 (daily after admission)	1 day before AKI onset/at admission/daily after admission
Tomašev et al. 2019[31]	114 VA centers Adult inpatients	703,782	KDIGO (sCr only) within next 48 h	RNN LR RF GBM MLP	Stage 1+: 0.863-0.921 Stage 2+: 0.870-0.957 Stage 3+: 0.930-0.980	Every 6 h
Zimmerman et al. 2019[32]	MIMIC-III Adult ICU stays	23,950	KDIGO (sCr only) within 72 h	LR RF MLP	Stage 1+: 0.772-0.796	24 h following ICU admission
Demirjian et al. 2022[33]	1 hospital Cardiac surgery adult patients	58,526	KDIGO (sCr only) within 72 h and 14 d Dialysis information from registry data	LR	Stage 2+: - Internal validation 0.876 (72 h) 0.854 (14 d) - External validation 0.860 (72 h) 0.842 (14 d) Dialysis: - Internal validation 0.916 (72 h) 0.900 (14 d) - External validation 0.879 (72 h) 0.873 (14 d)	When first postoperative metabolic panel results are available

LR: logistic regression
LASSO: least absolute shrinkage and selection operator
XGBoost: Extreme gradient boosting
RF: random forest
NB: naïve Bayes
NCDR: National Cardiovascular Data Registry
PCI: percutaneous coronary intervention
AKIN: Acute Kidney Injury Network

GBM: gradient boosting machine
NHS: National Health Service
RNN: recurrent neural network
MLP: multilayer perceptron
MIMIC-III: Medical Information Mart for Intensive Care III

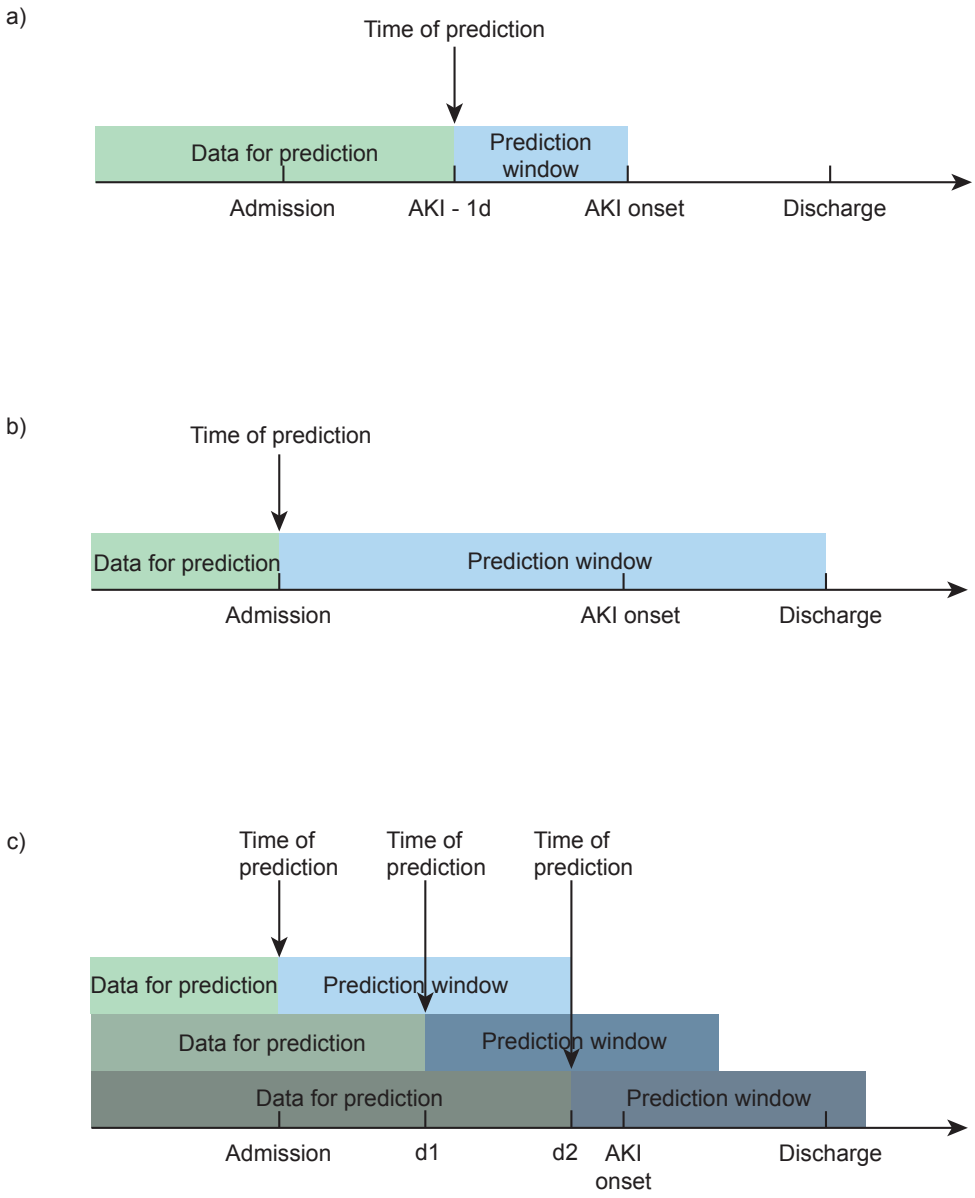


Figure 1.1 Different modeling strategies for AKI risk prediction.

Visual representation of modeling strategies used by published studies. a) First anchor the AKI event time and exclude data within a certain period of time (e.g. 1 day) before AKI onset. b) Make a one-time prediction at a pre-determined time (e.g. admission) for future AKI risk in a given period (e.g. from admission to discharge). c) Dynamic prediction. The model predicts the AKI onset in a fixed prediction window (e.g. next 24 hours) and the data collection window moves forward to include most recent data as new predictions are to be made. This figure is recreated and adapted from Figure 1 presented in He et al[30].

1.4 Gaps in Existing Work

AKI risk prediction models are needed to improve patient outcomes through better targeting of interventions. An ideal AKI model should run dynamically to meet the clinical need while also having sufficiently high ability to discriminate between high- and low-risk patients. The model performance for different stages of AKI should also be considered to understand a model's potential impact on clinical use and patient management. Despite recent improvements in model performance in the AKI domain, gaps remain in understanding and enhancing the transportability, clinical applicability, and scalability of AKI models.

Firstly, the transportability of most developed AKI models is unclear. Model transportability refers to the ability of the model to maintain its predictive performance when applied to a different but related population or in a different context than the one it was originally developed for. In other words, it assesses the generalizability of the developed model across different datasets or settings. As shown in **Table 1.2**, most studies used patients from a single hospital and reported results from internal validation only. When models were externally validated (i.e., applied to a cohort that is different from the training cohort), the performance was often lower, suggestive of overfitting and raising concerns about model generalizability across health systems. DeepMind's AKI model, although presenting the best AUC up to date, was trained based on VA population that is 94% male. Concerns have arisen about the generalizability of this model to females[35], and to non-VA contexts where practice patterns may differ.

Secondly, most AKI models exclude urine output as a predictor for AKI model in spite of it being an important AKI biomarker. This is primarily because it has historically been documented sparsely and inconsistently, especially in patients hospitalized in non-ICU settings[26,36–38]. When

AKI models are trained exclusively for ICU patients, urine output is more closely monitored, and thus urine output has historically been considered more useful. Zimmerman et al. used hourly rate of urine output during the first day of ICU admission as a predictor, but found it not significantly associated with increase in sCr or AKI event in their regression analyses[32]. As healthcare systems increasingly adopt EHR-based workflows, there is potential for urine output to emerge as a more valuable digital marker for AKI, with potential applicability across both ICU and non-ICU settings. Thus, the clinical utility of urine output in AKI risk prediction models remains an understudied but promising area worth investigating in the era of EHR-based documentation.

Lastly, most published studies describe locally developed models (i.e. use data from one single hospital) rather than multicenter modeling efforts, as shown in **Table 1.2**. This results from restricted data sharing for privacy and lack of model generalizability across centers. While the single-center modeling strategy may suffice for centers with a sufficiently large patient population, it may lead to inequity issues among smaller centers that do not have a cohort size needed to develop a reliable AKI risk prediction model. While it is possible that the inclusion of multiple centers will improve model performance at smaller centers, this has not been established. There are two common approaches to multicenter modeling, data pooling and federated learning. In the data pooling approach, a centralized cohort of data is curated by pooling data from multiple centers. While this approach effectively boosts model performance and generalizability, it requires a substantial amount of work, and the risk of re-identifying patient information remains[39] and is against patient privacy expectations[40,41]. Federated learning (FL) is an emerging approach to this multicenter modeling problem that aims to enhance privacy and improve model scalability. FL is a decentralized approach to train ML models, by removing

the need to pool data into a central repository while allowing the development of a model to be informed by the information shared from participating centers. Recent studies have shown its application in COVID-19-associated AKI[42] and AKI in the ICU[43]. It is worth noting that the FL approach that is most widely used typically requires numerous rounds of information exchange between participating centers and a central server and thus requires substantial infrastructure to perform.

1.5 Overview

Development and validation of clinically applicable risk prediction models to identify high-risk AKI patients is a compelling area in this big data era. Identifying gaps and selecting the modeling strategy accordingly based on existing work is essential to make AKI prediction one step closer to real-world model implementation for clinical AKI care. In this dissertation, I mainly focus on addressing three gaps identified in the field of AKI risk prediction modeling. In Chapter II, I evaluate the transportability of a reproduced version of DeepMind's GBDT AKI model to a more sex-balanced population at the University of Michigan (UM) and update the model to correct the model performance discrepancy in sex. In Chapter III, I describe the pattern of urine output documentation in the UM EHR system and assess the clinical applicability of urine output as a predictor in AKI models. In Chapter IV, I propose a new federated learning framework, federated stacked learning (FSL), to improve AKI model scalability. I compare the FSL framework against the pooled approach and single-center approach, drawing on data from a large national perioperative research and quality improvement network using the prediction of cardiac surgery associated-acute kidney injury as a prototypical clinical scenario. Finally, Chapter V concludes this dissertation by summarizing the main findings, discussing the

implication of these studies, and proposing future directions towards model improvement and clinical implementation.

Chapter 2 Generalizability of An Acute Kidney Injury Prediction Model Across Health Systems

2.1 Background

Delays in the identification of acute kidney injury (AKI) in hospitalized patients are a major barrier to the development of effective interventions for treatment[44]. By the time changes in typical kidney function biomarkers—serum creatinine (sCr) and blood urea nitrogen—are detected, damage that is not readily reversed is often already established. This is underscored by recent evidence that automated alerts generated upon AKI onset appear to be ineffective in changing the trajectory of AKI[45]. This has led to multiple efforts to develop early warning system scores that predict the onset of AKI with sufficient lead times to support potential intervention The most promising of these efforts was recent work by Tomašev and colleagues from DeepMind describing models for the continuous prediction of AKI that outperformed previously published models[31]. Developed and validated using data from 703,782 US veterans, the primary recurrent neural network model described in the paper achieved an area under the receiver operating characteristic curve (AUC) of 92.1% when predicting AKI in the next 48 h. This study was notable for several reasons, including its large sample size, high AUC and a longer lead time, all of which made this model a clear outlier as compared with previous studies. Despite its promise, the model has not been implemented within the Veterans Affairs (VA) health system. In this respect, this model represents an example of the 'artificial intelligence (AI) chasm', a term used to describe high-performing AI models that fail to reach the bedside due to challenges involved in real-world implementation[46]. The model described in this study is also not

publicly available, meaning that it cannot be readily reproduced and evaluated in other clinical settings despite knowledge of the underlying methods and software[47,48]. This lack of computational reproducibility among complex AI models in healthcare is a well recognized barrier to sustaining progress in clinical AI applications[49–53]. Because the model was developed in a veteran population that is 94% male, concerns have also arisen about the generalizability of this model to females[35], and to non-VA contexts where practice patterns may differ. Recent work in medical imaging demonstrates that models trained in predominantly male populations fail to perform well in females[54]. This was suggested by the DeepMind study, where a lower AKI episode-level sensitivity was observed in females as compared with males (44.8% versus 56.0%, respectively).

To address these concerns, we sought to evaluate the generalizability of an AKI model trained at the VA in patient populations with a more balanced sex composition. Due to computational constraints within the VA computing environment, we aimed to approximate the gradient-boosted decision tree (GBDT) model reported in the DeepMind study rather than the primary recurrent neural network, with the rationale that even the GBDT outperformed previous models with an AUC of 88.9%. Drawing on electronic health record (EHR) data from 278,813 US veterans, we approximated aspects of this model, including data preprocessing, feature selection, transformation of hospitalization data to 6 h person–period intervals, and outcome definitions. Areas where our approach differed from the original study are detailed in Differences between our model and the DeepMind GBDT model. We further assessed the model's generalizability in a large academic center using sex-balanced data from another 165,359 hospitalizations. Finding that the model performs worse in females, we evaluated an approach to updating the model to

correct for this discrepancy. Both our reconstructed model and the corrected model are publicly available[55].

2.2 Methods

2.2.1 Study Cohorts

Our study used data from two cohorts: a national VA cohort drawing on data from 118 VA hospitals, and a UM cohort. We have complied with all relevant ethical regulations. The study was approved by the institutional review boards of the VA Ann Arbor Healthcare System and the UM Medical School, and the need for informed consent was waived.

2.2.2 National VA Cohort

We collected clinical data on all adult patients admitted at a VA hospital between 1 October 2016 and 30 September 2017. Starting with a cohort of 280,985 US veterans hospitalized between 1 October 2016 and 30 September 2017, we excluded patients who did not have creatinine checked at baseline or during their stay (defined in Predictor Variables), had pre-existing end stage renal disease or had a baseline creatinine of >4.0 mg/dL (because they may have had pre-existing AKI stage 3). Only the first hospitalization for each patient was included in the analysis. The final VA cohort consisted of 278,813 patients, which was randomly divided into training (64%), validation (16%) and test (20%) sets at the patient level.

2.2.3 UM Cohort

We collected clinical data from all adult patients admitted to UM from 1 January 2016 to 31 December 2020. The same exclusion criteria as used in the VA cohort were applied to the UM cohort, though all hospitalizations (not only the first) were included. The final UM cohort

consisted of 165,359 hospitalizations. Anticipating the need for updating of the VA model at UM, we randomly selected 60% of hospitalizations (sampled at the patient level) for the test set, and set aside the remaining 40% for model updating, which was divided equally into a training (20%) and a validation (20%) set.

2.2.4 Predictor Variables

We collected both fixed predictors (that is, baseline variables) and time-varying predictors (that is, variables measured on a repeated basis during a hospitalization) in both cohorts. Fixed predictors included age, height, weight, body mass index, 17 comorbidities, admission to a surgical service, intensive care unit status and baseline sCr, all of which were captured at the time of admission. Age was top-coded at 89 yr. Baseline height and weight were calculated as the mean value from the 3 yr preceding admission for VA patients, and the most recent value within the past year for UM patients. If no recent value was identified for UM patients, the first inpatient measurement was used. Height and weight measurements were converted into inches and pounds, respectively, and extreme values were removed. Baseline body mass index was calculated using the baseline height and baseline weight. Comorbidities were calculated with the Charlson comorbidity index using data from 1 yr before admission for VA patients and from the current encounter for UM patents[56]. Baseline sCr was determined by the following order of preference: (1) mean outpatient sCr between 7 and 365 d before admission and (2) within 7 d before admission, and (3) first inpatient sCr test for VA patients or first documented sCr value within 24 h of admission for UM patients.

Time-varying predictors consisted of inpatient vital signs, laboratory test results and administration of medications. Twenty-six laboratory testing components (serum albumin, alkaline phosphatase, alanine aminotransferase, aspartate transaminase, total and direct bilirubin,

blood urea nitrogen, serum calcium, carbon dioxide, serum chloride, serum glucose, high-density lipoprotein cholesterol, hematocrit, hemoglobin A1c, hemoglobin, international normalized ratio, low-density lipoprotein cholesterol, microalbumin-to-creatinine ratio, serum phosphate, platelet count, serum potassium, sCr, serum sodium, total cholesterol, triglyceride and total white blood cell count) were selected due to universal use across different health systems. Eight vital signs (inpatient weight, systolic blood pressure, diastolic blood pressure, respiratory rate, temperature, pulse, blood oxygen level and central venous pressure) were pulled regardless of the frequency of measurement. Administration of medications was examined for 11 drug classes (aminoglycosides, sympathomimetics, beta blockers, alpha blockers, calcium channel blockers, antilipemic agents, loop diuretics, angiotensin-converting enzyme inhibitors, angiotensin II inhibitors, non-ionic contrast media and non-salicylate antirheumatic non-steroidal anti-inflammatory drugs) as opposed to individual medications.

2.2.5 Data Preprocessing and Feature Engineering

Physiologically infeasible values (for example, due to a laboratory error) were excluded. Microalbumin-to-creatinine ratios were set to 0 when values were reported only in a text field based on the observation that the text fields reported such values as being below the detectable range. Data elements were time-stamped using the time when values became available to the EHR (that is, the observation time). The description of variables, the associated units and valid ranges are shown in **Supplemental Table 2.1**.

After extracting the fixed and time-varying predictors, we captured patient states at 6 h intervals beginning with the time of admission for each patient in a manner similar to that of the DeepMind AKI study. Patient states were captured up until the final creatinine value, discharge or death, and truncated at 7 d of hospitalization due to computational constraints. For each 6 h

interval, summary statistics (length, minimum, mean, median, maximum) were calculated for the preceding 48 h divided into 6 h windows for vital signs and laboratory test results. Using these summary statistics, additional variables were created based on clinical relevance: the ratio of the most recent maximum sCr to baseline sCr, the difference between the most recent maximum sCr and baseline sCr, and the ratio of most recent maximum blood urea nitrogen to most recent maximum sCr. These three sCr-based predictors, time (h) from admission and current AKI stage, plus the summary statistics of temporal predictors in the given windowed lookback period, together with the fixed predictors, were used as the full set of 1,467 predictors. The preparation of predictors at the VA and UM followed the same procedures, with the only exception for central venous pressure predictors. Central venous pressure information is not available at UM. Hence, central venous pressure-based predictors were manually added to the predictor set and were all set to missing. The number of administered medications was calculated for the preceding week (7 d) divided into 24 h sliding windows. More details can be found in **Supplemental Table 2.1**. A visual representation of the feature engineering process is shown in **Figure 2.1**.

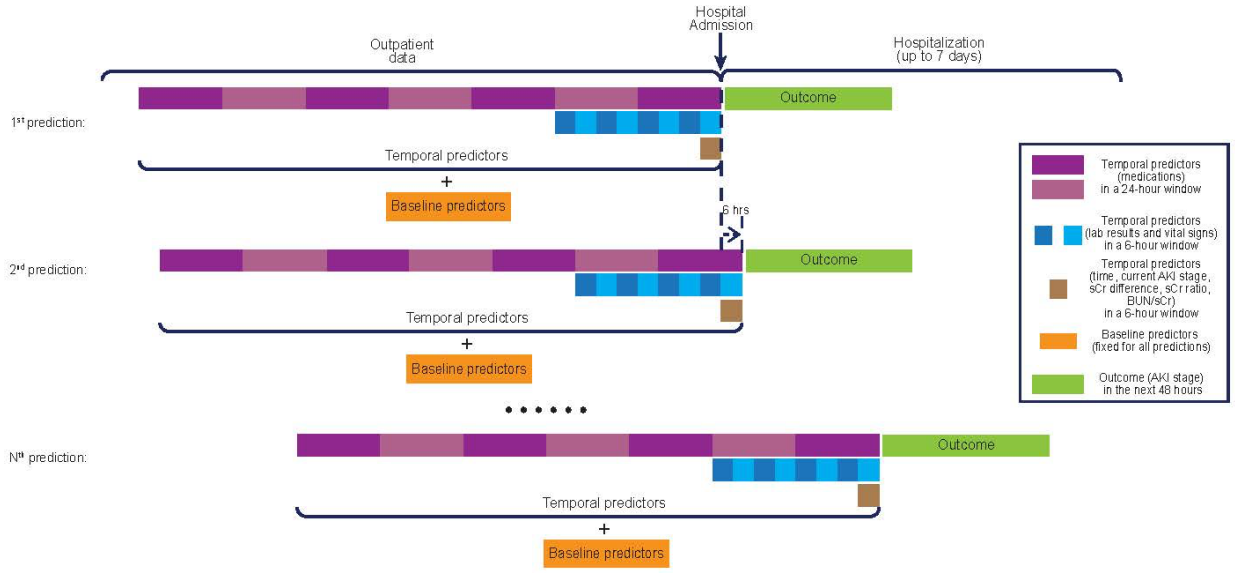


Figure 2.1 Representation of the EHR data for the proposed model.

Representation of the EHR data for our model. EHR data available for each hospitalization were prepared to make an AKI risk prediction every 6 h from the time of hospital admission and up to 7 d from admission. For each prediction, baseline predictors and temporal predictors (medications; laboratory results and vital signs; time, current AKI stage, sCr difference from baseline sCr, sCr ratio to baseline sCr, and blood urea nitrogen (BUN) to sCr ratio) were prepared and used together to estimate the outcome (AKI stage) in the next 48 h.

2.2.6 Outcome Definition

AKI was defined and staged for severity according to the Kidney Disease: Improving Global Outcomes international guidelines[16]. The outcome was calculated on a rolling basis at 6 h intervals by comparing the maximum sCr value in the 48 h prediction window with the baseline sCr. Stage 1 AKI was defined as a sCr level increase of $\geqslant$0.3 mg/dL, but less than twice the baseline sCr or an increase of 1.5 times baseline. Stage 2 AKI reflected an increase of two to three times the baseline, and stage 3 AKI was an sCr level increase greater than three times baseline or an increase to $\geqslant$4.0 mg/dL. Stage 3D was determined based on the need for dialysis, where the time of first dialysis was determined based on diagnosis, procedure and clinic stop codes during hospitalization at the VA, and using procedure codes at UM. Thus, for every 6 h interval in which patient states were captured, outcomes were defined as one of five classes

based on the 48 h prediction window: no AKI, AKI stage 1, AKI stage 2, AKI stage 3 or AKI stage 3D. While models were trained using this multinomial outcome, results reported by AKI stages were grouped according to level of severity. For example, AKI stage 1+ is used to refer to any AKI stage, and AKI stage 2+ refers to AKI stage 2 or greater (including stages 3 and 3D).

2.2.7 Model Development

In the original study, Tomašev et al. selected a 'simple' recurrent neural network as their primary model, which achieved an AUC of 92.1% in their test set. Tomašev and colleagues also evaluated 11 other neural network architectures, two tree ensembles and a logistic regression model, all of which performed better than previous studies. For example, the GBDT achieved an AUC of 88.9%, which still outperforms previously published models. While Tomašev et al. had access to a deidentified dataset, which allowed them to use DeepMind's computing infrastructure to train deep learning models, our team was restricted to using the VA's VINCI platform, which lacks the graphical processing units needed to efficiently train deep learning models. Thus, we opted to approximate the GBDT model from the Tomašev study.

The GBDT model was trained on the VA training set to predict AKI stage in the next 48 h as a multinomial outcome (that is, 'no AKI', 'AKI stage 1', 'AKI stage 2', 'AKI stage 3', 'AKI stage 3D') using 1,467 predictors at each 6 h step with a maximum of 1,000 trees and a maximum depth of 5. The VA validation set was used to determine the need for early stopping based on an improvement in log loss lower than 0.0005 on five consecutive rounds based on a moving average calculated after every ten trees. The categorical predictors were reordered by the mean response of each level for more efficient training. Internally, a separate one-versus-all tree was trained for each outcome class. Using a learning rate of 0.1, the trained VA AKI model stopped training at 160 trees (internally represented as 160 trees per class). Lower learning rates (0.01

and 0.001) produced more trees (because more trees were needed to achieve convergence) but achieved similar results (that is, AUC), so will not be presented here.

2.2.8 Differences between Our Model and the DeepMind GBDT Model

There are differences worth highlighting between our modelling process and that in the work of Tomašev et al. First, whereas Tomašev et al. trained separate models for each stage of AKI as a binary outcome (for example, no AKI versus AKI 1+, up to AKI stage 1 versus 2+, and so on), we modelled the outcome as a multinomial outcome with five possible outcome states. Patients who developed stage 1 AKI were not excluded from model training because they were still at risk for developing AKI stages 2, 3 and 3D. The inclusion of post-AKI states also allowed the model to account for AKI recovery, which is important for generating well calibrated outcome probabilities. Indeed, whereas the DeepMind recurrent neural network model required recalibration in the original paper, our model's probabilities were relatively well calibrated without recalibration (**Supplemental Figure 2.1**). Although we used a multinomial outcome, the H2O implementation of GBDT implements multinomial outcomes as a series of one-versus-all trees. Thus, internally, both our approach and the DeepMind GBDT considered the risk of each stage of AKI as a binary outcome.

Second, whereas Tomašev et al. used 3,599 engineered features in their final model, our model used 1,467 features. These differences can be attributed to which features were included as well as how they were represented. Our model lacked billing and procedure codes due to computational constraints. While the DeepMind model represented each feature in each window using a series of histogram bins, we represented each feature using summary statistics. Because the VA dataset is quite sparse (any given 6 h window has very few observations), we do not believe that this approximation substantially harms the performance.

2.2.9 Model Evaluation

The performance of the GBDT model was evaluated in both the VA test set and the UM test set. The model discrimination was assessed by using the AUC. The AUC was reported both as a multinomial outcome using the method of Hand and Till[57], and as a series of binary AUCs where at-risk individuals were evaluated on their risk of progression to a higher AKI stage. For example, patients without any AKI to date were evaluated on their risk of developing any AKI (that is, stage 1 or greater), patients with no AKI or AKI stage 1 were evaluated on their risk of developing AKI stage 2 or greater, and so on. The 95% CIs were generated using 200 bootstrap resamples for the multiclass AUCs and DeLong's method for binary AUCs[58]. Our primary finding is the AUC calculated when treating each prediction independently in its ability to predict AKI in the next 48 h, which is closely comparable to the way AUC was calculated in the DeepMind study. We evaluated model calibration by comparing deciles (ten bins) of predicted probabilities with observed risk. We also used expected calibration error (ECE) as a scalar metric to measure model calibration. ECE is defined as the weighted average over the absolute difference between observed risks and predicted probabilities, formulated as

$$ECE = \sum_{m=1}^{10} \frac{n_m}{N} |R_{observed} - P_{predicted}|$$

where m represents the number of the bin, nm, $R_{observed}$ and $P_{predicted}$ represent total number of predictions, observed risk and predicted probability within each bin, respectively, and N represents the total number of predictions in the group (all/female/male) examined.
Because the make-up of the VA population is different from other hospitals (for example, 94% male), we examined model performance across sexes and racial groups.

2.2.10 Updating the Model with UM Data

Given previous concerns that models trained at the VA may not generalize to broader populations, we updated the VA model using a UM training/validation set that was set aside before model evaluation (as described in Study Cohorts). Starting with the original 160-tree GBDT model trained only in the VA population, we continued to train it using the UM training set, with a similar early stopping strategy based on a lack of log loss improvement of 0.0001 after five consecutive rounds in the UM validation set. This updated model (which we refer to as the 'extended model' to indicate that it includes a portion of the original VA model) added 10 more trees on top of the original 160 trees, resulting in a total of 170 trees. The updated model was then evaluated in both the UM and VA test sets.

2.2.11 Matching UM Females to VA Female Patients

To address whether these differences in patient characteristics could explain this performance discrepancy, we matched female patients in the UM test with females at the VA. We used the tilted bootstrap method (as implemented in the tboot R package[59]) to match UM test set female patients to VA female patients by mean age, proportion of white patients, proportion of baseline chronic kidney disease and proportion of baseline congestive heart failure. Because the method involves bootstrapping, patients may be included in the matched cohort multiple times. Matching at UM was performed using summary statistics from the VA. We could not perform 1:1 patient-level matching because the VA and UM datasets could not be analyzed in the same computing environment due to security restrictions.

2.2.12 Variable Importance

We assessed variable importance using each variable's squared influence within the GBDT algorithm aggregated over the tree ensemble[60]. Variable importances for the original and extended models are provided in **Supplemental Figure 2.2**.

2.2.13 Software

SAS 9.3 was used to pull data from the VA, and R with dbplyr 2.1.1 was used to pull data from UM. All data processing and analyses were performed using R 4.0.5 at the VA and R 3.6.1 at UM[61]. Transformation of time-series data was performed using the Grammar of Prediction (gpmodels) R package[62]. H2O version 3.32.1.3 was used to fit the GBDT model[63]. We did not use XGBoost (which was used in the DeepMind study) because, while H2O and XGBoost achieve comparable performance for their respective GBDT implementations, H2O's implementation is more memory efficient[64], which was a requirement when using the VA's VINCI computing platform.

2.3 Results

2.3.1 Cohort Characteristics

We identified 278,813 VA hospitalizations (from 118 VA hospitals) and 165,359 University of Michigan (UM) hospitalizations meeting inclusion and exclusion criteria. Only the first hospitalization was included for VA patients, whereas all eligible hospitalizations were included for 97,506 UM patients. As compared with UM, patients with VA hospitalizations were more likely to be male (94% versus 50%), older (mean 69 versus 57) and Black (20% versus 11%) and to have diabetes (36% versus 29%). On the other hand, UM patients were more likely to have normal baseline kidney function (baseline estimated glomerular filtration rate (eGFR) ⩾ 60 ml min−1/1.73 m2, 81% versus 73%) and a longer length of stay (mean 6.6 versus 5.3 d), leading to

more 6 h periods per patient (18 versus 15) calculated over a maximum of 7 d of hospitalization (**Table 2.1**).

Table 2.1 Characteristics of the VA and UM cohorts.

Cohort		**VA**			**UM**	
	Training	**Validation**	**Test**	**Training**	**Validation**	**Test**
Characteristic	**N = 178,453**	**N = 44,614**	**N = 55,746**	**N = 33,077 (19,501 patients)**	**N = 33,034 (19,501 patients)**	**N = 99,248 (58,504 patients)**
Age (years)	68.8 (13.1)	68.9 (13.1)	68.8 (13.1)	57.3 (18.2)	57.2 (18.2)	56.8 (18.2)
Sex						
Female	10,083 (5.7%)	2,557 (5.7%)	3,119 (5.6%)	16,381 (49.5%)	16,605 (50.3%)	49,280 (49.7%)
Male	168,370 (94.3%)	42,057 (94.3%)	52,627 (94.4%)	16,696 (50.5%)	16,429 (49.7%)	49,968 (50.3%)
Race						
African American	35,171 (19.7%)	8,791 (19.7%)	10,990 (19.7%)	3,361 (10.2%)	3,794 (11.5%)	11,036 (11.1%)
Caucasian	120,962 (67.8%)	30,308 (67.9%)	37,835 (67.9%)	27,594 (83.4%)	27,277 (82.6%)	82,164 (82.8%)
Other	15,038 (8.4%)	3,742 (8.4%)	4,696 (8.4%)	1,733 (5.2%)	1,621 (4.9%)	4,899 (4.9%)
Unknown	7,282 (4.1%)	1,773 (4.0%)	2,225 (4.0%)	389 (1.2%)	342 (1.0%)	1,149 (1.2%)
Baseline BMI	29.6 (6.7)	29.6 (6.6)	29.5 (6.6)	28.8 (6.7)	28.8 (6.9)	28.9 (6.8)
Unknown	12,519 (7.0%)	3,033 (6.8%)	3,808 (6.8%)	1,898 (5.7%)	1,944 (5.9%)	5,921 (6.0%)
Baseline serum creatinine (mg/dL)	1.1 (0.4)	1.1 (0.4)	1.1 (0.4)	1.0 (0.4)	1.0 (0.4)	1.0 (0.4)
Baseline eGFR* (mL/min/1.73 m2)						
≥ 60	131,108 (73.5%)	32,836 (73.6%)	40,874 (73.3%)	27,016 (81.7%)	26,813 (81.2%)	80,530 (81.1%)
45-59	27,100 (15.2%)	6,761 (15.2%)	8,651 (15.5%)	3,192 (9.7%)	3,339 (10.1%)	9,894 (10.0%)
30-44	14,686 (8.2%)	3,631 (8.1%)	4,481 (8.0%)	1,957 (5.9%)	1,945 (5.9%)	5,988 (6.0%)
15-29	5,470 (3.1%)	1,365 (3.1%)	1,706 (3.1%)	872 (2.6%)	905 (2.7%)	2,727 (2.7%)
< 15	89 (0.0%)	21 (0.0%)	34 (0.1%)	40 (0.1%)	32 (0.1%)	109 (0.1%)
Baseline diabetes	64,844 (36.3%)	16,174 (36.3%)	20,143 (36.1%)	9,707 (29.3%)	9,558 (28.9%)	28,922 (29.1%)
Baseline congestive heart	25,905 (14.5%)	6,443 (14.4%)	8,019 (14.4%)	8,396 (25.4%)	8,747 (26.5%)	26,180 (26.4%)

failure						
Baseline liver disease	16,672 (9.3%)	4,214 (9.4%)	5,349 (9.6%)	6,469 (19.6%)	6,541 (19.8%)	19,606 (19.8%)
Surgical service	41,673 (23.4%)	10,367 (23.2%)	13,035 (23.4%)	5,773 (17.5%)	5,666 (17.2%)	16,815 (16.9%)
Admitted to ICU	13,075 (7.3%)	3,346 (7.5%)	4,180 (7.5%)	2,753 (8.3%)	2,917 (8.8%)	8,167 (8.2%)
Length of stay (days)	5.3 (12.2)	5.3 (12.5)	5.4 (14.2)	6.6 (7.8)	6.7 (8.6)	6.6 (8.3)
Number of 6-hour windows**	15.1 (9.0)	15.1 (8.9)	15.1 (9.0)	18.4 (8.6)	18.3 (8.7)	18.3 (8.7)

Statistics presented: mean (SD); n (%)

* Calculated based on CKD-EPI Creatinine Equation (2021)

** Calculated based on a maximum of 7-day hospitalization stay

Table 2.2 AKI incidence in the VA and UM cohorts, by acute kidney injury stage, by sex.

Outcome		**VA (all)**			**UM (all)**		
		All	**Female**	**Male**	**All**	**Female**	**Male**
Hospitalization level							
	AKI-1+	10.39% (25,978/250,103)	6.04% (890/14,741)	10.66% (25,088/235,362)	16.10% (26,529/164,774)	13.79% (11,307/82,009)	18.39% (15,222/82,765)
	AKI-2+	1.52% (4,127/271,850)	1.11% (171/15,463)	1.54% (3,956/256,387)	3.93% (6,494/165,192)	3.55% (2,917/82,184)	4.31% (3,577/83,008)
	AKI-3+	0.82% (2,244/275,030)	0.60% (94/15,613)	0.83% (2,150/259,417)	1.76% (2,914/165,276)	1.46% (1,198/82,227)	2.07% (1,716/83,049)
	AKI-3D	0.31% (278,799)	0.13% (21/15,758)	0.33% (856/263,041)	0.23% (388/165,338)	0.18% (152/82,260)	0.28% (236/83,078)
Multiclass predictions, every 6 hours		N = 4,213,375	N = 215, 923	N = 3,997,452	N = 3,033,165	N = 1,478,583	N = 1,554,582
	No AKI	3,277,669 (77.8)	185,228 (85.8)	3,092,441 (77.4)	2,723,535 (89.8)	1,350,169 (91.3)	1,373,366 (88.3)
	AKI-1	737,264 (17.5)	22,690 (10.5)	714,574 (17.9)	231,626 (7.6)	94,951 (6.4)	136,675 (8.8)
	AKI-2	87,500 (2.1)	3,801 (1.8)	83,699 (2.1)	39,700 (1.3)	18,614 (1.3)	21,086 (1.4)
	AKI-3	93,376 (2.2)	3,757 (1.7)	89,619 (2.2)	30,444 (1.0)	11,912 (0.8)	18,532 (1.2)
	AKI-3D	17,566 (0.4)	447 (0.2)	17,119 (0.4)	7,860 (0.3)	2,937 (0.2)	4,923 (0.3)

Binary predictions for each stage among patients who have not reached that stage, every 6 hours						
AKI-1+	3.55% (120,255/3,386,277)	2.22% (4,191/189,054)	3.63% (116,064/3,197,223)	3.76% (97,197/2,583,269)	3.18% (40,973/1,288,611)	4.34% (56,224/1,294,658)
AKI-2+	0.41% (16,323/4,029,154)	0.35% (722/208,549)	0.41% (15,601/3,820,605)	0.87% (25,286/2,921,450)	0.76% (10,922/1,428,256)	0.96% (14,364/1,493,194)
AKI-3+	0.21% (8,700/4,109,724)	0.19% (404/212,042)	0.21% (8,296/3,897,682)	0.39% (11,767/2,983,230)	0.31% (4,566/1,457,716)	0.47% (7,201/1,525,514)
AKI-3D	0.12% (5,116/4,200,925)	0.06% (140/215,616)	0.12% (4,976/3,985,309)	0.08% (2,367/3,027,672)	0.06% (925/1,475,646)	0.09% (1,442/1,549,659)

The incidence of AKI differed in the two cohorts and in individual sex groups (**Table 2.2**). Among patients without AKI on presentation to the hospital, 10.4% (25,978/250,103) developed AKI during their hospitalization at the VA, whereas at UM AKI occurred in 16.1% (26,529/164,774) of hospitalizations. Male patients were more likely to experience AKI than females in both cohorts (10.6% versus 5.6% at the VA, and 18.4% versus 13.8% at UM). While the model was trained on all windows (including those in which AKI had already occurred), the model was evaluated on only those windows in which patients had not yet experienced the outcome. At the 6 h window level, the incidence of new-onset AKI in the test set was 3.53% at the VA and 3.76% at UM (**Table 2.3**).

Table 2.3 Model performance (AUC) of the original VA model at VA and UM, by outcome stage, by sex.

Outcome	**VA Test AUC (95% CI)**			**UM Test AUC (95% CI)**		
	All	**Female**	**Male**	**All**	**Female**	**Male**
Multiclass						
Original VA model	0.9742 (0.9718, 0.9770)	0.9691 (0.9413, 0.9846)	0.9744 (0.9721, 0.9777)	0.8685 (0.8644, 0.8726)	0.8689 (0.8612, 0.8738)	0.8680 (0.8620, 0.8740)

AKI-1+						
Incidence	3.53% (23,957/678,516)	2.00% (745/37,226)	3.62 (23,212/641,290)	3.76% (58,382/1,551,354)	3.18% (24,585/772,665)	4.34% (33,797/778,689)
Original VA model	0.8196 (0.8168, 0.8223)	0.7943 (0.7770, 0.8116)	0.8194 (0.8166, 0.8222)	0.8469 (0.8453, 0.8484)	0.8477 (0.8453, 0.8501)	0.8439 (0.8419, 0.846)
AKI-2+						
Incidence	0.41% (3,277/806,465)	0.34% (139/40,855)	0.41% (3,138/765,610)	0.86% (15,076/1,753,474)	0.75% (6,472/857,809)	0.96% (8,604/895,665)
Original VA model	0.7741 (0.7656, 0.7825)	0.7636 (0.7191, 0.8080)	0.7749 (0.7663, 0.7835)	0.6550 (0.6494, 0.6606)	0.6504 (0.6419, 0.6590)	0.6622 (0.6549, 0.6695)
AKI-3+						
Incidence	0.22% (1,775/821,316)	0.21% (88/41,443)	0.22% (1,687/779,873)	0.39% (6,976/1,790,447)	0.32% (2,780/875,621)	0.46% (4,196/914,826)
Original VA model	0.8341 (0.8248, 0.8433)	0.7111 (0.6520, 0.7703)	0.8393 (0.8300, 0.8486)	0.7981 (0.7919, 0.8044)	0.7627 (0.7518, 0.7737)	0.8271 (0.8198, 0.8345)
AKI-3D						
Incidence	0.11% (940/839,964)	0.15% (61/42,071)	0.11% (879/797,893)	0.08% (1,412/1,817,604)	0.07% (586/887,574)	0.09% (826/930,030)
Original VA model	0.9497 (0.9429, 0.9565)	0.8927 (0.8251, 0.9602)	0.9537 (0.9487, 0.9588)	0.9558 (0.9507, 0.9609)	0.9560 (0.9480, 0.9641)	0.9550 (0.9483, 0.9618)

2.3.2 Model Discrimination and Calibration at the VA

Among eligible 6 h windows (those in which the outcome had not already occurred), our GBDT model predicted any AKI in the next 48 h with an AUC of 82.0% (95% confidence interval (CI) 81.7%, 82.2%) in the VA test set, which was lower than DeepMind's observed AUC for a similar GBDT of 88.9% (95% CI 88.6%, 89.2%). The rationale behind our selection of a GBDT model is provided in Methods. The model's AUCs for AKI stages 2+, 3+ and 3D were 77.4%, 83.4% and 95.0%, respectively (full details in **Table 2.3**). The performance substantially varied between VA hospitals in the test set, with AUCs ranging from 61.5% to 98.5%, suggesting that

even a high-performing model may not generalize across all VA sites (**Supplemental Figure 2.3**). Overall, the model was well calibrated for all levels of AKI (**Supplemental Figure 2.1a**). However, the model overestimated the risk of AKI-1+ in females as compared with males (**Supplemental Figure 2.1a**). The model also had worse discrimination in females when predicting AKI-3+ and 3D, with AUCs of 71.1% for AKI-3+ (as compared with 83.9% in males) and 89.3% for AKI-3D (as compared with 95.4% in males).

2.3.3 Generalizability of the AKI Model at UM

Among eligible windows, the GBDT model predicted any AKI in the next 48 h in the UM test set with an AUC of 84.7% (95% CI 84.5%, 84.8%), which was higher than the AUC of 82.0% we observed in the VA test set. In the UM test set, the model's AUCs for AKI stages 2+, 3+ and 3D were 65.5%, 79.8% and 95.6%, respectively (full details in **Table 2.3**). While the model appeared to generalize well overall, there was a marked difference in AUC for stage 2+ AKI (65.5% at UM versus 77.4% at the VA).

The model generally overestimated the risk of AKI at all stages, and this finding was worse in females as compared with males (**Supplemental Figure 2.1b** and **Supplemental Table 2.2**). Also, similar to our finding in the VA test set, the model performed worse in females when predicting AKI stage 3+, with an AUC of 76.3% in females as compared with 82.7% in males (**Table 2.3**).

2.3.4 Understanding the Differential Performance by Sex

Because the VA population consists of 94% males, one potential reason for the worse performance observed in females is the relatively small number of females who progressed to AKI stage 3+ (n = 94 in the entire VA cohort, **Table 2.2**). If the worse performance in females

was primarily attributable to the lower number of events observed during training, then updating the model using data from a sex-balanced population should improve the model's performance in females. Thus, starting with our 160-tree GBDT model, we continued to further train it using the UM training cohort (in which 50% are females), with early stopping determined based on the UM validation cohort (as described in Methods). This process added ten trees to the original model, and we refer to this updated model as the 'extended model' to highlight that this 170-tree model contains the original 160 trees within it, and is thus an extension of the original model. Remarkably, this small extension to the original model improved the performance in the UM test set both overall and between sexes (**Table 2.4**). Whereas the original model had poorly predicted AKI 2+ at UM (AUC 65.5%), the extended model performs much better on the UM test set (AUC 81.8%). At AKI stage 3+, where the original model exhibited the largest difference between females and males (AUC 76.3% versus 82.7%), the performances were much more similar in the extended model (AUC 85.5% for females and 88.6% for males). The overall calibration was also better in the UM test set (**Supplemental Figure 2.4** and **Supplemental Table 2.3**). While this mechanism of updating a base model in a local population is a promising approach to correcting issues related to model generalizability, the small sample of females used to train the original model does not entirely explain the differential performance by sex. When the extended model was re-evaluated on the VA test set, its performance was worse in females, with an AUC for AKI 3+ of 69.1% in females as compared with 82.8% in males (**Table 2.5**).

Table 2.4 Model performance (AUC) of the extended VA model at UM, by outcome stage, by sex.

Outcome	UM Test AUC (95% CI)		
	All	Female	Male

Multiclass				
	Extended VA model	0.8780 (0.8749, 0.8826)	0.8757 (0.8697, 0.8813)	0.8795 (0.8752, 0.8850)
AKI-1+				
	Extended VA model	0.8523 (0.8508, 0.8538)	0.8535 (0.8512, 0.8559)	0.8490 (0.8470, 0.8510)
AKI-2+				
	Extended VA model	0.8181 (0.8138, 0.8224)	0.8135 (0.8070, 0.8200)	0.8236 (0.8179, 0.8292)
AKI-3+				
	Extended VA model	0.8722 (0.8666, 0.8778)	0.8554 (0.8461, 0.8647)	0.8858 (0.8790, 0.8927)
AKI-3D				
	Extended VA model	0.9346 (0.9258, 0.9433)	0.9402 (0.9271, 0.9532)	0.9297 (0.9178, 0.9415)

Table 2.5 Model performance (AUC) of the extended VA models at VA, by outcome stage, by sex.

Outcome		**VA Test AUC (95% CI)**		
		All	**Female**	**Male**
Multiclass				
	Extended VA model	0.9530 (0.9501, 0.9574)	0.9474 (0.9208, 0.9653)	0.9531 (0.9506, 0.9579)
AKI-1+				
	Extended VA model	0.8178 (0.8150, 0.8178)	0.7892 (0.7717, 0.8067)	0.8178 (0.8150, 0.8206)
AKI-2+				
	Extended VA model	0.7593 (0.7507, 0.7679)	0.7432 (0.6976, 0.7888)	0.7602 (0.7515, 0.7690)
AKI-3+				
	Extended VA model	0.8230 (0.8131, 0.8329)	0.6907 (0.6318, 0.7495)	0.8284 (0.8184, 0.8384)
AKI-3D				
	Extended VA model	0.9355 (0.9261, 0.9450)	0.8925 (0.8252, 0.9599)	0.9385 (0.9298, 0.9472)

Differences in model performance were not observed between racial groups (**Table 2.6** and **Table 2.7**), potentially because the VA population includes a relatively high proportion of Black patients.

Table 2.6 Model performance (AUC) of the original and extended VA models at VA, by outcome stage, by race.

Outcome	VA Test AUC (95% CI)				
	All	**Caucasian**	**African American**	**Other**	**Unknown**
Multiclass					
Original VA model	0.9742 (0.9718, 0.9770)	0.9729 (0.9686, 0.9759)	0.9758 (0.9704, 0.9812)	0.9796 (0.9740, 0.9842)	0.9714 (0.9603, 0.9809)
Extended VA model	0.9530 (0.9501, 0.9574)	0.9506 (0.9450, 0.9544)	0.9563 (0.9498, 0.9633)	0.9591 (0.9528, 0.9664)	0.9526 (0.9375, 0.9638)
AKI-1+					
Incidence	3.53% (23,957/678,516)	3.47% (16,115/463,936)	3.78% (4,825/127,634)	3.43% (2,017/58,830)	3.56% (1,000/28,116)
Original VA model	0.8196 (0.8168, 0.8223)	0.8217 (0.8184, 0.8250)	0.8109 (0.8047, 0.8171)	0.8174 (0.8078, 0.8269)	0.8277 (0.8137, 0.8417)
Extended VA model	0.8178 (0.8150, 0.8206)	0.8196 (0.8162, 0.8230)	0.8109 (0.8046, 0.8171)	0.8145 (0.8048, 0.8241)	0.8264 (0.8124, 0.8404)
AKI-2+					
Incidence	0.41% (3,277/806,465)	0.36% (1,997/548,168)	0.49% (761/155,339)	0.57% (393/69,438)	0.38% (126/33,520)
Original VA model	0.7741 (0.7656, 0.7825)	0.7596 (0.7485, 0.7707)	0.7937 (0.7767, 0.8107)	0.8026 (0.7820, 0.8233)	0.8070 (0.7625, 0.8514)
Extended VA model	0.7593 (0.7507, 0.7679)	0.7463 (0.7351, 0.7575)	0.7815 (0.7644, 0.7986)	0.7752 (0.7525, 0.7980)	0.7898 (0.7423, 0.8373)
AKI-3+					
Incidence	0.22% (1,775/821,316)	0.19% (1,040/557,294)	0.27% (424/159,121)	0.34% (238/70,859)	0.21% (73/34,042)
Original VA model	0.8341 (0.8248, 0.8433)	0.8189 (0.8067, 0.8312)	0.8486 (0.8281, 0.8691)	0.8706 (0.8524, 0.8889)	0.8451 (0.8042, 0.8861)

Extended VA model	0.8230 (0.8131, 0.8329)	0.8103 (0.7974, 0.8231)	0.8375 (0.8162, 0.8588)	0.8442 (0.8196, 0.8688)	0.8418 (0.7941, 0.8895)
AKI-3D					
Incidence	0.11% (940/839,964)	0.10% (567/568,043)	0.14% (225/164,387)	0.12% (88/72,834)	0.17% (60/34,700)
Original VA model	0.9497 (0.9429, 0.9565)	0.9500 (0.9407, 0.9593)	0.9332 (0.9184, 0.9479)	0.9696 (0.9539, 0.9853)	0.9684 (0.9568, 0.9801)
Extended VA model	0.9355 (0.9261, 0.9450)	0.9350 (0.9221, 0.9479)	0.9153 (0.8940, 0.9366)	0.9644 (0.9490, 0.9797)	0.9632 (0.9506, 0.9758)

Table 2.7 Model performance (AUC) of the original and extended VA models at UM, by outcome stage, by race.

Outcome	**UM Test AUC (95% CI)**				
	All	**Caucasian**	**African American**	**Other**	**Unknown**
Multiclass					
Original VA model	0.8685 (0.8644, 0.8726)	0.8697 (0.8649, 0.8742)	0.8625 (0.8500, 0.8714)	0.8689 (0.8421, 0.8861)	0.8576 (0.8375, 0.8916)
Extended VA model	0.8780 (0.8749, 0.8826)	0.8799 (0.8755, 0.8850)	0.8733 (0.8622, 0.8806)	0.8759 (0.8529, 0.8936)	0.8565 (0.8385, 0.8885)
AKI-1+					
Incidence	3.76% (58,382/1,551,354)	3.71% (47,489/1,281,347)	4.22% (7,389/174,932)	3.49% (2,688/77,092)	4.54% (816/17,983)
Original VA model	0.8469 (0.8453, 0.8484)	0.8460 (0.8443, 0.8477)	0.8433 (0.8390, 0.8476)	0.8561 (0.8491, 0.8631)	0.8859 (0.8762, 0.8957)
Extended VA model	0.8523 (0.8508, 0.8538)	0.8514 (0.8497, 0.8531)	0.8488 (0.8446, 0.8530)	0.8624 (0.8555, 0.8693)	0.8905 (0.8811, 0.8999)
AKI-2+					
Incidence	0.86% (15,076/1,753,474)	0.83% (12,064/1,445,319)	1.01% (2,033/201,650)	0.80% (686/86,207)	1.44% (293/20,298)
Original VA model	0.6550 (0.6494, 0.6606)	0.6519 (0.6456, 0.6581)	0.6535 (0.6383, 0.6687)	0.6680 (0.6412, 0.6948)	0.7646 (0.7312, 0.7980)
Extended VA model	0.8181 (0.8138, 0.8224)	0.8158 (0.8110, 0.8205)	0.8215 (0.8101, 0.8330)	0.8406 (0.8210, 0.8601)	0.8342 (0.8040, 0.8644)
AKI-3+					

	Incidence	0.39% (6,976/1,790,447)	0.37% (5,451/1,475,447)	0.50% (1,038/206,309)	0.40% (350/87,868)	0.66% (137/20,742)
	Original VA model	0.7981 (0.7919, 0.8044)	0.7925 (0.7853, 0.7998)	0.8187 (0.804, 0.8333)	0.8063 (0.7776, 0.8349)	0.8585 (0.8190, 0.8979)
	Extended VA model	0.8722 (0.8666, 0.8778)	0.8763 (0.8701, 0.8826)	0.8518 (0.8366, 0.8670)	0.8626 (0.8367, 0.8885)	0.8980 (0.8632, 0.9327)
AKI-3D						
	Incidence	0.08% (1,412/1,817,604)	0.07% (981/1,496,642)	0.12% (258/210,914)	0.15% (134/89,093)	0.19% (39/20,955)
	Original VA model	0.9558 (0.9507, 0.9609)	0.9546 (0.9483, 0.9609)	0.9584 (0.9503, 0.9666)	0.9540 (0.9354, 0.9725)	0.9581 (0.9082, 1)
	Extended VA model	0.9346 (0.9258, 0.9433)	0.9375 (0.9276, 0.9475)	0.9332 (0.9118, 0.9546)	0.9299 (0.9023, 0.9575)	0.8748 (0.7809, 0.9687)

2.3.5 Role of Patient Characteristics in Performance Discrepancy

The extended model's worse performance in the VA population could be attributable either to differences in care patterns between females and males at the VA, or to differences in female patient characteristics at the VA and UM. As compared with females at the VA, females at UM were younger (UM, 55.2 (s.d. 19.1); VA, 58.4 yr (s.d. 14.6)) and less diverse (UM, 81.4% white; VA, 58.8% white), and were more likely to have baseline chronic kidney disease (eGFR < 30 at UM, 3.1%; VA, 2.0%) and congestive heart failure (UM, 23.2%; VA, 6.5%), but had similar body mass indices (UM, 29.1 (s.d. 7.4); VA, 30.7 (s.d. 7.4)) and a similar rate of diabetes mellitus (UM, 26.2%; VA, 24.0%) (**Supplemental Table 2.4**).

To address whether these differences in patient characteristics could explain this performance discrepancy, we matched female patients in the UM test with females at the VA. Details of the matching process are discussed in Methods. We then compared the extended model's performance (which was updated on the UM training set) in the subgroup of UM test set females

who most closely resembled VA females (**Supplemental Table 2.4**). If the differences in patient characteristics accounted for the differences in model performance, then we would expect the model performance in this UM test set subpopulation to mirror the VA test set. On the contrary, the model performance in this matched UM test set was much more similar to the overall UM test set than to the VA population (**Supplemental Table 2.5**).

2.4 Discussion

In our study, drawing on a population of US veterans from over 100 VA hospitals, we observed an AUC for predicting any AKI in the next 48 h of 82.0% in a national VA cohort, which was lower than the AUC of 88.9% for a similar GBDT described in the DeepMind paper. At lower stages of AKI, we found the model to be miscalibrated in females, which could align with the DeepMind team's finding of a lower sensitivity in females as compared with males. However, we also uncovered a lower AUC in females as compared with males in higher stages of AKI, a difference that was not evaluated in the DeepMind study. This difference persists when the VA-trained model is transported to a large academic hospital. While further training on a sex-balanced cohort improved the discrepancy in model performance at the academic hospital, it worsened the discrepancy in model performance at the VA, suggesting that the lower performance in females is related to factors other than simply a low number of events observed during training at the VA.

Our finding that a modelling strategy relying on only VA data results in worse performance in females is troubling. Had the differences been attributable solely to the small sample size of females observed during training, these differences should have been correctable by updating the model using information from a sex-balanced cohort as was present at our academic hospital. However, updating the model actually worsened this discrepancy at the VA, which suggests that

other factors such as practice patterns or patient characteristics for females treated at the VA may account for this difference in the VA context. Practice patterns appear to be a more likely explanation because the updated model continued to perform better at UM even when the analysis was limited to a UM subgroup that most closely resembled VA females. A low number of events in the VA test set remains a possible source of measurement error.

Our work has limitations that may affect our findings. While we approximated aspects of the DeepMind study, including similar inclusion and exclusion criteria, a similar modelling strategy and the inclusion of many of the same predictors (**Supplemental Table 2.4**), we were unable to include International Classification of Diseases, Ninth Revision (ICD-9) codes and clinical note headings as predictors in our model due to computational constraints within the VA computing environment. Billing codes also undergo periodic updates, which can result in models becoming outdated. By the time the DeepMind study was published, ICD-9 codes had been replaced with ICD-10 codes, and the implementation of ICD-11 codes is already underway[65]. ICD-9 codes were known to be an important component of the DeepMind AKI model. For example, 'malignant neoplasm of [the] kidney' was reported as one of the top features in the original study[31], possibly because this billing code foreshadows an imminent nephrectomy or renal artery embolization.

Our work also has important implications. While sex and gender inequalities in healthcare machine learning models have long been suspected, we provide definitive evidence that this phenomenon can and does occur, and that it is complex, not simply explained away by a low sample size seen during model training. We also show promising results that some of these differences attributable to models trained in imbalanced populations can be mitigated through further training on a balanced population, which means that base models trained in a large

population may be capable of being fine-tuned through a relatively simple mechanism in tree ensemble models. In the interest of promoting transparency, we have made our original and extended models publicly available[55].

2.5 Data Availability

This study used data from the national Veterans Health Administration's Corporate Data Warehouse and the University of Michigan. Analyses were performed in secure locations within the VA and UM information systems, respectively. The data in this study are not publicly available because they contain protected health information, and restrictions apply to their use. A sample of processed data from six patients has been made available online[55]. Researchers interested in obtaining deidentified Michigan Medicine patient data should contact PHDataHelp@umich.edu to obtain guidance on which regulatory and compliance requirements need to be fulfilled to obtain access to the Precision Health data resources. More details about the data and the access process are available at https://precisionhealth.umich.edu/. Source data are provided with this paper.

2.6 Code Availability

Data preparation code, an example of prepared data, the original and extended models trained in this study, and code to generate predictions from the provided data are available online[55]. Data preparation requires the gpmodels R package[62].

2.7 Supplemental Materials

2.7.1 Supplemental Tables

Supplemental Table 2.1 Description of predictors used in the presented model.

Variable	Description	Fixed/Temporal	Category/EHR domain	Unit	Valid Range	Time windows used (in hours)	Summary Statistics	Number of predictors
age	Age	Fixed	Demographics	year				1
sex	Sex	Fixed	Demographics		F, M			1
baseline_scr	Baseline Serum Creatinine	Fixed	Lab Results	mg/dL	0-4			1
ht	Baseline Height	Fixed	Vital Signs	inch	48-96			1
wt	Baseline Weight	Fixed	Vital Signs	pound	60-700			1
bmi	Baseline BMI	Fixed	Vital Signs	kg/m2	15-50			1
myocardial_infarction	Myocardial Infarction	Fixed	Comorbidities		0, 1			1
congestive_heart_failure	Congestive Heart Failure (CHF)	Fixed	Comorbidities		0, 1			1
peripheral_vascular_disease	Peripheral Vascular Disease (PVD)	Fixed	Comorbidities		0, 1			1
cerebrovascular_disease	Cerebrovascular Disease	Fixed	Comorbidities		0, 1			1
chronic_pulmonary_disease	Chronic Pulmonary Disease	Fixed	Comorbidities		0, 1			1
dementia	Dementia	Fixed	Comorbidities		0, 1			1
paralysis	Paralysis	Fixed	Comorbidities		0, 1			1
diabetes_uncomplicated	Diabetes Mellitus, Uncomplicated	Fixed	Comorbidities		0, 1			1
diabetes_complicated	Diabetes Mellitus, Complicated	Fixed	Comorbidities		0, 1			1
kidney_disease	Kidney Disease	Fixed	Comorbidities		0, 1			1
mild_liver_disease	Mild Liver Disease	Fixed	Comorbidities		0, 1			1
mod_severe_liver_disease	Moderate or Severe Liver Disease	Fixed	Comorbidities		0, 1			1
peptic_ulcer_disease	Peptic Ulcer Disease	Fixed	Comorbidities		0, 1			1
rheumatic_disease	Rhematoid Disease	Fixed	Comorbidities		0, 1			1
hiv_aids	HIV/AIDS	Fixed	Comorbidities		0, 1			1
metastatic_cancer	Metastatic Cancer	Fixed	Comorbidities		0, 1			1
non-metastatic_cancer	Non-Metastatic Cancer	Fixed	Comorbidities		0, 1			1
surgical_service	Surgical Service	Fixed	Service		0, 1			1
icu_location	ICU location	Fixed	Location		0, 1			1

time	Time (hours) from Admission	Temporal				1
drug_AM300	Aminoglycosides	Temporal	Medications	-24 to 0, -48 to -24, -96 to -48, -120 to -96, -144 to -120, -168 to -144, -192 to -168	length	7
drug_AU100	Sympathomimetics (Adrenergics)	Temporal	Medications	-24 to 0, -48 to -24, -96 to -48, -120 to -96, -144 to -120, -168 to -144, -192 to -168	length	7
drug_CV100	Beta Blockers/related	Temporal	Medications	-24 to 0, -48 to -24, -96 to -48, -120 to -96, -144 to -120, -168 to -144, -192 to -168	length	7
drug_CV150	Alpha Blockers/related	Temporal	Medications	-24 to 0, -48 to -24, -96 to -48, -120 to -96, -144 to -120, -168 to -144, -192 to -168	length	7
drug_CV200	Calcium Channel Blockers	Temporal	Medications	-24 to 0, -48 to -24, -96 to -48, -120 to -96, -144 to -	length	7

drug_CV350	Antilipemic Agents	Temporal	Medications	120, -168 to -144, -192 to -168 -24 to 0, -48 to -24, -96 to -48, -120 to -96, -144 to -120, -168 to -144, -192 to -168 -24 to 0, -48 to -24, -96 to -48, -120 to -96, -144 to -120, -168 to -144, -192 to -168 -24 to 0, -48 to -24, -96 to -48, -120 to -96, -144 to -120, -168 to -144, -192 to -168 -24 to 0, -48 to -24, -96 to -48, -120 to -96, -144 to -120, -168 to -144, -192 to -168 -24 to 0, -48 to -24, -96 to -48, -120 to -96, -144 to -120, -168 to -144, -192 to -168 -24 to 0, -48 to -24, -96 to -48,	length	7
drug_CV702	Loop Diuretics	Temporal	Medications		length	7
drug_CV800	ACE Inhibitors	Temporal	Medications		length	7
drug_CV805	Angiotensin II Inhibitor	Temporal	Medications		length	7
drug_DX101	Non-Ionic Contrast Media	Temporal	Medications		length	7

drug_MS102	Nonsalicylate NSAIs, Antirheumatic	Temporal	Medications			-120 to -96, -144 to -120, -168 to -144, -192 to -168 -24 to 0, -48 to -24, -96 to -48, -120 to -96, -144 to -120, -168 to -144, -192 to -168 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -	length	7
Albumin	Serum Albumin	Temporal	Lab Results	g/dL	0.5-8		length, min, mean, median, max	40
Alkaline	Alkaline Phosphatase	Temporal	Lab Results	IU/L	1-10000		length, min, mean, median, max	40
ALT	Alanine Aminotransferase	Temporal	Lab Results	IU/L	1-10000		length, min, mean, median, max	40
AST	Aspartate Transaminase	Temporal	Lab Results	U/L	1-10000		length, min, mean,	40

Bilirubin_D	Bilirubin (Direct)	Temporal	Lab Results	mg/dL	0-50	30 to -24, -36 to -24, -42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24,	median, max length, min, mean, median, max	40
Bilirunbin	Bilirubin (Total)	Temporal	Lab Results	mg/dL	0-50		length, min, mean, median, max	40
BUN	Blood Urea Nitrogen	Temporal	Lab Results	mg/dL	1-300		length, min, mean, median, max	40
Calcium	Serum Calcium	Temporal	Lab Results	mg/dL	3-20		length, min, mean, median, max	40
Carbon	Carbon Dioxide	Temporal	Lab Results	meq/dL	1-60		length, min, mean, median, max	40

Chloride	Serum Chloride	Temporal	Lab Results	meq/dL	60-150	-42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42	length, min, mean, median, max	40
Glucose	Serum Glucose	Temporal	Lab Results	mg/dL	10-1200	-42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42	length, min, mean, median, max	40
HDLC	High-Density Lipoprotein Cholesterol	Temporal	Lab Results	mg/dL	10-150	-42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42	length, min, mean, median, max	40
Hematocrit	Hematocrit	Temporal	Lab Results	%	10-80	-42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42	length, min, mean, median, max	40
Hemo_A1C	Hemoglobin A1c (Glycohemoglobin)	Temporal	Lab Results	%	0-24	-42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42	length, min, mean, median, max	40

Hgb	Hemoglobin	Temporal	Lab Results	g/dL	2-20	-6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42	length, min, mean, median, max	40
INR	International Normalized Ratio	Temporal	Lab Results	ratio	0.3-10	-6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42	length, min, mean, median, max	40
LDLC	Low-Density Lipoprotein Cholesterol	Temporal	Lab Results	mg/dL	10-300	-6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42	length, min, mean, median, max	40
MC_Ratio	Microalbumin-to-Creatinine Ratio	Temporal	Lab Results	mg/g	0-30000	-6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42	length, min, mean, median, max	40
Phosphate	Serum Phosphate	Temporal	Lab Results	mg/dL	0.1-20	-6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42	length, min, mean, median, max	40
Platelet	Platelet Count	Temporal	Lab Results	count/volume	0-1000	-6 to 0, -12 to -6, -18 to -12,	length, min, mean,	40

						-24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24,	median, max	
Potassium	Serum Potassium	Temporal	Lab Results	meq/L or mmol/L	1-10		length, min, mean, median, max	40
sCr	Serum Creatinine	Temporal	Lab Results	mg/dL	0.4-20		length, min, mean, median, max	40
Sodium	Serum Sodium	Temporal	Lab Results	meq/L or mmol/L	90-190		length, min, mean, median, max	40
Tot chole	Total Cholesterol	Temporal	Lab Results	mg/dL	10-1000		length, min, mean, median, max	40
Triglyceride	Triglyceride	Temporal	Lab Results	mg/dL	10-10000		length, min, mean, median, max	40

WBC	Total White Blood Cell Count	Temporal	Lab Results	k/uL or k/mm^3	0-50000	-42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42	length, min, mean, median, max	40
WT	Inpatient Weight	Temporal	Vital Signs	pound	60-700	-42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42	length, min, mean, median, max	40
Systolic	Systolic Blood Pressure	Temporal	Vital Signs	mmHg	50-240	-42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42	length, min, mean, median, max	40
Diastolic	Diastolic Blood Pressure	Temporal	Vital Signs	mmHg	30-150	-42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42	length, min, mean, median, max	40
R	Respiratory Rate	Temporal	Vital Signs	breaths/minute	6-50	-42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42 -6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42	length, min, mean, median, max	40

T	Temperature	Temporal	Vital Signs	Fahrenheit	80-110	-6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42	length, min, mean, median, max	40
Pulse	Pulse	Temporal	Vital Signs	beats/minute	30-200	-6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42	length, min, mean, median, max	40
PO2	Partial Pressure of Oxygen	Temporal	Vital Signs	mmHg	30-100	-6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42	length, min, mean, median, max	40
CVP	Central Venous Pressure	Temporal	Vital Signs	mmHg	0.5-30	-6 to 0, -12 to -6, -18 to -12, -24 to -18, -30 to -24, -36 to -24, -42 to -36, -48 to -42	length, min, mean, median, max	40
cr_ratio_to_baseline	Fold Change from Baseline Creatinine	Temporal	Lab Results			-6 to 0		1
cr_diff_to_baseline	Absolute Value Change from Baseline Creatinine	Temporal	Lab Results	mg/dL		-6 to 0		1
bun_to_cr_ratio	BUN-to-Serum Creatine Ratio	Temporal	Lab Results			-6 to 0		1
current_aki_stage	Current AKI Stage	Temporal			no_aki, aki_1, aki_2,		1	

	aki_3,
	aki_3d
Total	1467

Supplemental Table 2.2 Expected calibration error (ECE) for the original VA model at VA and at UM.

ECE	VA			UM		
	All	**Female**	**Male**	**All**	**Female**	**Male**
AKI-1+	0.20%	0.43%	0.21%	0.63%	0.67%	0.64%
AKI-2+	0.05%	0.10%	0.06%	0.59%	0.60%	0.58%
AKI-3+	0.02%	0.10%	0.02%	0.12%	0.16%	0.10%
AKI-3D	0.01%	0.07%	0.02%	0.16%	0.14%	0.18%

Supplemental Table 2.3 Expected calibration error (ECE) for the extended VA model at UM.

ECE	UM		
	All	**Female**	**Male**
AKI-1+	0.48%	0.47%	0.55%
AKI-2+	0.37%	0.43%	0.32%
AKI-3+	0.09%	0.14%	0.05%
AKI-3D	0.11%	0.09%	0.12%

Supplemental Table 2.4 Patient characteristics of females at the University of Michigan (UM), the Veteran Affairs (VA), and a subpopulation of UM test set females matched to the VA females.

UM test set female patients were matched to VA female patients by mean age, proportion of white race, proportion of baseline chronic kidney disease (CKD), and proportion of baseline congestive heart failure (CHF). Body mass index (BMI) and baseline diabetes mellitus (DM) were not used for matching but are shown below.

	UM **N = 82,266** **(49,743 patients)**	**VA** **N = 15,759**	**UM test set** **(matched to VA)** **N = 3,119** (same number as females in the VA test set)
Age, mean (SD)	55.2 (19.1)	58.4 (14.6)	58.8 (18.4)
White race	66,949 (81.4%)	9,259 (58.8%)	1,815 (58.2%)
Baseline CKD (eGFR* < 30 mL/min/1.73 m2)	2,536 (3.1%)	311 (2.0%)	58 (1.9%)
Baseline CHF	19,056 (23.2%)	1,026 (6.5%)	191 (6.1%)
BMI, mean (SD)	29.1 (7.4)	30.7 (7.4)	28.8 (7.1)
Baseline DM	21,594 (26.2%)	3,788 (24.0%)	825 (26.5%)

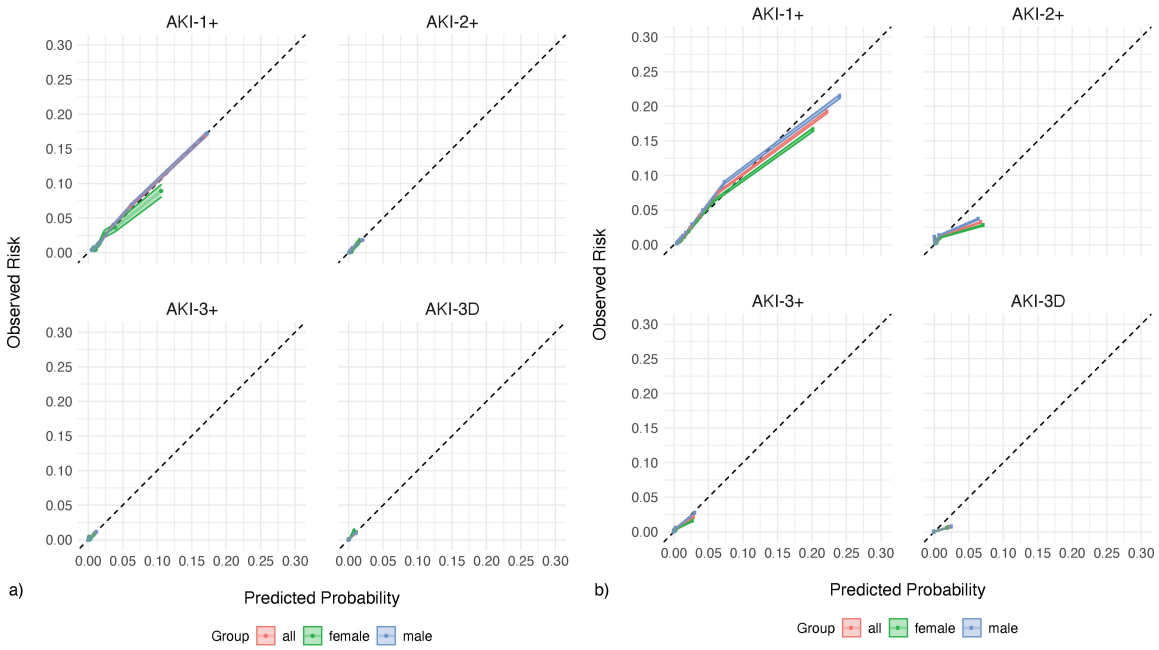
Statistics presented: mean (SD); n (%)

* Calculated based on CKD-EPI Creatinine Equation (2021)

Supplemental Table 2.5 Model performance for females in the University of Michigan (UM) test set, the Veterans Affairs (VA) test set, and a subpopulation of UM test set females matched to the VA females.

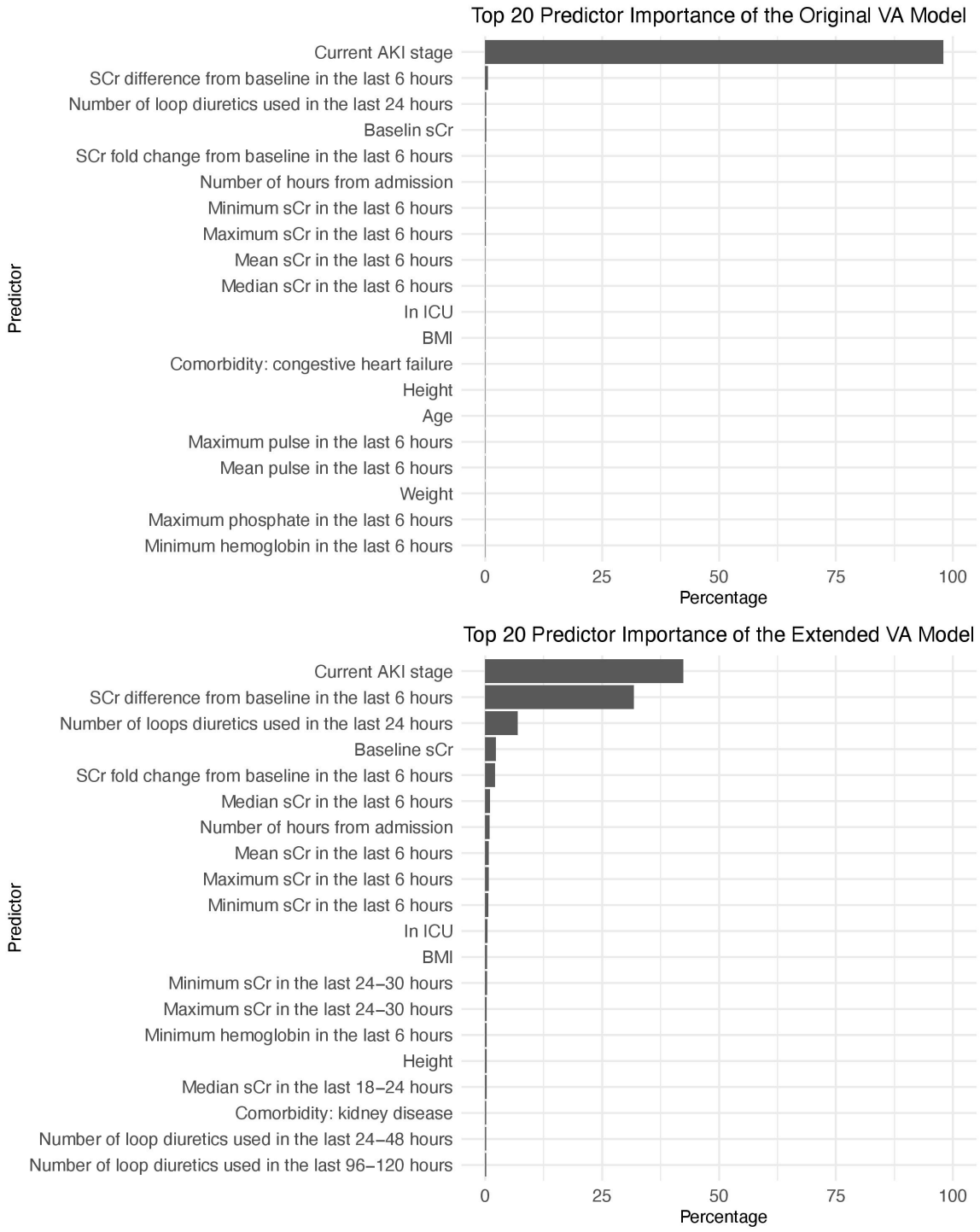
Outcome	Extended VA Model AUC (95% CI)		
	UM test set N = 49,280	**VA test set N = 3,119**	**UM test set (matched to VA) N = 3,119**
Multiclass	0.8757 (0.8697, 0.8813)	0.9474 (0.9208, 0.9653)	0.8781 (0.8699, 0.8945)
AKI-1+	0.8535 (0.8512, 0.8559)	0.7892 (0.7717, 0.8067)	0.8402 (0.8291, 0.8512)
AKI-2+	0.8135 (0.8070, 0.8200)	0.7432 (0.6976, 0.7888)	0.7980 (0.7674, 0.8286)
AKI-3+	0.8554 (0.8461, 0.8647)	0.6907 (0.6318, 0.7495)	0.8564 (0.8218, 0.8910)
AKI-3D	0.9402 (0.9271, 0.9532)	0.8925 (0.8252, 0.9599)	0.9773 (0.9583, 0.9963)

2.7.2 Supplemental Figures



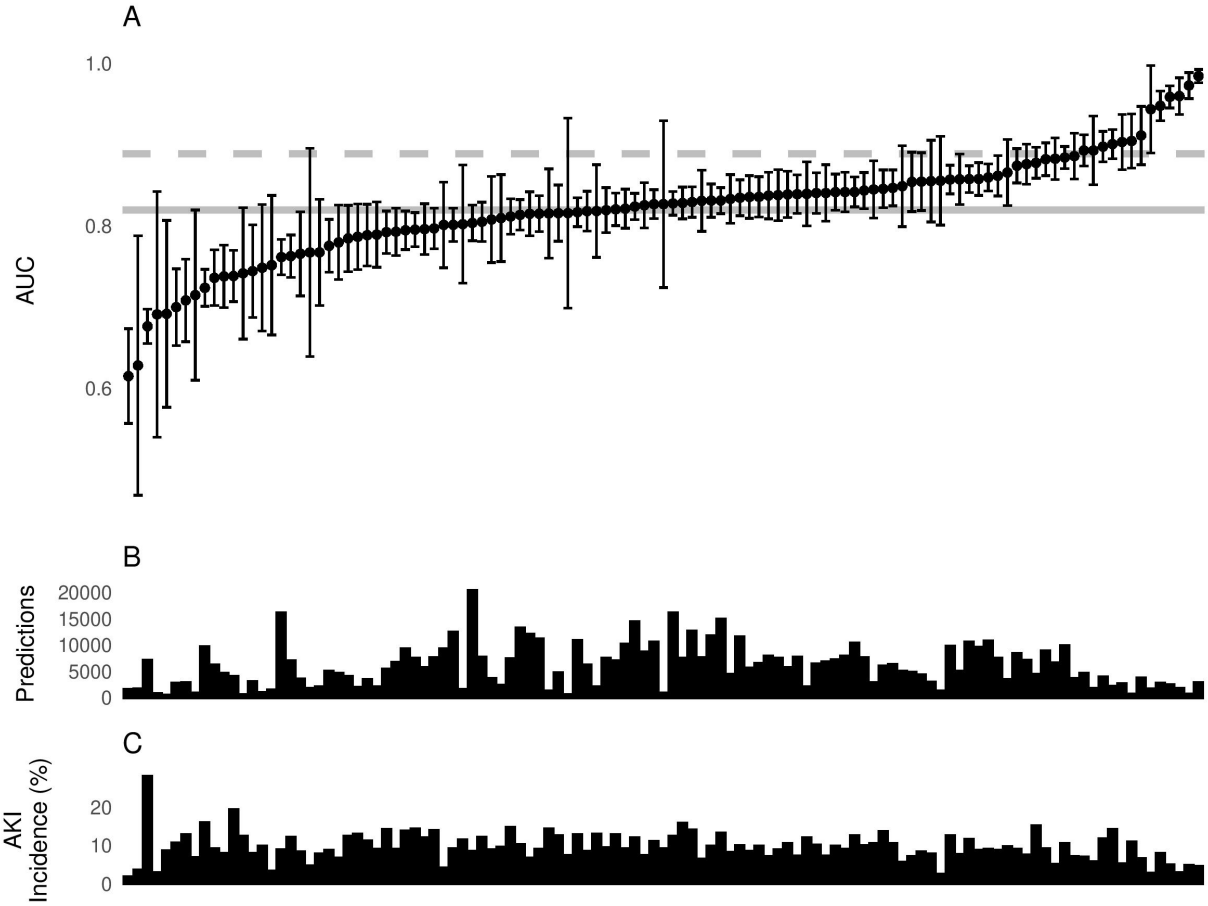
Supplemental Figure 2.1 Calibration of the original VA model a) VA test set b) UM test set.

The calibration of the original model on the a) VA test set and b) UM test set. The predicted probabilities (deciles) are plotted against the observed probabilities with 95% confidence intervals. The diagonal line demonstrates the ideal calibration. The model calibration is examined for all patients (red), females only (green), and males only (blue).



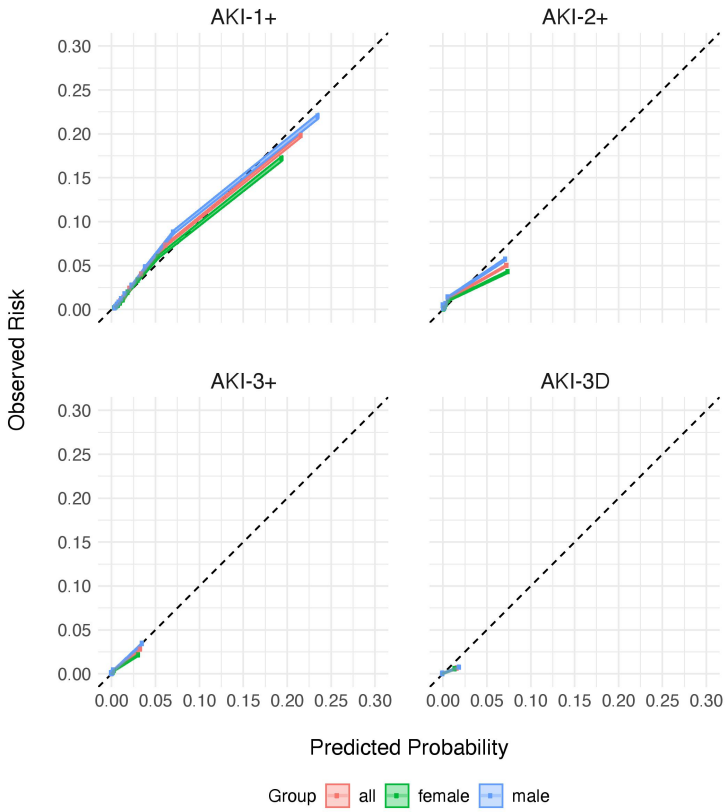
Supplemental Figure 2.2 Predictor importance plot of the original and extended VA model.

Top 20 important predictors of the original VA model (top) and the extended VA model (bottom). Predictors are ranked by their relative importance and expressed as a percentage.



Supplemental Figure 2.3 Model performance (AUC) of the original VA model at each VA hospital.

Model performance of the original model at each VA hospital in the test set, along with characteristics of each VA hospital. A. Model performance with respect to area under the curve (AUC) with 95% CI of the original VA model for predicting AKI-1+ at each VA hospital. The center dot represents the AUC when the original model is applied to the hospital, and the 95% CI is calculated by the DeLong's method24. B. Number of predictions (after excluding those with AKI-1+ at baseline) at each VA hospital. C. Hospitalization level AKI-1+ incidence in the test set (after excluding those with AKI-1+ at baseline) at each VA hospital. Five VA hospitals are not shown here due to small cohort sizes (<30 patients).



Supplemental Figure 2.4 Calibration of the extended VA model at UM.

The calibration of the extended model in the UM test set. The predicted probabilities (deciles) are plotted against the observed probabilities with 95% CI. The diagonal line demonstrates the ideal calibration. The model calibration is examined for all patients (red), females only (green), and males only (blue).

Chapter 3 Assessing the Role of Urine Output in Acute Kidney Injury Risk Prediction

3.1 Introduction

Acute kidney injury (AKI) is a clinically significant condition associated with worse patient outcomes, including increased morbidity and mortality, imposing a substantial burden on healthcare resources[1,8–11,13]. The absence of a universally accepted definition for AKI has hindered research progress for a long period until the Kidney Disease Improving Global Outcomes (KDIGO) criteria emerged as the standard for AKI definition. KDIGO uses both serum creatinine (sCr) and urine output (UO) to define and stage AKI. UO, as a rapid bedside test for kidney function, is the oldest known biomarker for AKI, with a rapid reduction indicating potential kidney function decline[66].

Researchers have explored whether a rapid reduction of UO can serve as an early biomarker for AKI. In a prospective observation study, Macedo et al. assessed hourly UO in ICU patients, discovering oliguria as a sensitive and early marker for AKI, significantly associated with adverse outcomes[67]. Kellum et al. analyzed electronic health records (EHR) of ICU patients, classifying them by levels of sCr and/or UO, and found that UO assessment was a necessity for staging severe AKI, and low UO was associated with a long-term hazard for AKI 2+ patients. Conversely, Md Ralib et al. found that a UO criterion of 0.5 ml/kg/hour for 6 hours was not predictive of survival; instead, a 6-hour UO threshold of 0.3 ml/kg/hour demonstrated a better association with mortality and dialysis in the studied ICU patients[68,69]. Notably, several other studies also showed discordance between the UO-defined AKI and sCr-defined AKI[70–73]. It is

worth noting that these studies focused exclusively on critically ill patients, leaving uncertainty regarding the utility of UO in identifying high-risk AKI patients outside of the ICU setting. Many of these studies employed prospective designs, collecting hourly UO data for enhanced completeness, though this may not be practical in clinical settings. A more pragmatic understanding of the role of routinely collected UO data in AKI may be gleaned by exploring routinely documented EHR data.

In addition to epidemiological studies exploring the association between UO criterion and AKI outcomes, recent efforts have turned to machine learning tools to further leverage EHR data for identifying high-risk AKI patients. However, despite the consensus of using the KDIGO definition for AKI, most of these studies developing AKI risk prediction models opt for a simplified version that relies solely on sCr, neglecting the UO criterion in defining the ground truth for AKI (**Table 1.2**)[23–26,28,30–33,74]. Researchers transparently acknowledge this limitation, often qualitatively citing it due to the incomplete and inconsistent documentation of UO in electronic health records (EHR)[26,36–38] but provide no quantitative description of UO documentation availability and patterns.

More recently, some studies have started to incorporate UO as a predictor, albeit in a limited set of hospitalized patients. Alfieri et al. developed a random forest model based on UO, biochemical and hematologic data collected during ICU stays, predicting AKI 2+ episodes[75], and demonstrated that urine output trend was predictive of AKI 2+ at least 12 hours in advance[76]. Zhao et al. assessed intraoperative UO in patients undergoing major thoracic surgery but found its poor predictive ability for postoperative AKI[77]. However, these models utilized UO because data is closely monitored in critically ill or surgically treated patients, raising questions about the generalizability of UO's benefits in AKI risk prediction for the broader hospitalized population.

In this study, we used a five-year inpatient cohort at a large academic health system to examine the pattern of urine output documentation in EHR and assess its significance in predicting AKI. The objectives of the studies were to: 1) offer a quantitative description of the completeness and consistency of urine output documentation in the EHR for all hospitalized patients; 2) provide insights into the phenotyping of urine output documentation within the EHR; and 3) quantify the value of utilizing urine output documented in the EHR in AKI risk prediction models.

3.2 Methods

3.2.1 Study Cohort

Clinical data from all adult patients admitted to the University of Michigan (UM) from January 1, 2016 to December 31, 2020 were collected for the study. Patients who did not have creatinine checked at baseline or during their stay (defined in Predictor Variables), had pre-existing end stage renal disease or had a baseline creatinine of >4.0 mg/dL (pre-existing AKI stage 3) were excluded. Notably, in cases where a patient experienced multiple admissions during the study period, all hospitalizations were considered for inclusion, contributing to the final cohort of 165,359 encounters. For each encounter within the cohort, clinical data spanning from arrival to the hospital through discharge or up to 7 days from admission, whichever occurred earlier, were utilized in the study.

3.2.2 Urine Output and Urine Occurrence in EHR

We retrieved both quantitative and qualitative UO documentation from the EHR system used by UM. The UO of patients who underwent surgical procedures were closely monitored during the perioperative period, and the amount was quantitatively documented in the perioperative notes.

When patients were in the general ward or the ICU, both quantitative urine output (with actual volume measured) and qualitative urine occurrence (where no specific volume was measured) were consistently documented in nursing flowsheets.

3.2.3 AKI definition

AKI was defined and staged for severity according to the Kidney Disease: Improving Global Outcomes international guidelines[16]. Stage 1 AKI was defined as a sCr level increase of $\geqslant$0.3 mg/dL, but less than twice the baseline sCr or an increase of 1.5 times baseline. Stage 2 AKI reflected an increase of two to three times the baseline, and stage 3 AKI was an sCr level increase greater than three times baseline or an increase to $\geqslant$4.0 mg/dL. Stage 3D was determined based on the need for dialysis, where the time of first dialysis was determined using procedure codes.

3.2.4 Clustering Analysis

Clustering analysis was conducted on encounters with a minimum of two UO measurements with documented volumes. Each hospitalization was divided into 6-hour intervals, and the cumulative volume of UO recorded within these periods was computed. Summary statistics (minimum, mean, median, and max) of the number of UO measurements documented each day, time (h) between two consecutive UO measurements, total UO volume in the 6-h intervals were calculated as features. Each feature was scaled to the range from 0 to 1 by using the min-max normalization.

We used density-based spatial clustering and application with noise (DBSCAN) for our clustering analysis because it is robust to outliers and the does not require the number of clusters to be specified[78]. Two parameters are required for DBSCAN, epsilon (ε) and minimum points

(MinPts). The parameter ε defines the radius of neighborhood around a data point, while MinPts designates the minimum number of neighbors within the ε radius required for a data point to be considered part of a cluster. If a data point lacks a neighbor count equal to or exceeding the specified MinPts and does not belong to the ε-neighborhood of any core point, it is categorized as an outlier in the analysis.

Considering the extensive size of our dataset, we set the value of MinPts to 50 to ensure robust clustering. Epsilon was then determined by calculating the average of the distances of every point to its k (as specified by MinPts) nearest neighbors, plotting these k-distances in an ascending order on a plot, and finding the "knee" (sharp turning point).

Since UO is typically monitored in ICU stays, and previous AKI predictions studies primarily focused on ICU patients, our clustering analysis focused on patients without ICU stays in this study.

3.2.5 Evaluate the Role of Urine Output in AKI prediction

3.2.5.1 Data Split

In **Chapter 2**, we described an AKI model developed at the Veterans Affairs (VA) can be further trained to be a valuable model at UM. Given that the primary objective of the earlier study was to validate the generalizability of the VA model, we only allocated 20% of the entire UM cohort for training purposes. While this percentage may seem smaller compared to conventional training datasets, it is important to note that this subset of the UM cohort comprised 33,077 hospitalizations, a sufficiently large size for robust model training. Hence, to facility comparability with the previous study, we maintained the same data split for this study—20% for training, 20% for validation, and the remaining 60% of hospitalizations for the test set. To prevent any information leakage, random splits were sampled at the patient level.

3.2.5.2 Predictor Variables

Similar to the approach outlined in **2.2.4 Predictor Variables**, we collected both fixed variables (i.e., baseline variables or those remain constant during a hospitalization) and time-varying variables (i.e., variables measured on a repeated basis and changed values during a hospitalization) for the construction of AKI predictors.

Baseline variables included age, height, weight, body mass index (BMI), comorbidities, and baseline sCr. Note that we excluded two variables—admission to a surgical service and ICU status—that were part of the extended VA model discussed in **Chapter 2**. Age was top-coded at 89 years old. Baseline height and weight were derived from the most recent values within the past year. In instances where no recent value was available, the first inpatient measurement was utilized. Height and weight measurements were converted into inches and pounds, respectively, with extreme values removed. Baseline BMI was calculated using the baseline height and baseline weight. Comorbidities were determined using the Charlson comorbidity index calculated using information from the current hospitalization. Baseline sCr was determined following a hierarchical order of preference: (1) mean outpatient sCr between 7 and 365 days before admission and (2) within 7 days before admission, and (3) first documented sCr value within 24 hours of admission.

Time-varying variables consisted of inpatient vital signs, laboratory test results and administration of medications. Twenty-six laboratory testing components (serum albumin, alkaline phosphatase, alanine aminotransferase, aspartate transaminase, total and direct bilirubin, blood urea nitrogen, serum calcium, carbon dioxide, serum chloride, serum glucose, high-density lipoprotein cholesterol, hematocrit, hemoglobin A1c, hemoglobin, international normalized ratio, low-density lipoprotein cholesterol, microalbumin-to-creatinine ratio, serum phosphate, platelet

count, serum potassium, sCr, serum sodium, total cholesterol, triglyceride and total white blood cell count) were selected due to their universal use across different health systems. Seven vital signs (inpatient weight, systolic blood pressure, diastolic blood pressure, respiratory rate, temperature, pulse, blood oxygen level) were included, irrespective of the frequency of measurement. Administration of medications was examined for 11 drug classes (aminoglycosides, sympathomimetics, beta blockers, alpha blockers, calcium channel blockers, antilipemic agents, loop diuretics, angiotensin-converting enzyme inhibitors, angiotensin II inhibitors, non-ionic contrast media and non-salicylate antirheumatic non-steroidal anti-inflammatory drugs) as opposed to individual medications.

3.2.5.3 Data Preprocessing Feature Engineering

Physiologically infeasible values, potentially arising from laboratory errors, were systematically excluded from the dataset. Microalbumin-to-creatinine ratios were set to 0 when values were reported only in a text field based on the observation that the text fields reported such values as being below the detectable range. Data elements were time-stamped using the time when values became available to the EHR (i.e., the observation time).

After extracting the baseline and time-varying variables, we captured patient states at 6-hour intervals beginning with the time of admission for each patient. Patient states were captured up until the final creatinine value, discharge or death, and truncated at 7 days of hospitalization due to computational constraints.

For each 6-hour interval, baseline variables were directly used as predictors since their values do not change during hospitalization. Time-varying variables were processed in three different ways (**Supplemental Figure 3.1**) to explore simplified versions of the AKI model without compromising performance.

As demonstrated in **Supplemental Figure 3.1**, we first replicated the strategy used in the VA AKI model. Summary statistics were calculated for time-varying variables on a rolling basis. Specifically, number of values, and first, last, minimum, mean, median and maximum values were calculated for the preceding 48 hours (lookback period) divided into 6-hour windows. The number of administered medications was calculated in the same manner. Secondly, temporal predictors were calculated on a rolling basis without breaking the lookback window into smaller windows (i.e., a 48-hour lookback period with a 48-hour window). Thirdly, summary statistics for time-varying variables were calculated in a cumulative manner, using data accumulated from arrival until the time of prediction.

Based on clinical relevance, additional variables were created, including the ratio of the most recent maximum sCr to baseline sCr, the difference between the most recent maximum sCr and baseline sCr, and the ratio of most recent maximum blood urea nitrogen to most recent maximum sCr. These three sCr-based predictors, time (h) from admission and current AKI stage, plus the summary statistics of temporal predictors in the given windowed lookback period, together with the baseline predictors, made up a set of 1,467 predictors that were used in the extended VA AKI model.

Due to the sparsity of microalbumin-to-creatinine ratio data and the unavailability of central venous pressure information in the UM dataset, relevant temporal predictors were removed, resulting in a starting model at UM consisting of 1,393 predictors.

For UO predictors, the same lookback period (48 hours) and the same window intervals (6 hours or 48 hours) were applied to generate rolling UO predictors for the rolling strategy. The rolling UO predictors with a 48-hour lookback period windowed in 6-hour intervals was employed for models with growing predictors due to clinical relevance.

3.2.5.4 Outcome Definition

The calculation of AKI outcomes was performed on a rolling basis at 6-hour intervals, comparing the maximum sCr value within a 48-hour prediction window with the baseline sCr. Each 6-hour interval in which patient states were captured resulted in outcomes categorized into one of five classes based on the 48-hour prediction window: no AKI, AKI stage 1, AKI stage 2, AKI stage 3, or AKI stage 3D. While models were trained using this multinomial outcome, results reported by AKI stages were grouped according to level of severity. For example, AKI stage 1+ is used to refer to any AKI stage, and AKI stage 2+ refers to AKI stage 2 or greater (including stages 3 and 3D).

3.2.5.5 Model Training

The gradient-boosted decision tree (GBDT) model was trained on training set to predict AKI stage in the next 48 h as a multinomial outcome (that is, 'no AKI', 'AKI stage 1', 'AKI stage 2', 'AKI stage 3', 'AKI stage 3D') using different combinations of predictors (rolling predictors with and without smaller windowed intervals, growing predictors, UO predictors, see details discussed in **3.2.5.3**) at each 6 h step with a maximum of 1,000 trees and a maximum depth of 5. The validation set was used to determine the need for early stopping based on an improvement in log loss lower than 0.0005 on five consecutive rounds based on a moving average calculated after every ten trees. The categorical predictors were reordered by the mean response of each level for more efficient training. Internally, a separate one-versus-all tree was trained for each outcome class.

3.2.5.6 Model Evaluation

The performance of the GBDT model was evaluated in the UM test set. The model discrimination was assessed by using the area under the receiver operating curve (AUROC or

AUC). The AUC was reported as a series of binary AUCs where at-risk individuals were evaluated on their risk of progression to a higher AKI stage. For example, patients without any AKI to date were evaluated on their risk of developing any AKI (that is, stage 1 or greater), patients with no AKI or AKI stage 1 were evaluated on their risk of developing AKI stage 2 or greater, and so on.

3.2.5.7 Explore the Association between Urine Output and Other Predictors

To investigate the potential relationship between urine output and other predictors used in the AKI model, we applied the same modeling strategy described above in Model Training. However, in this instance, we substituted the outcome variable with UO predictors, namely the count of UO measurements within the 6-hour/48-hour interval and the cumulative volume of UO documented during the same period, respectively, to build two separate regression tree models. Deviance was selected as the stopping metric because of the continuous nature of the outcome variable.

3.2.6 Software

All data processing and analyses were performed using R 4.1.2[61]. Dbplyr 2.1.1 was used to pull clinical data from UM EHR database. Transformation of time-series data was performed using the Grammar of Prediction (gpmodels) R package[62]. H2O version 3.36.0.3 was used to fit the GBDT model[79].

3.3 Results

3.3.1 Description of Urine Output Documentation in EHR

In our 5-year analysis cohort, 96.7% (159,823/165,359) of encounters had at least one urine output or urine occurrence documented within 7 days following admission. This includes 90.3%

(149,339/165,359) encounters with at least one recorded urine output volume, and 76.0% (125,640/165,359) with documented urine occurrence. As depicted in **Supplemental Table 3.1**, on average, the average frequency of urine output documentation (with or without volume) during a hospital stay was 6.3 ± 3.6 times per day, with a predominant focus on capturing the volume of urine output (5.3 ± 3.8 times per day). The mean duration between two urine output measurements (with or without volume) was 3.2 ± 3.7 hours, while the interval between two documented urine output volumes averaged 3.4 ± 4.7 hours (**Supplemental Table 3.2**). **Figure 3.1** illustrates the distribution of time intervals between two consecutive records for different types of urine output measurements, emphasizing the close monitoring of urine output volume during the perioperative period. When patients were not undergoing surgery, urine output was observed less frequently, with peaks in documentation of volume happening every hour, and peaks in documentation of occurrence every two hours.

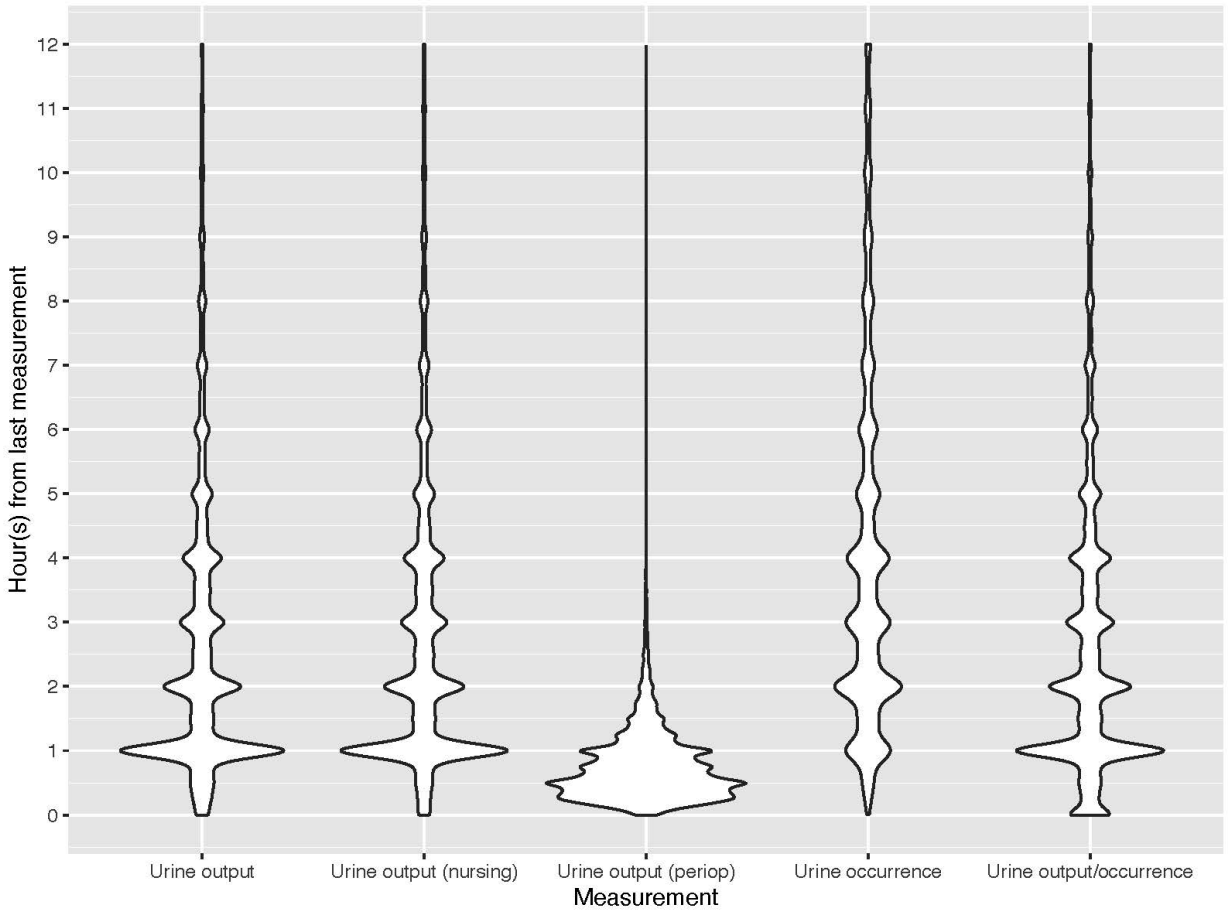


Figure 3.1 Distribution of time (hours) between two urine output measurements, by measurement type.

Violin plot demonstrating distribution of time interval between two documented urine output measurements. Types of urine output measurement include urine output (with volume), urine output – nursing (with volume), urine output – periop (without volume), urine occurrence (without volume), urine output/occurrence (with or without volume).

3.3.2 Phenotypes of Urine Output Documentation and Associated Patient Characteristics in non-ICU patients

Supplemental Figure 3.2 provides a graphical representation of selected hospital stays, depicting the diverse patterns of urine output documentation observed during these periods While most encounters included at least one documented urine output measurement, the consistency and type of measurement greatly varied from stay to stay. Consistent with our previous results, the urine output volume was checked more frequently in perioperative patients.

During stays in the ICU, the consistency of urine output volume documentation, while mostly frequent, was not uniformly guaranteed. Furthermore, when patients were in general wards, the documentation pattern of urine output experienced significant variations. This ranged from systematic documentation of urine output volume, a blend of urine output volume and occurrence recordings, to sparse narrations of urine occurrence only.

While previous studies predominantly focused on predicting AKI in ICU patients, we conducted clustering analysis for patients who never stayed in ICU to facilitate better understanding phenotypes of UO documentation in general wards. **Figure 3.2** visually illustrates representative UO documentation for non-ICU patients across the three clusters identified by our analysis, with 30 patients depicted for each cluster. The three clusters we identified were ordered based on membership size from largest to smallest. Patients in the second cluster, who underwent the most frequent UO monitoring, exhibited a higher incidence of AKI events. In contrast, patients in the first cluster received less frequent but still notable documentation of UO, while patients in the third cluster had minimal documentation. Upon comparing characteristics of patients in the first two clusters (**Table 3.1**), we observed that those receiving the most frequent urine output monitoring were more likely to be older (61.5 years vs. 57.5 years), males (65.1% vs. 50.0%), and had poorer baseline kidney function (baseline eGFR < 60 mL/min/1.73 m^2, 28.6% vs. 19.3%). Additionally, this group had a higher prevalence of baseline congestive heart failure (42.2% vs. 26.1%), received surgical service (35.2% vs. 16.9%) and experienced prolonged hospital stays (13.9 days vs. 5.9 days). The occurrence rate of AKI was highest in the second cluster, reaching 35.5%. The third cluster, albeit smaller in size (N = 44), contained a higher proportion of African American patients (18.2% vs. 11.2%), a greater prevalence of baseline

diabetes (38.6% vs. 29.7%), more baseline liver disease (29.5% vs. 20.1%), longer hospital stays (10.8 days compared to 5.9 days in the first cluster), and the lowest AKI rate at 9.1%.

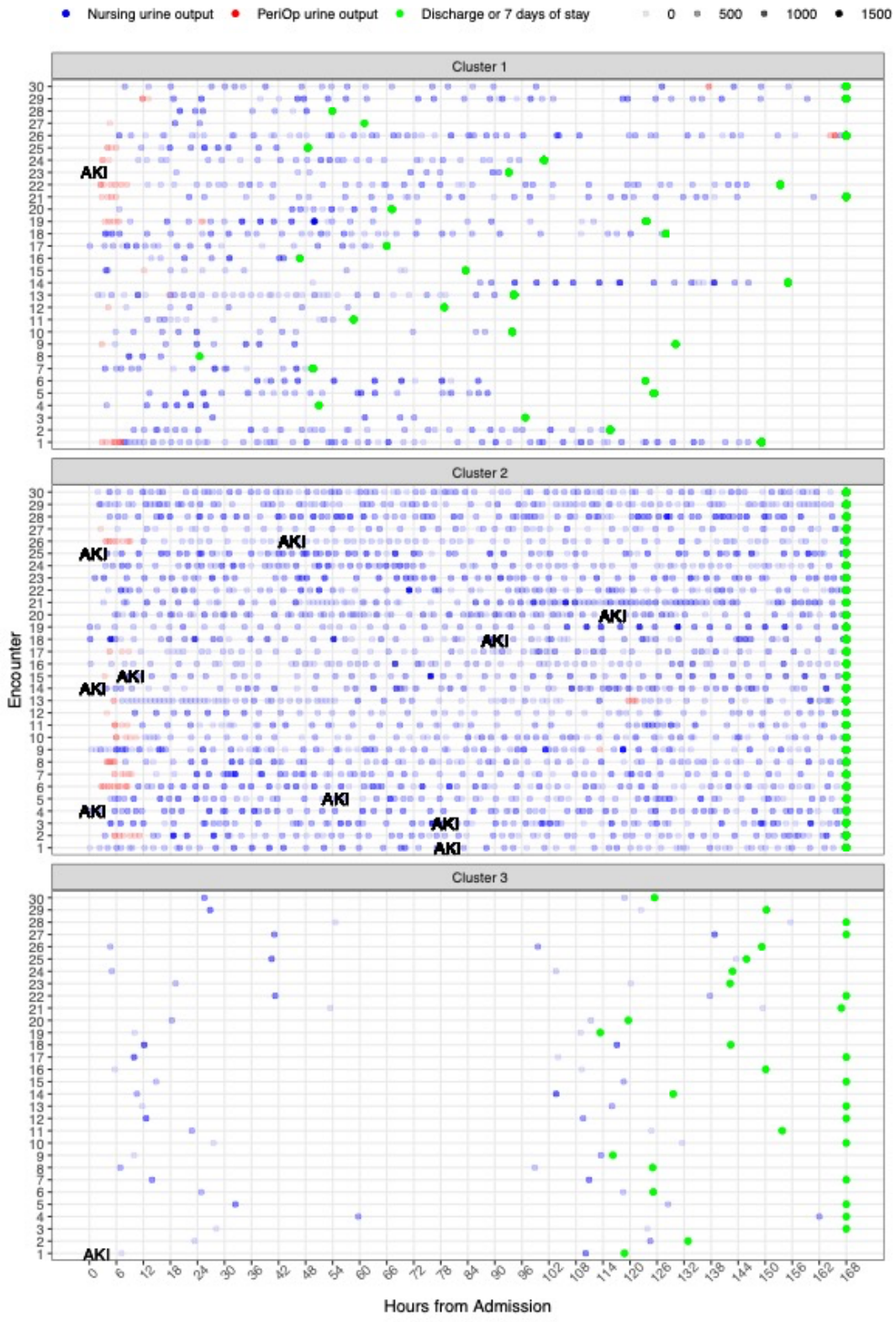


Figure 3.2 Representative visual representation of urine output documentation for three clusters of non-ICU patients.

Table 3.1 Characteristics of non-ICU patients by different clusters.

Characteristic	Overall (N = 138,812)	Cluster 1 (N = 115,224)	Cluster 2 (N = 1,231)	Cluster 3 (N = 44)
Age (years), mean (SD)	56.6 (18.5)	57.5 (18.1)	61.5 (14.7)	54.0 (18.2)
Female	71,039 (51.2%)	57,582 (50.0%)	430 (34.9%)	25 (56.8%)
Race, N (%)				
African American	15,714 (11.3%)	12,883 (11.2%)	130 (10.6%)	8 (18.2%)
Caucasian	114,770 (82.7%)	95,608 (83.0%)	1,049 (85.2%)	33 (75.0%)
Other	6,985 (5.0%)	5,612 (4.9%)	37 (3.0%)	3 (6.8%)
Unknown	1,343 (1.0%)	1,121 (1.0%)	15 (1.2%)	0 (0.0%)
Baseline BMI, mean (SD)	28.8 (6.8)	29.0 (6.8)	29.2 (6.8)	26.9 (6.3)
Unknown	7,628 (5.5%)	6,251 (5.4%)	62 (5.0%)	4 (9.1%)
Baseline serum creatinine (mg/dL), mean (SD)	1.0 (0.4)	1.0 (0.4)	1.1 (0.5)	1.0 (0.3)
Baseline eGFR (mL/min/1.73 m^2), N (%)				
>= 60	113,376 (81.7%)	93,002 (80.7%)	879 (71.4%)	33 (75.0%)
45-59	13,562 (9.8%)	11,770 (10.2%)	150 (12.2%)	8 (18.2%)
30-44	8,151 (5.9%)	7,153 (6.2%)	134 (10.9%)	2 (4.5%)
15-29	3,581 (2.6%)	3,178 (2.8%)	65 (5.3%)	1 (2.3%)
< 15	142 (0.1%)	121 (0.1%)	3 (0.2%)	0 (0.0%)
Baseline diabetes, N (%)	40,484 (29.2%)	34,273 (29.7%)	437 (35.5%)	17 (38.6%)
Baseline congestive heart failure, N (%)	34,157 (24.6%)	30,062 (26.1%)	520 (42.2%)	6 (13.6%)
Baseline liver disease, N (%)	27,819 (20.0%)	23,155 (20.1%)	222 (18.0%)	13 (29.5%)
Surgical service, N (%)	20,318 (14.6%)	19,477 (16.9%)	433 (35.2%)	0 (0.0%)
Length of stay (days), mean (SD)	5.8 (6.9)	5.9 (7.0)	13.9 (9.7)	10.8 (9.8)
Number of 6-hour windows, mean (SD)	17.5 (8.5)	17.8 (8.3)	29.0 (0.2)	25.9 (3.4)
AKI rate, N (%)	21,021 (15.1%)	18,820 (16.3%)	437 (35.5%)	4 (9.1%)

Note: Patients who did not have at least two urine output measurements or were identified as outliers by the DBSCAN algorithm are not included in the three clusters.

3.3.3 Role of Urine Output in AKI prediction

In our cohort, 82.8% encounters had at least one episode where the time between two consecutive urine measurements (urine output or urine occurrence) was greater than 6 hours. That being said, at least 82.8% of encounters would have been categorized as having AKI 1+ if UO is used in the AKI outcome definition.

To assess the role of UO as predictors in AKI model, we initially explored the feasibility of a simplified version of a previously reported AKI model. The original model utilized rolling predictors, employing multiple smaller windowed periods within a 48-hour lookback period, resulting in a total of 1,393 predictors and achieving an AUC of 0.86 for predicting any AKI

(AKI 1+). In our simplified versions of the AKI models, either utilizing rolling predictors without breaking the lookback period into smaller windows or directly using growing predictors, we significantly reduced the number of predictors to 273. Despite the reduction in complexity, the AUCs remained consistent at 0.86, and the model performance exhibited a similar level as the original model did for both moderate and severe AKI stages (**Table 3.2**).

Given the comparable performance to the previously reported AKI model, we further investigated the impact of adding UO predictors to the simplified AKI models. **Table 3.2** illustrates that the inclusion of UO predictors led to marginal increase in AUC for all staged AKI outcomes.

Notably, when using UO predictors exclusively, the model demonstrated good performance in predicting severe AKI (0.75 for AKI 3+ and 0.84 for AKI 3D), as detailed in **Table 3.2**. This suggests that UO predictors alone contribute valuable information for predicting severe AKI outcomes. However, the predictive value of UO predictors alone did not surpass the performance achieved by the baseline predictors.

Table 3.2 Model performance (AUC) of AKI models with different combination of predictors.

Predictors	**Baseline + Rolling (48h lookback, 6h window)**		**Baseline + Rolling (48h lookback, 48h window)**		**Baseline + Growing**		**Baseline only**	**UO (rolling) only**
	Extended VA model	**UM trained model**	**Without UO predictors**	**With UO predictors**	**Without UO predictors**	**With UO predictors**		
No. of predictors	1467	1393	273	297	273	297	120	24
AKI 1+	0.8523 (0.8508, 0.8538)	0.8626 (0.8611, 0.8641)	0.8620 (0.8605, 0.8634)	0.8634 (0.8619, 0.8649)	0.8593 (0.8578, 0.8608)	0.8622 (0.8608, 0.8637)	0.7077 (0.7056, 0.7097)	0.6267 (0.6246, 0.6288)
AKI 2+	0.8181 (0.8138, 0.8224)	0.8944 (0.8916, 0.8972)	0.8935 (0.8906, 0.8963)	0.8953 (0.8926, 0.8981)	0.8907 (0.8880, 0.8935)	0.8950 (0.8923, 0.8977)	0.6905 (0.6863, 0.6948)	0.7052 (0.7010, 0.7094)

AKI 3+	0.8722 (0.8666, 0.8778)	0.9412 (0.9382, 0.9442)	0.9393 (0.9363, 0.9424)	0.9419 (0.9390, 0.9449)	0.9382 (0.9352, 0.9412)	0.9411 (0.9382, 0.9439)	0.7744 (0.7688, 0.7801)	0.7504 (0.7445, 0.7563)
AKI 3D	0.9346 (0.9258, 0.9433)	0.9628 (0.9588, 0.9668)	0.9538 (0.9487, 0.9590)	0.9606 (0.9560, 0.9652)	0.9613 (0.9563, 0.9664)	0.9698 (0.9666, 0.9729)	0.8742 (0.8651, 0.8834)	0.8382 (0.8262, 0.8503)

To explore potential factors contributing to the limited additive value of urine output in predicting AKI, we identified top features associated with prediction of AKI, number of urine output measurements and the total volume of urine output in the last 6 hours, respectively (**Supplemental Table 3.3**). While most of the top features predicting AKI are sCr-related, a noteworthy feature with significance in both the prediction of AKI prediction and urine output is the use of diuretics. Diuretics, a class of drugs facilitating salt and water excretion through urine, exhibit a close relationship with patients' urine output and contribute significantly to the predictive value in the AKI model. This association may potentially overshadow the unique contribution of urine output in predicting AKI. Additionally, a set of features that emerged as top predictors for urine output includes the frequency of vital sign measurements (e.g., temperature, pulse, respiratory rate). These observations may indicate close patient monitoring. Patients under intensive monitoring are likely at a higher risk of certain conditions, potentially contributing to an increased risk of AKI. Consequently, the marginal benefit derived from monitoring urine output in such cases may explain the limited additional value observed.

3.4 Discussion

In this study, we provide a quantitative description of UO documentation for all hospitalized patients within a large academic hospital. We found that UO is documented frequently in the

EHR, but there is evident room to improve the consistency and completeness of UO documentation. Additionally, we identified three clusters in non-ICU patients. The cluster characterized by the most frequent UO documentation exhibited the highest occurrence of AKI. Finally, we assessed the value of utilizing UO in AKI risk prediction models. It was discerned that UO, while a valuable parameter, demonstrates limited additive value when integrated into a well-established and effective AKI prediction model.

While previous research showed that UO is an important predictor to include for AKI prediction[75–77,80], our study suggests that while it may be useful in predicting AKI in the absence of other information, addition of UO information to other predictors yields only marginal predictive value. This may be attributed to variations in AKI outcomes used and the patient population studied. Unlike previous works that predominantly focused on ICU patients to predict AKI 2+, our study encompassed all hospitalized patients, predicting any AKI (AKI 1+). A recent study found that hourly UO was not significantly associated with the outcome AKI 1+ within ICU patients[32]. Our study echoed that the predictive performance of UO predictors alone for AKI 1+ was notably poor, demonstrated by an AUC of 0.63. Fluctuations in UO can be confounded by factors unrelated to AKI. The absence of UO documentation for 6 hours, as frequently observed in our study (82.8% of encounters), may not necessarily indicate kidney injury. It could be attributed to normal physiological responses[81], transient fluid balance disturbances, or gaps in checking and/or documentation.

Our study did reveal a more promising performance of UO predictors in predicting AKI 2+ and demonstrated good discriminatory power for AKI stage 3+. However, when evaluated against baseline predictors and other established predictors, considering the marginal gain in model performance, the value of incorporating UO as AKI predictors was limited.

Based on the prevailing UO documentation practices in EHR, our study suggests caution in utilizing UO to define AKI and questions the imperative of including it as a predictor for AKI prediction. Difficulties in measuring, monitoring, and accurately recording UO lead to a lack of a standardized approach to assessing changes in UO. A recent study showed that intensive monitoring of UO is associated with improved patient outcomes[82]. Further investigations into optimizing UO documentation practices and refining AKI prediction strategies in diverse clinical contexts are warranted.

While our study provides valuable insights into UO documentation in the EHR and its value in AKI prediction, several limitations should be considered. The use of data from a single large academic hospital may limit generalizability. Variations in practices, patient populations, and documentation procedures across different hospitals could impact the external validity of our results. Our study was done retrospectively using data collected from the EHR, which may be subject to inaccuracies or missing information. The retrospective nature also limits our ability to establish causation and may introduce selection bias. Furthermore, while we explored the limited additive value of UO in AKI prediction models, the design and architecture of the AKI model used could impact the generalizability of this finding. It is possible a choice of different models may yield varying results.

Despite these limitations, our findings underscore the need for ongoing efforts to enhance the consistency and completeness of UO documentation in EHR. The exploration of UO's role in predictive models highlights the complexities involved and suggests that there remains ample room to optimize AKI prediction strategies in clinical practice.

3.5 Supplemental Materials

3.5.1 Supplemental Tables

Supplemental Table 3.1 Summary of number of urine output measurements per day.

	Encounters with at least 1 measurement		**Number of measurements per day**				
	N	**%**	**min**	**mean**	**median**	**max**	**std**
Urine output	149,339	90.3	0.1	5.3	4.7	51.1	3.8
nursing	149,066	90.1	0.1	5	4.4	51.1	3.6
perioperative	44,637	27	0.1	1.1	0.8	29.6	1
Urine occurrence	125,640	76	0.1	1.7	1.3	22.9	1.5
Urine output/urine occurrence	159,823	96.7	0.1	6.3	5.7	51.1	3.6

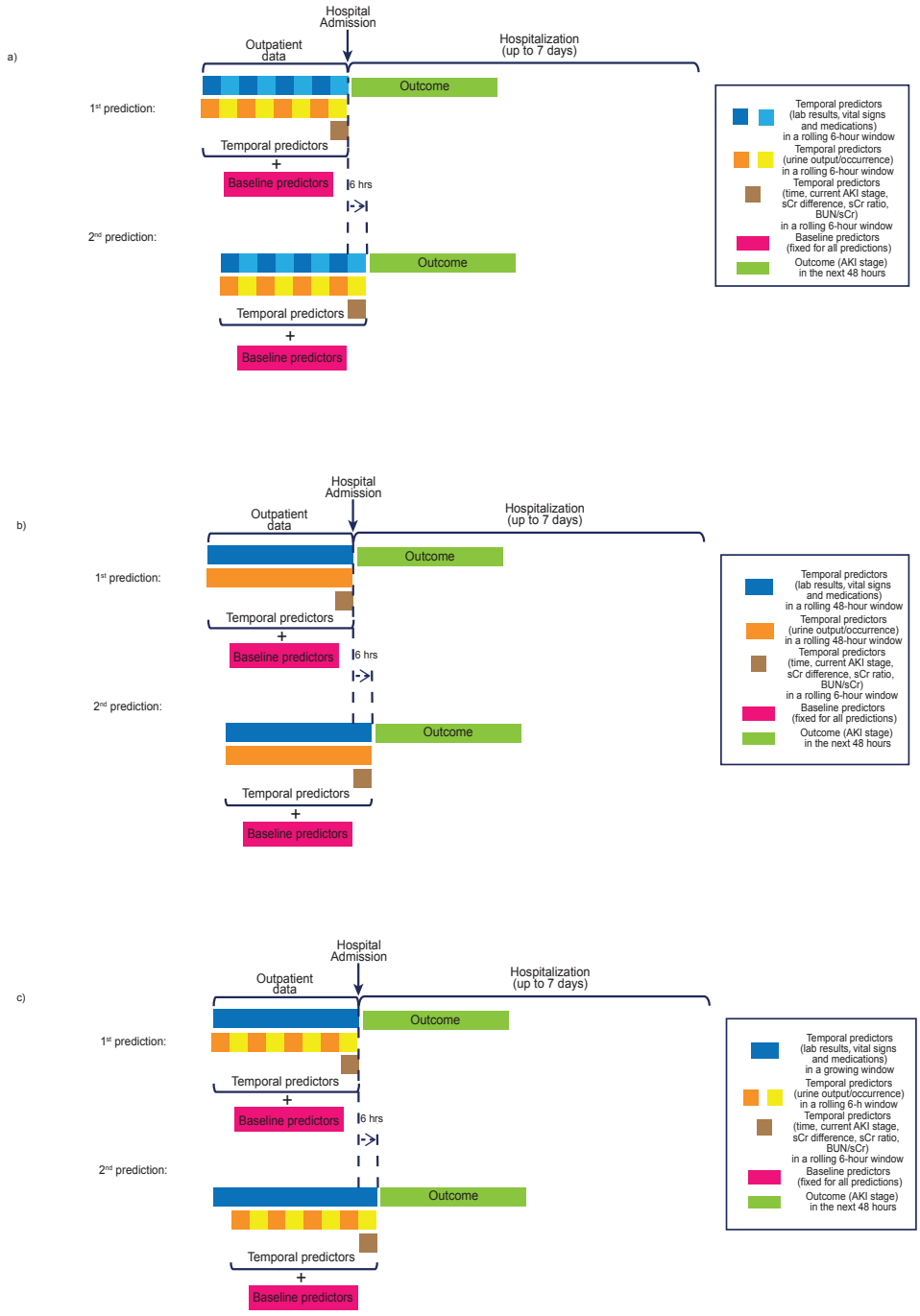
Supplemental Table 3.2 Summary of time (in hours) between urine output measurements.

	Encounters with at least 2 measurements		**Time between measurements (hours)**				
	N	**%**	**min**	**mean**	**median**	**max**	**std**
Urine output	143,196	86.6	0	3.4	2	162.4	4.7
nursing	142,714	86.3	0	3.6	2.1	162.4	4.8
perioperative	34,102	20.6	0	1.5	0.6	162.7	7.3
Urine occurrence	104,721	63.3	0.01	8.1	4	166.5	12.3
Urine output/urine occurrence	157,784	95.4	0	3.2	2	153.3	3.7

Supplemental Table 3.3 Top 20 features for models predicting AKI, number of urine output measurements, and total volume of urine output.

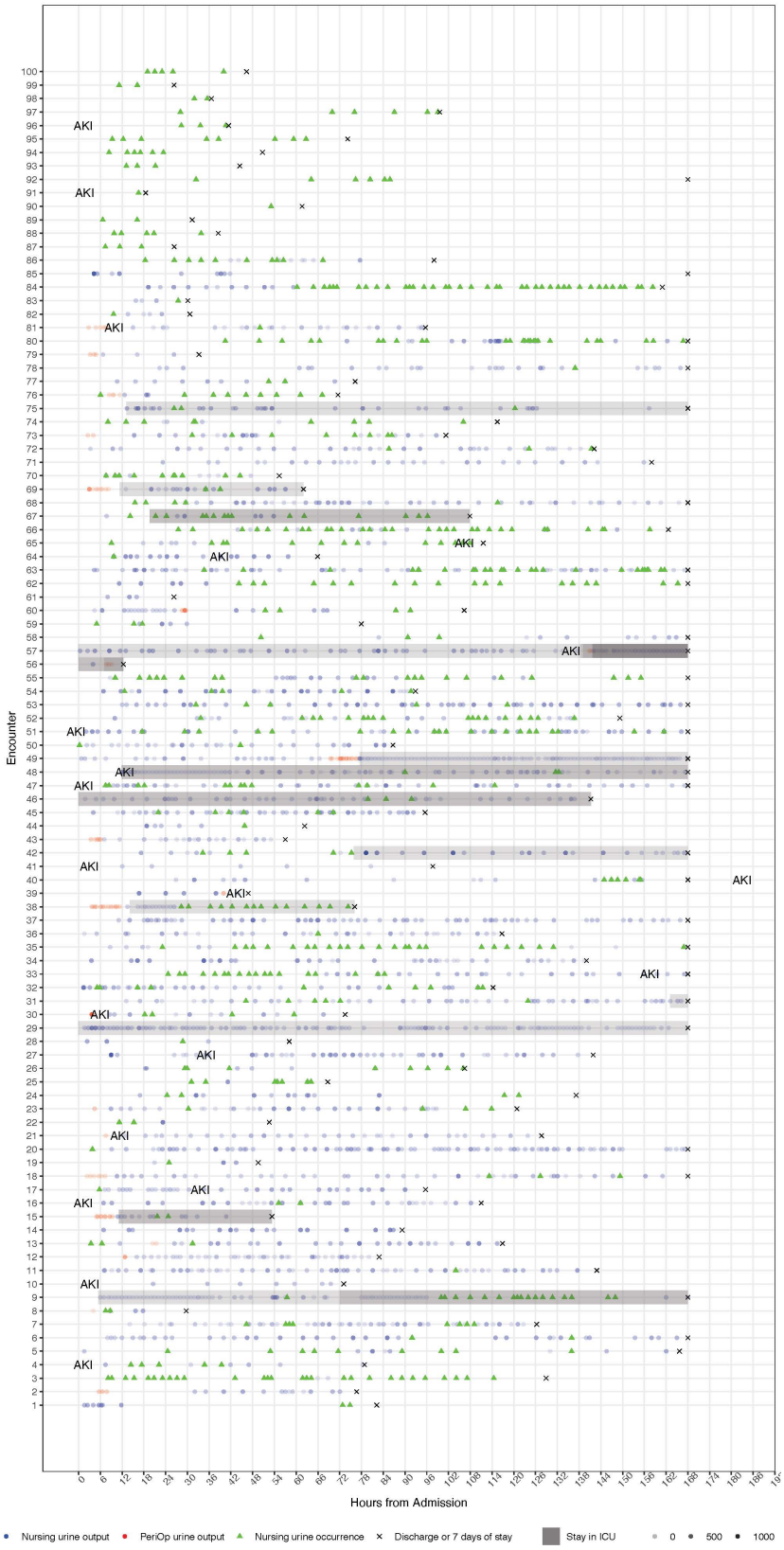
	Rolling (48h lookback, 48h window)			Growing		
Outcome	**AKI**	**No. or urine output measurements**	**Total volume of urine output**	**AKI**	**No. or urine output measurements**	**Total volume of urine output**
Feature Importance Rank						
1	Current AKI stage (67.1%)	Number of temperature measurements in the past 48 hours (26.8%)	Number of temperature measurements in the past 48 hours (29.1%)	Current AKI stage (65.6%)	Number of adrenergics administered in the past 48 hours (16.7%)	Number of diuretics administered (9.6%)
2	sCr difference from baseline (10.6%)	Number of partial pressure of oxygen in the past 48 hours (11.0%)	Number of diuretics administered in the last 48 hours (8.7%)	sCr difference from baseline (11.4%)	Hours from admission (in 6h block) (7.5%)	Number of pulse measurements (7.5%)
3	sCr ratio to baseline (2.7%)	Number of diuretics administered in the last 4 hours (5.6%)	Hours from admission (in 6h block) (8.2%)	sCr ratio to baseline (2.8%)	Number of diuretics administered (5.3%)	Hours from admission (in 6h block) (5.9%)
4	Last sCr in the past 48 hours (1.6%)	Number of pulse measurements in the past 48 hours (5.1%)	Age (3.9%)	Last sCr (1.9%)	First temperature result (3.2%)	Age (4.1%)
5	Hours from admission (in 6h block) (1.0%)	Number of respiratory rate measurements in the past 48 hours (5.1%)	Height (2.2%)	Comorbidity – chronic kidney disease (1.0%)	Number of pulse measurements (2.7%)	Number of temperature measurements (4.0%)
6	Comorbidity – chronic kidney disease (0.9%)	Number of phosphate results in the past 48 hours (4.4%)	Number of pulse measurements in the past 48 hours (2.0%)	Number of diuretics administered (0.8%)	Minimum diastolic (2.3%)	Number of inpatient weight measurements (2.1%)
7	Number of diuretics administered in the last 48 hours (0.8%)	Hours from admission (in 6h block) (4.2%)	Weight (1.6%)	Baseline sCr (0.7%)	Number of inpatient weight measurements (2.0%)	Height (2.1%)
8	Baseline sCr (0.7%)	Number of adrenergics administered in the past 48 hours (3.2%)	Number of carbon results in the past 48 hours (1.3%)	Maximum sCr (0.5%)	Maximum inpatient weight (1.8%)	Number of adrenergics administered in the past 48 hours (1.8%)
9	Minimum sCr in the past 48 hours (0.4%)	Number of diastolic pressure measurements in the past 48 hours (2.0%)	Number of potassium results in the past 48 hours (1.2%)	Comorbidity – congenital heart failure (0.3%)	Number of temperature measurements (1.7%)	First temperature (1.8%)
10	Maximum sCr in the past 48 hours (0.4%)	Height (0.9%)	BMI (1.1%)	Number of urine output measurements in the last 6 hours (0.3%)	Last respiratory rate (1.6%)	Maximum inpatient weight (1.6%)

3.5.2 Supplemental Figures



Supplemental Figure 3.1 Visual representation of three modeling strategies.

Visual representation of three different ways employed in the study to prepare features for modeling. a) Time-varying variables, including urine output/occurrence are prepared on a rolling basis, with 48-hour lookback period and windowed into 6-hour intervals. b) Time-varying variables, including urine output/occurrence are prepared on a rolling basis, with 48-hour lookback period and not windowed into smaller intervals. c) Time-varying variables (not including urine output/occurrence) are prepared in a growing manner. Urine output/occurrence are prepared on a rolling basis, with 48-hour lookback period and windowed into 6-hour intervals.



Supplemental Figure 3.2 Visual representation of urine output documentation pattern for selected hospital stays.

A visual representation of urine output documentation for selected 100 hospital stays, including the ICU status and AKI outcomes. Types of urine output measurements are differentiated by color (blue dot: nursing urine output, red dot: perioperative urine output, green triangle: nursing urine output). The intensity of the dots corresponds to the volume of the documented urine output. Patterns are presented from admission to discharge or up to 7 days following admission. ICU stays are highlighted by shaded grey areas. AKI outcome is determined based on KDIGO criteria, but only serum creatinine is used in the definition.

Chapter 4 Participation in Multicenter Prediction Modeling to Improve Generalizability Across a National Research Network

4.1 Introduction

A key requirement for the responsible use of artificial intelligence (AI) is to ensure that such models generalize to the clinical settings in which they are intended to be used[65–67]. Failure of models to generalize may present to varying degrees due to a lack of sufficient sample size or diversity,[34,86] or due to differences in underlying patient populations, health behaviors, or technology between settings where models were trained versus deployed[87,88]. Building models with data from multiple centers is an effective way to increase the sample size and potentially the sample diversity, which may enable generalizability to a broader range of centers exhibiting similar diversity. Multicenter modeling is typically accomplished by pooling data across multiple centers, often through a data coordinating center. However, training and validating models on pooled data is problematic for two reasons. First, a single model trained on pooled data may not effectively capture risk when there is substantial heterogeneity across centers, leading to varying model performance and the potential for decreased generalizability. Second, even if pooling data is expected to produce better models, it may not always be possible due to ethical or legal concerns related to patient privacy and data security[89,90]. Training multicenter models without pooling data is possible due to recent advances in federated learning, where models share parameters across multiple centers without directly sharing patient data[91]. Federated learning is becoming more common in clinical research in the prediction of intensive care unit (ICU)

outcomes,[92,93] COVID-19 diagnosis and outcomes,[94–96] breast density classification,[97] and rare cancer boundary detection[98]. However, its adoption in clinical settings has been limited. Despite recognizing the importance of generalizable models, large health systems may not see a clear reason to participate in multicenter modeling efforts, either through pooling data or federated learning, because they often have a sufficiently large and representative dataset to train models that are robust in their local setting[85,99]. In contrast, smaller centers lack the ability to develop single-center prediction models due to an insufficient sample size and lack of methodological expertise. Multicenter research networks may help smaller centers overcome these limitations by bringing the benefits of large and diverse datasets to all participating centers. However, the degree to which multicenter research networks improve generalizability across individual centers is not known. Network participation may disproportionately benefit smaller centers through their gaining access to larger datasets.

In this study, we used data from a national perioperative network comprising 31 academic and community hospitals across a wide range of sizes to investigate three important generalizability issues related to participation in multicenter AI modeling efforts. We sought to: 1) compare clinical prediction model performance of single-center approaches with pooled-data and federated-learning approaches; 2) quantify differences in performance for models trained using federated learning versus those trained using pooled data; and 3) characterize how the optimal strategy may change based on the size of a hospital network. We answered these questions through a representative clinical scenario: predicting cardiac surgery-associated acute kidney injury (AKI).

4.2 Methods

4.2.1 Study Design

We conducted a retrospective study using data from 31 hospitals within the Multicenter Perioperative Outcomes Group (MPOG) network, a national perioperative research and quality improvement network[100], to examine the optimal modeling strategy for hospitals in the prediction of AKI. We compared a single-center predictive modeling strategy against two strategies that require participation in a research network – pooling data and federated learning. We followed the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) guidelines for conducting and reporting the findings from this study[101]. Institutional review board approval (HUM00209313) was obtained for this observational study and patient consent was waived. An a priori study protocol for inclusion criteria, data collection and handling, and statistical methods was approved and registered by the MPOG network's Perioperative Clinical Research Committee.

4.2.2 Study Population

MPOG collects perioperative data from academic and community hospitals across 23 states in the United States. Methods for extraction of local electronic health record (EHR) data, validation, mapping to semantically interoperable concepts, and secure transfer to the MPOG data coordinating center have been previously described and used in multiple published studies[100,102–104].

Open cardiac surgical procedures using cardiopulmonary bypass performed on adult patients at US institutions from January 1, 2014 to February 1, 2022 were eligible for this study. Cases without preoperative (within 180 days) or postoperative (within 7 days) creatinine laboratory values, not meeting minimum data quality standards (**Supplemental Methods**), or from institutions contributing less than 20 cases annually meeting eligibility criteria, were also excluded. Finally, patients with pre-existing severe chronic kidney disease were excluded (Stage

4 or 5 based on an estimated glomerular filtration rate [eGFR] 15-29 or <15 mL/min/1.73m2 respectively); eGFR was computed using the baseline creatinine applied to the 2021 updated creatinine-based race-neutral equation[105]. For patients undergoing repeated cardiac surgical procedures meeting the above inclusion criteria, only the index case was used.

The cardiac case volume for each center was calculated as the number of cardiac surgery cases contributed by the center to the Multicenter Perioperative Outcomes Group (MPOG) Network. Because not all cases may have been captured in the MPOG database, our calculated value may underestimate the true case volume.

4.2.3 Predictor Variables

We collected preoperative patient and surgical characteristics and time-varying intraoperative and immediate postoperative measures for each case. Patient characteristics included demographics, anthropometrics, comorbidities, preoperative laboratory values and vital signs, home medications, American Society of Anesthesiologists Physical Status classification, the baseline kidney function including eGFR and presence of preoperative AKI (as defined in the **Supplemental Methods**), and the first postoperative serum creatinine[33] (within 24 hours). Surgical characteristics included emergent versus non-emergent, surgical procedure type (non-mutually exclusive), anesthesiology staffing model (presence of resident, nurse anesthetist, both, or neither with solely anesthesiology attending), weekday versus weekend start time, and academic versus community hospital.

Intraoperative time-varying variables consisted of arterial blood gas values, physiologic monitors (systolic/mean/diastolic arterial blood pressure, central venous pressure, oxygen saturation, and heart rate), and intravenous cardiovascular medications administered intraoperatively (**Supplemental Table 4.1**). Given the dynamic nature of patient physiology surrounding

initiation and separation from cardiopulmonary bypass (CPB), summary statistics for intraoperative variables were separately calculated as candidate predictors within models from each of three distinct phases: pre-CPB, intra-CPB, and post-CBP. More details can be found in **Supplemental Table 4.1**.

4.2.4 Outcome: Cardiac Surgery Associated AKI

AKI was defined based upon the maximum sCr level recorded between 2 and 7 days after the procedure. AKI was then defined and staged for severity according to the KDIGO international guidelines: no AKI, AKI stage 1, AKI stage 2, and AKI stage 3[16]. While our models were trained using this multinomial outcome, results reported by AKI stages were grouped into binary outcomes according to the level of severity. For example, AKI stage 1+ refers to any AKI stage, and AKI stage 2+ refers to AKI stage 2 and stage 3.

4.2.5 Development and Validation of Single-Center, Pooled, and Federated Models

After setting aside four centers for external validation (two randomly selected academic hospitals and two randomly selected community hospitals), we divided the remaining 27 institutions into a training set (January 1, 2014 to February 29, 2020) and a temporal validation set (March 1, 2020 to February 1, 2022) based on the timing of elective case scheduling changes induced by the COVID-19 pandemic[106]. Because not all hospitals had eligible cases during the entire time period, some hospitals selected for temporal validation were only included in the training set or in the temporal validation set. A visual representation of the data split is shown in **Figure 4.1**.

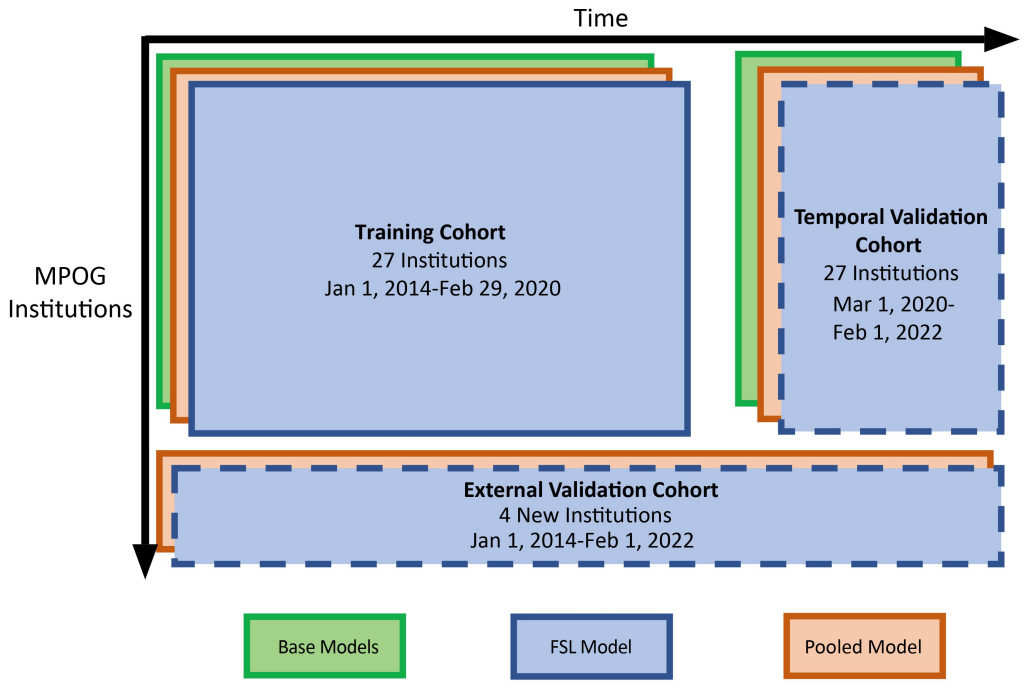


Figure 4.1 Visual representation of study data split.

Among all 31 centers participated in the study, four centers were set aside for external validation. The remaining 27 centers were split into a training set and a temporal validation set based on the timing of elective case scheduling changes induced by the COVID-19 pandemic. Single-center (base) models and multicenter models (pooled and federated) were trained using the training set. Base models were only tested on temporal validation set, while pooled and federated models were tested on both temporal validation and external validation sets.

4.2.5.1 Single-center (Base) Models

While single-center (or base) models may not generalize from one hospital to another, the use of models both trained and deployed at the same institution are growing and may offer predictive advantages regarding capture of center-specific characteristics when the goal of the model is limited to use at that specific site. As electronic health record (EHR) vendors tailor models to individual institutions[107], larger health systems may opt out of multicenter AI modeling efforts in favor of individually tailored models.

To evaluate this approach, gradient-boosted decision tree (GBDT) models were separately trained and evaluated for each of the hospitals present in both the training and temporal

validation sets. We opted for GBDT models because of their high empirical performance in prior applications to AKI prediction[31,86]. Details for the training and early stopping are presented in the **Supplemental Methods**. Centers in the external validation set were excluded from this evaluation.

4.2.5.2 Pooled Model

To assess the value of pooling data across multiple centers, we trained a pooled GBDT model in the training set and evaluated its performance in both the temporal and external validation cohorts. In contrast to the single-center models, the pooled model was evaluated in centers either absent from the training set entirely or with too few patients in the training set for a model to be adequately trained using data from that center only ($n < 20$ total cases, or no cases meeting an outcome definition).

4.2.5.3 Federated Model

We assessed the value of a federated learning approach by implementing a novel federated stacked learning (FSL) framework, which uses a two-stage training process based upon work developed for model stacking[108,109]. In the first stage, base models are trained at each center and placed on a central server to be shared with all centers. In the second stage, predictions from all base models are used to train the final meta-model. A visual representation of the FSL algorithm is shown in **Figure 4.2a**. Each center first partitions its data randomly into training, weighting, and testing (if model evaluation is desired) sets. In the first (base model building) stage, the following actions are taken: a) each site uses its own training set to train a base model; b) each site sends its base model to the central server; and c) once the central server has all base models, it sends the collection of base models to all sites. Upon receiving all base models, the second (meta-model building) stage starts and follows these steps: a) each site applies all base models to

its own weighting set to generate its weighting predictions; b) each site sends its weighting predictions (one number per patient) to the central server; c) the central server learns a meta model using the weighting predictions, and d) once the meta model is learned, the central server sends the meta model back to all sites. The meta-model is used to weight the predictions generated by different base models. To facilitate understanding of how data structure changes in FSL algorithms, **Figure 4.2b** highlights which data are used in each training stage. In contrast to existing federated learning approaches, FSL requires fewer rounds of model sharing across centers and is thus simpler to implement as centers are added or removed from the network.

Our FSL algorithm also allows center-level metadata to be added when training the meta-model. The value remains the same for cases from the same site. When each site generates its weighting predictions, a column indicating whether the center is an academic hospital and three additional columns showing their rate of different AKI outcomes (AKI stage 1 rate, AKI stage 2 rate, and AKI stage 3 rate) calculated from the weighting data are added to the weighting prediction dataset and used to train the meta-model.

In our study, we used GBDT to train both the base models and the meta-model. We trained and compared two federated models: one using patient data only and one additionally incorporating center-level metadata. The FSL model with center-level metadata was trained using both university affiliation of the sites and site-level AKI rates. However, when it was evaluated on the external validation set, only university affiliation was available and the AKI rates of the four held-out hospitals were set to missing.

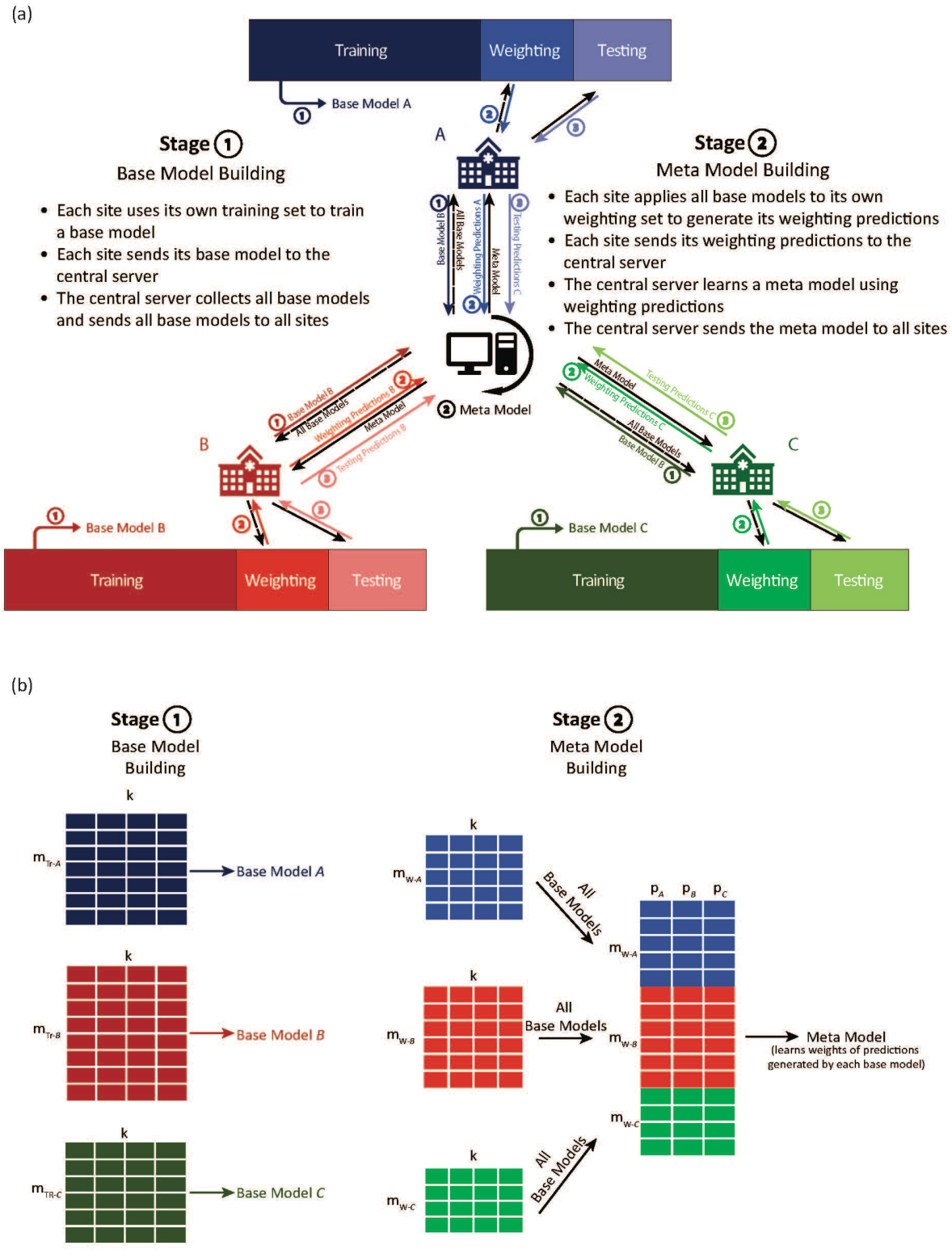


Figure 4.2 Visual representation of the FSL algorithm.

(a) Data exchanges between centers associated with each stage of FSL algorithm. (b) Changes of data structure used in each stage of FSL algorithm.

4.2.6 Model Evaluation

Model discrimination was assessed using the area under the receiver operating characteristic curve (AUC). A separate AUC was reported for individuals at-risk for each AKI stage. For example, patients without any AKI prior to surgery were evaluated on their risk of developing any AKI (i.e., stage 1 or greater), and patients with no AKI or AKI stage 1 were evaluated on their risk of developing AKI stage 2 or greater, and so on. The 95% confidence intervals were generated using DeLong's method[58]. Model calibration was evaluated by comparing deciles of predicted probabilities with observed risks for temporal and external validation sets, for all AKI outcome severities.

Models were evaluated in aggregate (across all centers in both validation sets) and then individually for each hospital. For each hospital, we determined the optimal modeling strategy by comparing the performance of single-center (base) models against pooled and federated models. The individual hospital analysis was performed using the temporal validation set only because there were no base models available for use in the external validation sets.

4.2.7 Studying the Role of Network Size in Resulting Model Performance

Building AI models as part of a research network may involve more investment as the network expands, but larger networks may produce more stable and generalizable models due to greater sample size and diversity. To examine the role of network size on model performance, we performed a learning curve analysis in which we compared the performance of pooled and federated models predicting AKI 1+ by varying the size of the network from 1 hospital (no

network) up to 23 hospitals. For each network size, hospitals were randomly selected without replacement, and this process was repeated 100 times.

4.2.8 Feature Importance

To develop an understanding of which variables most strongly contributed to the predictive performance of models developed, evaluated the feature importance. Feature importance of variables within the pooled model were assessed using each feature's squared influence within the GBDT algorithm aggregated over the tree ensemble. Feature importance for the pooled model is provided in **Supplemental Figure 4.1**.

4.2.9 Software

All data processing and analyses were performed using R 4.2.1[61]. Transformation of time-series data and calculation of summary statistics were performed using the Grammar of Prediction (gpmodels) R package[62]. H2o version 3.38.0.1 was used to fit all GBDT models, including the pooled model, base models and the meta model of the FSL model. **Figure 4.5** with axis breaks was prepared using the ggbreak R package[110,111].

4.3 Results

4.3.1 Cohort Characteristics

We identified a total of 66,166 cardiac surgery cases across 31 US academic and community hospitals meeting inclusion criteria (**Figure 4.3**). After applying inclusion criteria, 43,926 cases across 23 hospitals (n = 43,926) were included in the training set, 18,132 across 25 hospitals in the temporal validation set, and 4,108 cases across 4 hospitals in the external validation set (see **Supplemental Table 4.2** for details). Our overall cohort had a mean age of 62.0±13.5 years old

and consisted of more males (68.4%) and non-Hispanic whites (79.6%) (**Table 4.1**). Surgical cases had a mean duration of 6.98±2.22 hours in the operating room and 2.33±1.47 hours of CPB. The temporal validation set had similar baseline characteristics as the training set, except for more non-smokers (29.1% vs 18.6%) and a higher burden of comorbidities (**Supplemental Table 4.3**). As compared to the training set, the external validation cases were older (mean age 64.4±12.3 vs 61.9±13.6), were less diverse in regard to sex, race, and ethnicity (males 72.4% vs 68.0%, non-Hispanic whites 84.9% vs 79.3%), had fewer comorbidities, and had shorter operating room duration (6.19±2.02 hours vs 6.95±2.22 hours).

Among all cases in the overall cohort, 25.5% developed any AKI: 24.8% in the training set, 28.8% in temporal validation, and 18.6% in external validation. The median institution-level AKI was 29.7%, with an interquartile range of 23.6 to 34.0%. Prior to surgery, 5.0% of included cases had preoperative AKI, which was similar across the cohorts.

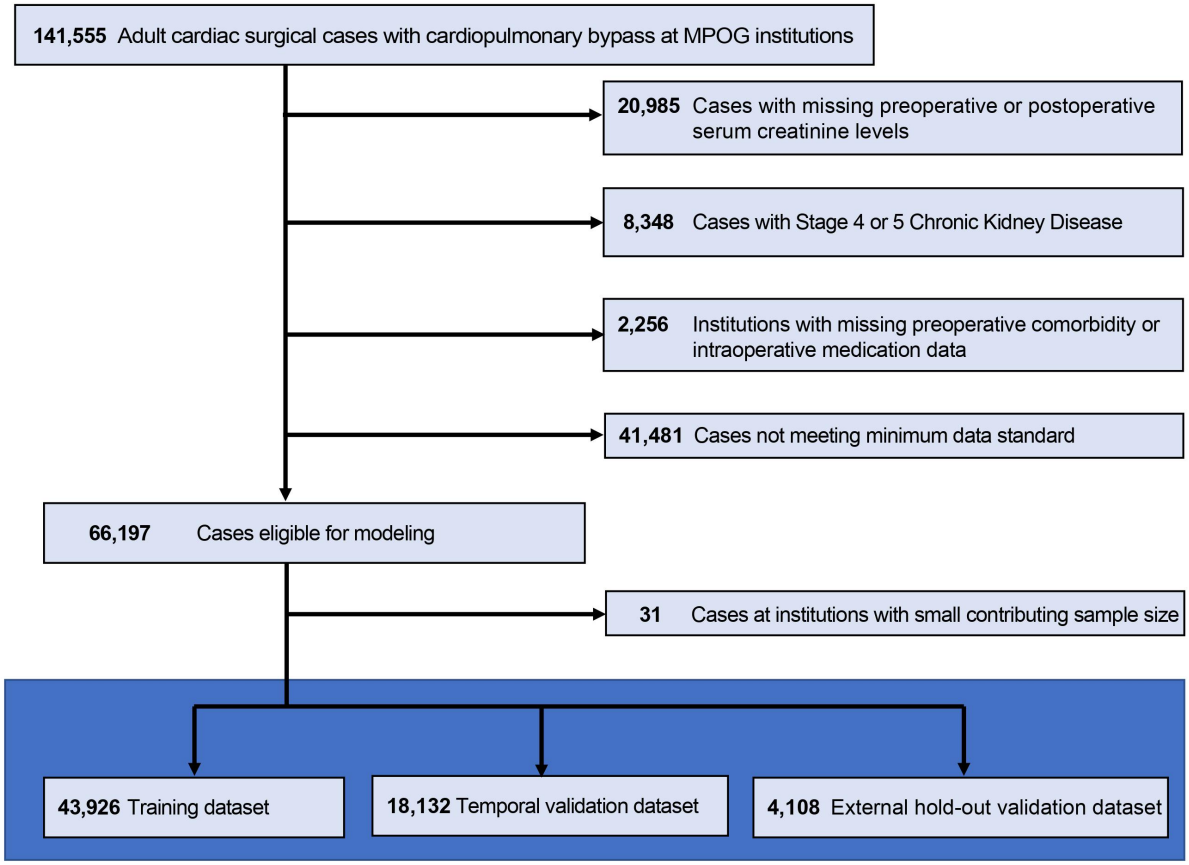


Figure 4.3 Study flow diagram.

The flow diagram shows the case inclusion and exclusion criteria for the analysis of the study. Number of remaining cases after each step are shown. Numbers of cases included in each data split for subsequent training and evaluation are also shown.

Table 4.1 Patient and surgical characteristics.

Characteristic	Overall (N = 66,166)	Training (N = 43,926)	Temporal Validation (N = 18,132)	External Validation (N = 4,108)
Preoperative Patient Characteristics				
Age (years)	62.0 (13.5)	61.9 (13.6)	61.7 (13.3)	64.4 (12.3)
Sex				
Female	20,921 (31.6%)	14,062 (32.0%)	5,726 (31.6%)	1,133 (27.6%)
Male	45,245 (68.4%)	29,864 (68.0%)	12,406 (68.4%)	2,975 (72.4%)
Race / Ethnicity				
White not of hispanic origin	52,643 (79.6%)	34,822 (79.3%)	14,335 (79.1%)	3,486 (84.9%)

Black not of hispanic origin	4,304 (6.5%)	2,720 (6.2%)	1,494 (8.2%)	90 (2.2%)
Asian or Pacific Islander	2,068 (3.1%)	1,215 (2.8%)	624 (3.4%)	229 (5.6%)
Bi or Multi Racial	569 (0.9%)	413 (0.9%)	156 (0.9%)	0 (0.0%)
American Indian or Alaska Native	180 (0.3%)	102 (0.2%)	66 (0.4%)	12 (0.3%)
Hispanic white	514 (0.8%)	267 (0.6%)	206 (1.1%)	41 (1.0%)
Hispanic black	38 (0.1%)	18 (0.0%)	19 (0.1%)	1 (0.0%)
Middle Eastern	38 (0.1%)	38 (0.1%)	0 (0.0%)	0 (0.0%)
Missing	5,812 (8.8%)	4,331 (9.9%)	1,232 (6.8%)	249 (6.1%)
Height (cm)	171.9 (14.1)	172.5 (10.9)	170.2 (19.9)	172.4 (10.4)
Weight (kg)	87.0 (20.8)	86.7 (20.7)	87.7 (21.2)	87.2 (21.0)
Body Mass Index (kg/m^2)	29.1 (6.2)	29.0 (6.2)	29.2 (6.3)	29.3 (6.3)
ASA Physical Status Classification				
ASA Class 1	59 (0.1%)	54 (0.1%)	3 (0.0%)	2 (0.0%)
ASA Class 2	397 (0.6%)	267 (0.6%)	72 (0.4%)	58 (1.4%)
ASA Class 3	13,838 (20.9%)	8,806 (20.0%)	3,130 (17.3%)	1,902 (46.3%)
ASA Class 4	50,982 (77.1%)	34,286 (78.1%)	14,614 (80.6%)	2,082 (50.7%)
ASA Class 5	890 (1.3%)	513 (1.2%)	313 (1.7%)	64 (1.6%)
Preoperative Laboratory Values				
White Blood Cell Count (per mL)	7.6 (3.1)	7.5 (3.1)	7.7 (3.1)	7.9 (3.4)
Platelet Count, (K/mL)	218.6 (72.2)	216.9 (70.2)	221.9 (77.2)	223.0 (70.4)
Hemoglobin (g/dL)	13.3 (2.0)	13.3 (2.0)	13.2 (2.1)	13.3 (1.9)
Sodium (mEq/L)	138.8 (3.2)	139.1 (3.2)	138.3 (3.1)	138.5 (3.1)
Potassium (mEq/L)	4.2 (0.4)	4.2 (0.4)	4.2 (0.4)	4.1 (0.4)
Bicarbonate (mmol/L)	25.6 (3.1)	25.8 (3.1)	25.3 (3.3)	24.8 (2.8)
Glucose (g/dL)	115.8 (39.3)	115.1 (39.6)	116.5 (38.7)	119.0 (38.0)
Creatinine-Related Variables				
Preoperative Baseline Serum Creatinine, g/dL	1.0 (0.3)	1.0 (0.3)	0.9 (0.3)	0.9 (0.3)
Preoperative Most Recent Serum Creatinine, g/dL	1.0 (0.5)	1.0 (0.6)	1.0 (0.3)	1.0 (0.3)

Preoperative Serum Creatinine Ratio (Most Recent/Baseline)	1.1 (0.5)	1.1 (0.6)	1.1 (0.2)	1.1 (0.2)
Preoperative Serum Creatinine Difference (Most Recent - Baseline)	0.1 (0.5)	0.1 (0.5)	0.1 (0.1)	0.1 (0.1)
First Post-operative Serum Creatinine Within 24h	1.0 (0.3)	1.0 (0.3)	1.0 (0.3)	0.9 (0.3)
Preoperative AKI				
No Preoperative AKI	62,831 (95.0%)	41,831 (95.2%)	17,081 (94.2%)	3,919 (95.4%)
Preoperative AKI-1	3,066 (4.6%)	1,934 (4.4%)	967 (5.3%)	165 (4.0%)
Preoperative AKI-2	218 (0.3%)	136 (0.3%)	66 (0.4%)	16 (0.4%)
Preoperative AKI-3	51 (0.1%)	25 (0.1%)	18 (0.1%)	8 (0.2%)
Summary Patient Comorbidities (Elixhauser)				
Cardiac Arrhythmia	42,934 (64.9%)	27,245 (62.0%)	12,912 (71.2%)	2,777 (67.6%)
Chronic Pulmonary Disease	15,116 (22.8%)	10,143 (23.1%)	4,008 (22.1%)	965 (23.5%)
Coagulopathy	25,473 (38.5%)	15,989 (36.4%)	8,898 (49.1%)	586 (14.3%)
Congestive Heart Failure	31,495 (47.6%)	19,952 (45.4%)	9,910 (54.7%)	1,633 (39.8%)
Diabetes	18,570 (28.1%)	11,661 (26.5%)	5,321 (29.3%)	1,588 (38.7%)
Fluid and Electrolyte Disorders	39,655 (59.9%)	25,740 (58.6%)	12,602 (69.5%)	1,313 (32.0%)
Hypertension	24,791 (37.5%)	14,337 (32.6%)	8,981 (49.5%)	1,473 (35.9%)
Liver Disease	4,758 (7.2%)	2,850 (6.5%)	1,609 (8.9%)	299 (7.3%)
Peripheral Vascular Disorders	24,588 (37.2%)	15,781 (35.9%)	7,557 (41.7%)	1,250 (30.4%)
Pulmonary Circulation Disorders	12,034 (18.2%)	7,682 (17.5%)	3,632 (20.0%)	720 (17.5%)
Valvular Disease	46,092 (69.7%)	31,145 (70.9%)	12,513 (69.0%)	2,434 (59.3%)
Surgical Characteristics - Procedure Type				
Valve Only	21,670 (32.8%)	15,371 (35.0%)	5,232 (28.9%)	1,067 (26.0%)
Coronary Artery Bypass Only	18,573 (28.1%)	11,350 (25.8%)	5,356 (29.5%)	1,867 (45.4%)
Aortic	9,316 (14.1%)	6,114 (13.9%)	2,852 (15.7%)	350 (8.5%)
Valve + Coronary Artery Bypass Only	6,455 (9.8%)	4,433 (10.1%)	1,492 (8.2%)	530 (12.9%)
Myectomy	2,156 (3.3%)	1,588 (3.6%)	541 (3.0%)	27 (0.7%)
Ventricular Assist Device	1,680 (2.5%)	1,161 (2.6%)	463 (2.6%)	56 (1.4%)
Heart Transplant	1,669 (2.5%)	953 (2.2%)	664 (3.7%)	52 (1.3%)

Pulmonary Thromboendarterectomy	385 (0.6%)	242 (0.6%)	143 (0.8%)	0 (0.0%)
Other	4,264 (6.4%)	2,716 (6.2%)	1,389 (7.7%)	159 (3.9%)
Additional Surgical Characteristics				
Anesthesia Duration (min)	419 (133)	417 (133)	434 (134)	371 (121)
Cardiopulmonary Bypass Duration (min)	140 (88)	133 (85)	157 (99)	137 (65)
Emergent	4,059 (6.1%)	2,613 (5.9%)	1,165 (6.4%)	281 (6.8%)
Institutional Characteristics				
Academic Hospital	64,920 (98.1%)	43,198 (98.3%)	18,111 (99.9%)	3,611 (87.9%)
Outcome Characteristics				
CSA-AKI Stage				
No CSA-AKI	49,264 (74.5%)	33,019 (75.2%)	12,901 (71.2%)	3,344 (81.4%)
CSA-AKI-1	11,759 (17.8%)	7,771 (17.7%)	3,449 (19.0%)	539 (13.1%)
CSA-AKI-2	3,457 (5.2%)	2,161 (4.9%)	1,142 (6.3%)	154 (3.7%)
CSA-AKI-3	1,686 (2.5%)	975 (2.2%)	640 (3.5%)	71 (1.7%)
** Non-mutually exclusive*				

Statistics presented as mean (SD) for numeric variables; N(%) for categorical variables. AIDS/HIV = acquired immunodeficiency syndrome / human immunodeficiency virus; AKI = acute kidney injury; ASA = American Society of Anesthesiologists; CPB = cardiopulmonary bypass; CSA-AKI = cardiac surgery-associated acute kidney injury; ETT = endotracheal tube; LMA = laryngeal mask airway

4.3.2 Aggregate Model Performance

In the temporal validation set, the pooled models demonstrated the highest AUCs for all AKI severity levels (AKI 1+: 0.856; AKI 2+: 0.890; AKI 3+: 0.911), while single-center base models had a lowest AUCs (AKI 1+: 0.770; AKI 2+: 0.796; AKI 3+: 0.821) (full results in **Table 4.2**). Federated model AUCs were approximately 0.03 lower than the pooled models (AKI 1+: 0.826; AKI 2+: 0.861; AKI 3+: 0.887), although the differences were attenuated by the inclusion of center-level metadata.

In the external validation set, single-center model performance was not calculated because the external validation centers were excluded from the training set. The pooled models once again demonstrated the highest AUCs (AKI 1+: 0.882; AKI 2+: 0.925; AKI 3+: 0.950). The federated model AUCs were approximately 0.02 lower (AKI 1+: 0.865; AKI 2+: 0.906; AKI 3+: 0.933), and the inclusion of metadata did not substantially change the results.

The FSL models, with and without metadata, and the pooled model were generally well-calibrated in both temporal and external validation for all AKI severities (**Figure 4.4**). The first postoperative serum creatinine was the most important variable (**Supplemental Figure 4.1**).

Table 4.2 Temporal and external validation AUCs.

	Temporal Validation			
	Base Model*	**FSL Model**	**FSL with Metadata** Model**	**Pooled Model**
AKI-1+	0.7703 (0.7609, 0.7796)	0.8261 (0.8190, 0.8332)	0.8317 (0.8248, 0.8386)	0.8557 (0.8493, 0.8621)
AKI-2+	0.7960 (0.7824, 0.8096)	0.8610 (0.8521, 0.8698)	0.8651 (0.8566, 0.8736)	0.8899 (0.8823, 0.8974)
AKI-3+	0.8214 (0.8016, 0.8411)	0.8873 (0.8744, 0.9001)	0.8938 (0.8819, 0.9057)	0.9108 (0.9004, 0.9211)
Note: * Centers where a Base Model could not be trained are excluded from this temporal validation set. ** Metadata used: university affiliation, AKI stage 1 rate, AKI stage 2 rate and AKI stage 3 rate at each site.				
	External Validation			
	Base Model	**FSL Model**	**FSL with Metadata*** Model**	**Pooled Model**
AKI-1+	NA	0.8647 (0.8496, 0.8798)	0.8623 (0.8472, 0.8774)	0.8824 (0.8685, 0.8964)
AKI-2+	NA	0.9063 (0.8887, 0.924)	0.9012 (0.8829, 0.9196)	0.9249 (0.9087, 0.9412)
AKI-3+	NA	0.9333 (0.904, 0.9626)	0.9312 (0.8994, 0.9631)	0.9496 (0.9287, 0.9705)
Note: *** Metadata used: university affiliation				

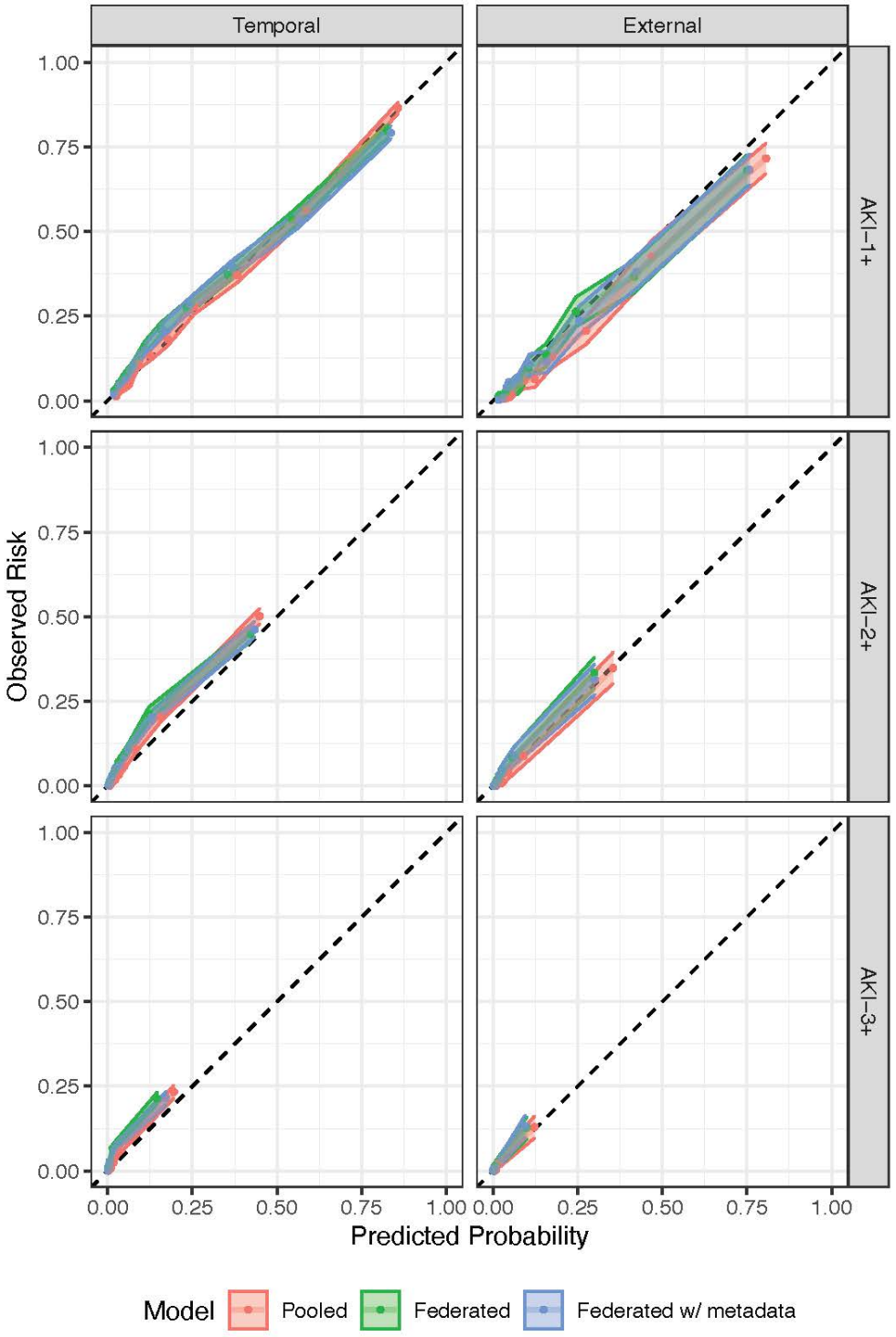


Figure 4.4 Calibration plot of the multicenter models.

The calibration of the multicenter models in temporal validation and external validation, for all AKI severities. The predicted probabilities (deciles) are plotted against the observed probabilities with 95% confidence intervals. The diagonal line demonstrates the ideal calibration. The model calibration is examined for pooled model (red), FSL model without center-level metadata (green), and FSL model incorporating center-level metadata (blue).

4.3.3 Model Performance at Individual Centers

While we found that the pooled model performs better in aggregate in both validation sets, not all centers may benefit equally from participating in a network, either through pooling or federating. To examine the potential benefits to individual centers in building AI models as part of a network, we evaluated each hospital's optimal modeling strategy using the temporal validation set for each center (**Supplemental Table 4.4** and **Supplemental Figure 4.2**). Among the 23 centers in the validation set, using a single-center model was the optimal strategy (based on the AUC point estimate) for none of the hospitals in predicting either AKI 1+, 2+, or 3+. Even if pooling data were not an option due to data sharing restrictions, single-center base models outperformed a simple FSL approach (without metadata) for only 1 hospital for AKI 1+, 5 hospitals for AKI 2+, and 4 hospitals for AKI 3+ (**Figure 4.5**). The single-center approach generally only performed better in the largest hospitals, and the magnitude of difference was small.

Five out of these 23 centers did not have a sufficient number of cases in the training set to even train a base model for AKI 1+ using the number of predictors in our model. Despite this limitation, 4 of these 5 centers achieved an AUC > 0.80 for AKI 1+ when pooling data, and 3 achieved an AUC > 0.80 when applying a federated learning approach.

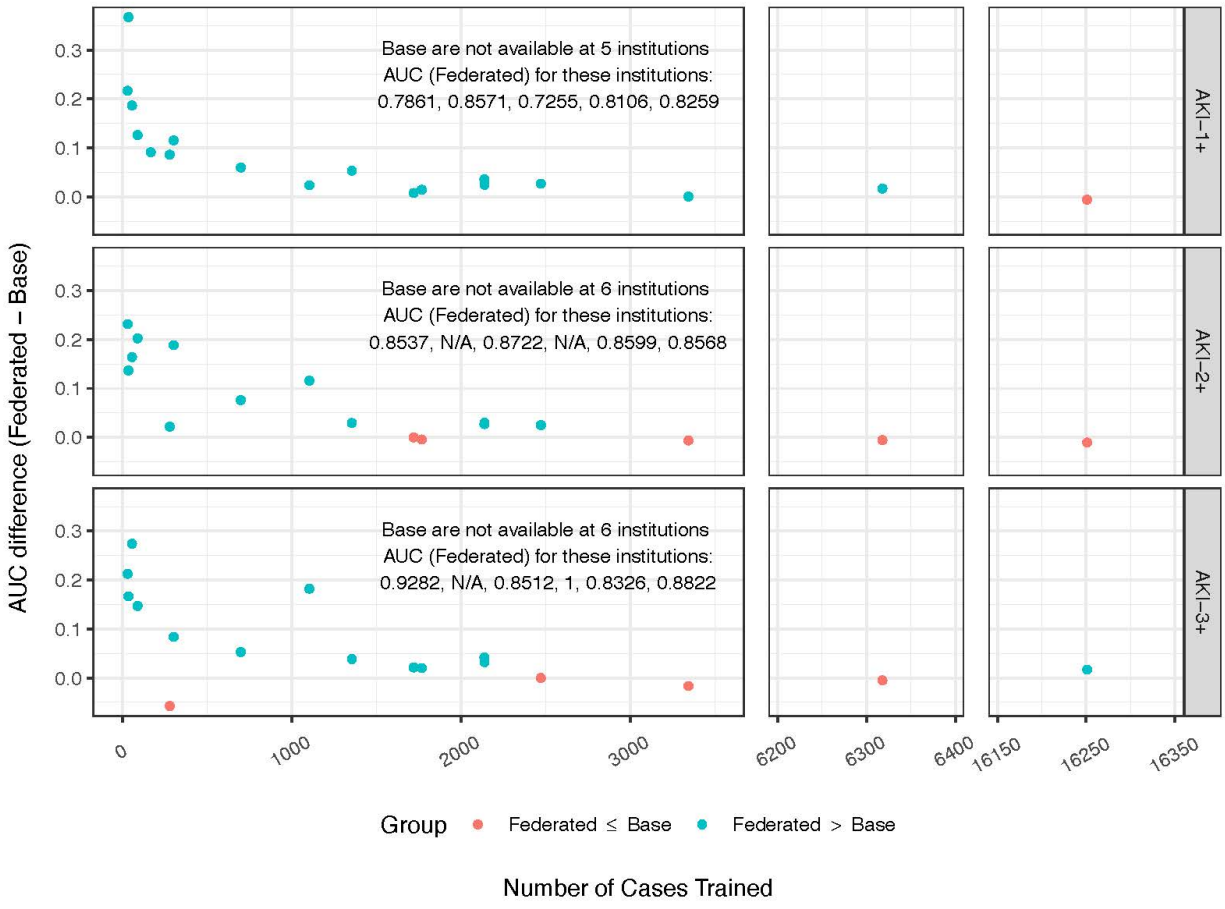


Figure 4.5 AUC difference between FSL and base models.

Difference in AUC between FSL (without metadata) and base models at each center in temporal validation, for all AKI severities. Better performing model is indicated by color (red: single-center base model, green: FSL model).

4.3.4 Network Size and Model Performance

While our results derive from a large national network of 31 hospitals, the value derived from building multicenter AI models may vary based on a network size. We found that the lowest AUC in both the temporal validation and external validation cohorts was for a network size of 1 (no network) and the highest AUC was for a network size of 23 (the full network available for study, **Figure 4.6** and **Supplemental Table 4.5**). The AUC increased the most with the addition of the first few hospitals, and the magnitude became smaller as the final few hospitals were

added to the network. About half of the increase in AUC from a single-center to the full 23-hospital models was observed with the models learning from only 4 hospitals.

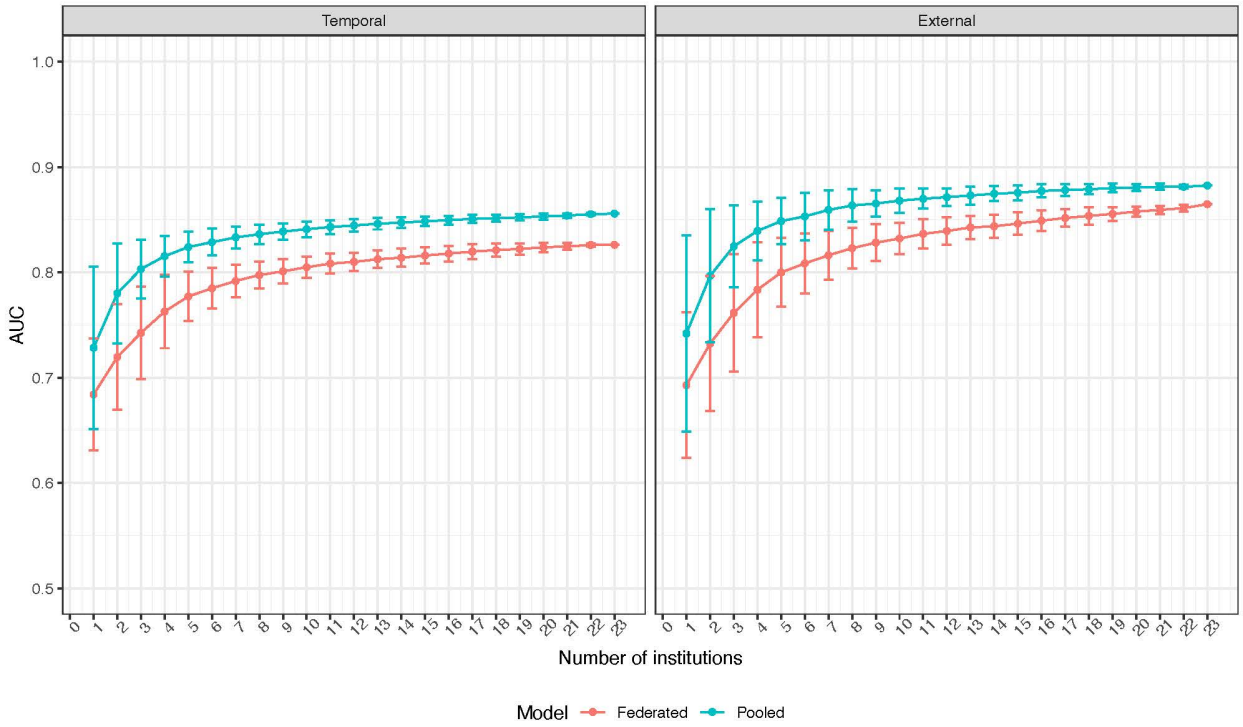


Figure 4.6 Learning curve of multicenter model performance in predicting AKI 1+ as network expands.

Changes in model performance (AUC) in predicting AKI 1+ as the network size increases from none (no network) to 23 hospitals. Results from both temporal validation and external validation are shown. Different multicenter modeling approaches are compared (red: FSL, green: pooled). For each network size, hospitals were randomly selected without replacement, and this process was repeated 100 times. The center dot and error bar at each network size represents the mean AUC and standard deviation, respectively, across 100 experiments.

4.4 Discussion

In this retrospective study conducted using a national perioperative research network, we found that participating in a research network produced a better model (substantially higher AUC) in aggregate for both the temporal and external validation sets as compared to single-center models. When examining the model performance at each individual center, a multicenter model had

higher performance for all individual centers and all AKI outcomes versus a single-center approach. This finding is particularly salient given the alternative conventional assumption that a single center's predictors might be expected to perform better temporally on its own population without potential contamination by center-specific effects from other centers. We found that this conventional wisdom does not hold.

Even if data sharing were not an option, a federated model without hospital metadata outperformed a single-center modeling approach at the vast majority of hospitals for all outcomes, and the benefits of federating were the largest for hospitals with fewer cases. The benefits of building multicenter AI models were greatest for smaller networks, with diminishing returns observed with each additional hospital being added to the network.

Our study suggests that contrary to a recent decision to build separate sepsis models for each hospital by an EHR vendor[107], sharing data or model parameters across centers generally produces superior models than an approach relying on a bespoke model for each center. Even hospitals with high case volumes in our study obtained a higher performing model from either pooling or federating, although this difference was marginal for the largest hospitals. On the other hand, the hospitals with the fewest training set cases either were unable to produce models or produced poorly performing models. Despite not contributing many cases to training, these hospitals generally saw high performance from pooled or federated models. Viewed through an equity lens, the centers with the largest case volume reap the smallest benefit from participating in an AI modeling research network, but their contribution of either data or base models to the research network greatly benefits the hospitals with smaller case volumes.

We also introduce an efficient two-stage federated learning approach inspired by model stacking. Our federated stacked learning approach achieved similar performance to the pooled models,

with an AUC lower than a pooled approach by approximately 0.03 in temporal validation and 0.02 in external validation across all outcomes. One of the challenges with implementing federated learning models in clinical practice is that the federation procedure generally becomes more complex as the network size grows. Along with a lack of federation infrastructure in the EHR, this may explain why federated learning models are rarely deployed in the EHR despite robust evaluations in the literature. It is thus important to emphasize that the complexity of the FSL approach used in this study does not change as additional centers are added. For example, adding a center requires the new center to train a base model, for the base model's predictions to be disseminated to a central server, after which the meta-model is retrained with the addition of the new predictions. Centers already in the federated network do not have to do any additional work or retraining. We hope that simplifying the federated learning approach will make it more tractable to implement within research networks.

Our study has limitations. We only examined one clinical scenario (AKI), although we examined different severities of AKI with varying incidence and found largely concordant results regardless of the chosen severity of AKI outcome. While we did have a relatively large sample size, cardiac surgery occurs less commonly than conditions like sepsis or acute kidney injury among general inpatients. Thus, while our findings are robust to a relatively large sample size, our findings may not generalize to situations where individual centers can accumulate a much larger sample size than what we studied. Lastly, while we looked at calibration in the aggregate validation cohorts, we did not evaluate hospital-specific calibration. This will be important to consider in future work because multicenter AI models may need to be recalibrated for specific hospitals to ensure generalizability[74,112].

Despite these limitations, our study has important implications for health systems considering whether the investment in research or quality networks focused on building multicenter AI models is sufficient to justify their participation. While prior evaluations of federated learning or pooling of data or coefficients have largely focused on aggregate performance from these approaches, we additionally show that the benefit to individual centers varies based on their sample size. The benefits are largest for centers with the smallest sample size and made possible by the participation of the larger centers. Thus, while the largest health systems may not see substantial benefits from pooling or federating with other systems, their participation in such efforts should be encouraged and incentivized as it leads to direct benefits to smaller health systems.

4.5 Data Sharing

The datasets involved in this study are defined as limited datasets per United States Federal Regulations and require execution of a data use agreement for transfer or use of the data. They are derived from data shared within the Multicenter Perioperative Outcomes Group (MPOG). The investigative team is able to share data securely and transparently conditional on: (i) receipt of a detailed written request identifying the requestor, purpose and proposed use of the shared data, (ii) use of a secure enclave for the sharing of personally identifiable information and (iii) the request is permissible within the confines of existing data use agreements executed between MPOG members.

4.6 Supplemental Materials

4.6.1 Supplemental Methods

Minimum Data Quality Standards

To ensure high data quality, each included case required a minimum data quality standard for inclusion in the dataset, defined as the presence of date of surgery, comorbidities, preoperative laboratory values, intraoperative arterial blood gas values and physiologic monitoring data, baseline kidney function, and postoperative creatinine values (to derive CSA-AKI outcomes). Values of intraoperative time-varying variables were restricted to physiologically plausible valid ranges.

Preoperative AKI

Preoperative AKI was calculated by comparing the lowest baseline sCr value in the 60 days prior to surgery to the most recent serum creatinine value closest to surgery, staged for severity according to the Kidney Disease: Improving Global Outcomes (KDIGO) international guidelines: Stage 1 AKI was defined as a sCr level increase $\geqslant$ 0.3 mg/dL or $\geqslant$1.5 times baseline. Stage 2 AKI was defined as an increase of $\geqslant$2 times the baseline, and Stage 3 AKI $\geqslant$3 times baseline or an increase to $\geqslant$4.0 mg/dL[16].

Gradient-Boosted Decision Tree Training and Early Stopping

Gradient-boosted decision tree models (GBDT) were trained on a random 80% split of the training set to predict the outcome AKI stage as a multinomial outcome (i.e. “No AKI”, “AKI stage 1”, “AKI stage 2”, “AKI stage 3”, “AKI stage 3D”) for each case using 426 predictors with a maximum of 1000 trees and a maximum depth of 5. The remaining 20% of the training set was used to determine the need for early stopping based on an improvement in log loss lower than 0.0005 on 5 consecutive rounds based on a moving average calculated after every 10 trees. Categorical predictors were reordered by the mean response of each level for more efficient training. Internally, a separate one-versus-all tree was trained for each outcome class and averaged to produce probabilities for achieving each AKI stage postoperatively.

4.6.2 Supplemental Tables

Supplemental Table 4.1 Description of variables and predictors in the study models

Variable	Description	Variable Type	Category	Valid Range	Phase	Summary statistics	Number of predictors
age	Age (years)	Fixed	Patient demographics		-	-	1
gender	Sex	Fixed	Patient demographics		-	-	1
race	Race	Fixed	Patient demographics		-	-	1
height	Height (cm)	Fixed	Patient demographics		-	-	1
weight	Weight (kg)	Fixed	Patient demographics		-	-	1
bmi	BMI (kg/m2)	Fixed	Patient demographics		-	-	1
smoking_classification	Smoking Classification	Fixed	Patient demographics		-	-	1
asa_class	ASA class	Fixed	Case characteristics		-	-	1
university_affiliated	University affiliation	Fixed	Case characteristics		-	-	1
weekend	Case performed on a weekend	Fixed	Case characteristics		-	-	1
holiday	Case performend on a holiday	Fixed	Case characteristics		-	-	1
emergent	Emergent case ((ASA “E” status)	Fixed	Case characteristics		-	-	1
resident_present	Resident present	Fixed	Case characteristics		-	-	1
crna_present	CRNA present	Fixed	Case characteristics		-	-	1
procedure_type_aortic	Non-hypothermia circulatory arrest aortic	Fixed	Case characteristics		-	-	1
procedure_type_circ_arrest	Circulatory arrest	Fixed	Case characteristics		-	-	1
procedure_type_heart_transplant	Heart transplant	Fixed	Case characteristics		-	-	1
procedure_type_pte	Pulmonary thromboendarterectomy	Fixed	Case characteristics		-	-	1
procedure_type_myectomy	Myectomy	Fixed	Case characteristics		-	-	1
procedure_type_vad	Ventricular assist device (VAD) pre-existing or implanted	Fixed	Case characteristics		-	-	1
procedure_type_iabp	Intra-aortic balloon pumps (IABP) pre-existing or placed	Fixed	Case characteristics		-	-	1

procedure_type_other_mech_support	Other mechanical support device pre-existing or placed	Fixed	Case characteristics		-	-	1
procedure_type_cab_only	Coronary artery bypass (CAB) only	Fixed	Case characteristics		-	-	1
procedure_type_valve_only	Valve replacement only	Fixed	Case characteristics		-	-	1
procedure_type_valve_cab_only	CAB and valve replacement	Fixed	Case characteristics		-	-	1
baseline_bp_map	Baseline BP MAP	Fixed	Pre-operative lab/phys results		-	-	1
preop_platelets	Pre-operative platelets	Fixed	Pre-operative lab/phys results		-	-	1
preop_wbc	Pre-operative WBC	Fixed	Pre-operative lab/phys results		-	-	1
preop_sodium	Pre-operative sodium	Fixed	Pre-operative lab/phys results	[90, 190]	-	-	1
preop_potassium	Pre-operative potassium	Fixed	Pre-operative lab/phys results	[0, 50]	-	-	1
preop_glucose	Pre-operative glucose	Fixed	Pre-operative lab/phys results	[0, 600]	-	-	1
preop_hemoglobin_combined	Pre-operative hemoglobin or hematocrit/3	Fixed	Pre-operative lab/phys results	[0, 30]	-	-	1
preop_hco3_or_co2_serum	Pre-operative bicarbonate or serum CO2	Fixed	Pre-operative lab/phys results	[0, 55]	-	-	1
bl110_count	Anticoagulants	Fixed	Pre-operative home medications		-	-	1
bl117_count	Platelet aggregation inhibitors	Fixed	Pre-operative home medications		-	-	1
cv050_count	Digitalis glycosides	Fixed	Pre-operative home medications		-	-	1
cv100_count	Beta blockers	Fixed	Pre-operative home medications		-	-	1
cv150_count	Alpha blockers	Fixed	Pre-operative home medications		-	-	1
cv200_count	Calcium channel blockers	Fixed	Pre-operative home medications		-	-	1
cv250_count	Anti-anginals	Fixed	Pre-operative home medications		-	-	1
cv300_count	Anti-arrhythmics	Fixed	Pre-operative home medications		-	-	1
cv350_count	Anti-lipemics	Fixed	Pre-operative home medications		-	-	1
cv701_count	Thiazide diuretics	Fixed	Pre-operative home medications		-	-	1
cv702_count	Loop diuretics	Fixed	Pre-operative home medications		-	-	1
cv704_count	Potassium sparing diuretics	Fixed	Pre-operative home medications		-	-	1

cv800_count	Angiotensin-converting enzyme (ACE) inhibitors	Fixed	Pre-operative home medications	-	-	1
cv805_count	Angiotensin II (ATII) inhibitors	Fixed	Pre-operative home medications	-	-	1
hs502_count	Oral hypoglycemic agents	Fixed	Pre-operative home medications	-	-	1
elixhauser_aids_hiv	Elixhauser Comorbidity - AIDS/HIV	Fixed	Comorbidities	-	-	1
elixhauser_alcohol_abuse	Elixhauser Comorbidity - Alcohol abuse	Fixed	Comorbidities	-	-	1
elixhauser_blood_loss_anemia	Elixhauser Comorbidity - Blood loss anemia	Fixed	Comorbidities	-	-	1
elixhauser_cardiac_arrhythmia	Elixhauser Comorbidity - Cardiac arrhythmia	Fixed	Comorbidities	-	-	1
elixhauser_chronic_pulmonary_disease	Elixhauser Comorbidity - Chronic pulmonary disease	Fixed	Comorbidities	-	-	1
elixhauser_coagulopathy	Elixhauser Comorbidity - Coagulopathy	Fixed	Comorbidities	-	-	1
elixhauser_congestive_heart_failure	Elixhauser Comorbidity - Congestive heart failure	Fixed	Comorbidities	-	-	1
elixhauser_deficiency_anemia	Elixhauser Comorbidity - Deficiency anemia	Fixed	Comorbidities	-	-	1
elixhauser_depression	Elixhauser Comorbidity - Depression	Fixed	Comorbidities	-	-	1
elixhauser_diabetes_with_complications	Elixhauser Comorbidity - Diabetes with complications	Fixed	Comorbidities	-	-	1
elixhauser_diabetes_without_complications	Elixhauser Comorbidity - Diabetes without complications	Fixed	Comorbidities	-	-	1
elixhauser_drug_abuse	Elixhauser Comorbidity - Drug abuse	Fixed	Comorbidities	-	-	1
elixhauser_fluid_and_electrolyte_disorders	Elixhauser Comorbidity - Fluid and electrolyte disorders	Fixed	Comorbidities	-	-	1
elixhauser_hypertension_with_complications	Elixhauser Comorbidity - Hypertension with complications	Fixed	Comorbidities	-	-	1
elixhauser_hypertension_without_complications	Elixhauser Comorbidity - Hypertension without complications	Fixed	Comorbidities	-	-	1
elixhauser_hypothyroidism	Elixhauser Comorbidity - Hypothyroidism	Fixed	Comorbidities	-	-	1
elixhauser_liver_disease	Elixhauser Comorbidity - Liver disease	Fixed	Comorbidities	-	-	1
elixhauser_lymphoma	Elixhuaser Comorbidity - Lymphoma	Fixed	Comorbidities	-	-	1
elixhauser_metastatic_cancer	Elixhauser Comorbidity - Metastatic cancer	Fixed	Comorbidities	-	-	1
elixhauser_obesity	Elixhauser Comorbidity - Obesity	Fixed	Comorbidities	-	-	1
elixhauser_other_neurological_disorders	Elixhauser Comorbidity - Other neurological disorders	Fixed	Comorbidities	-	-	1

elixhauser_paralysis	Elixhauser Comorbidity - Paralysis	Fixed	Comorbidities	-	-		1
elixhauser_peptic_ulcer_disease_excluding_bleeding	Elixhauser Comorbidity - Peptic ulcer disease excluding bleeding	Fixed	Comorbidities	-	-		1
elixhauser_peripheral_vascular_disorders	Elixhauser Comorbidity - Peripheral vascular disorders	Fixed	Comorbidities	-	-		1
elixhauser_psychoses	Elixhauser Comorbidity - Psychoses	Fixed	Comorbidities	-	-		1
elixhauser_pulmonary_circulation_disorders	Elixhauser Comorbidity - Pulmonary circulation disorders	Fixed	Comorbidities	-	-		1
elixhauser_rheumatoid_arthritis_collagen	Elixhauser Comorbidity - Rheumatoid arthritis collagen	Fixed	Comorbidities	-	-		1
elixhauser_solid_tumor_without_metastasis	Elixhauser Comorbidity - Solid tumor without metastasis	Fixed	Comorbidities	-	-		1
elixhauser_valvular_disease	Elixhauser Comorbidity - Valvular disease	Fixed	Comorbidities	-	-		1
elixhauser_weight_loss	Elixhauser Comorbidity - Weight loss	Fixed	Comorbidities	-	-		1
preop_creatinine_baseline	Pre-operative baseline sCr (lowest within 60 days prior to surgery)	Fixed	Baseline kidney function	-	-		1
preop_creatinine_most_recent	Pre-operative most recent sCr (closest to surgery, within 60 days prior to surgery)	Fixed	Baseline kidney function	-	-		1
ratio_creatinine_most_recent_to_baseline	Ratio of most recent sCr to baseline sCr	Fixed	Baseline kidney function	-	-		1
diff_creatinine_baseline_to_most_recent	Increase from baseline sCr to most recent sCr	Fixed	Baseline kidney function	-	-		1
baseline_aki_stage	Baseline AKI stage (no AKI, AKI-1, AKI-2, AKI-3)	Fixed	Baseline kidney function	-	-		1
first_postop_creatinine_within_24h	First post-operative sCr within 24h	Fixed	First post-operative sCr	-	-		1
bp_sys	Intra-operative systolic BP	Time-varying	Intra-operative BP	[0, 400]	pre-CPB	first, last, length, min, mean, median, max, slope	8
bp_sys	Intra-operative systolic BP	Time-varying	Intra-operative BP	[0, 400]	intra-CPB	first, last, length, min, mean, median, max, slope	8
bp_sys	Intra-operative systolic BP	Time-varying	Intra-operative BP	[0, 400]	post-CPB	first, last, length, min, mean, median, max, slope	8
bp_dias	Intra-operative diastolic BP	Time-varying	Intra-operative BP	[0, 300]	pre-CPB	first, last, length, min, mean, median, max, slope	8
bp_dias	Intra-operative diastolic BP	Time-varying	Intra-operative BP	[0, 300]	intra-CPB	first, last, length, min, mean, median, max, slope	8
bp_dias	Intra-operative diastolic BP	Time-varying	Intra-operative BP	[0, 300]	post-CPB	first, last, length, min, mean, median, max, slope	8

bp_map	Intra-operative MAP	Time-varying	Intra-operative BP	[0, 200]	pre-CPB	first, last, length, min, mean, median, max, slope	8
bp_map	Intra-operative MAP	Time-varying	Intra-operative BP	[0, 200]	intra-CPB	first, last, length, min, mean, median, max, slope	8
bp_map	Intra-operative MAP	Time-varying	Intra-operative BP	[0, 200]	post-CPB	first, last, length, min, mean, median, max, slope	8
cvp	Intra-operative CVP	Time-varying	Intra-operative CVP	[-10, 40]	pre-CPB	first, last, length, min, mean, median, max, slope	8
cvp	Intra-operative CVP	Time-varying	Intra-operative CVP	[-10, 40]	intra-CPB	first, last, length, min, mean, median, max, slope	8
cvp	Intra-operative CVP	Time-varying	Intra-operative CVP	[-10, 40]	post-CPB	first, last, length, min, mean, median, max, slope	8
Bicarbonate	Intra-operative bicarbonate	Time-varying	Intra-operative lab	[0, 55]	pre-CPB	first, last, length, min, mean, median, max, slope	8
Bicarbonate	Intra-operative bicarbonate	Time-varying	Intra-operative lab	[0, 55]	intra-CPB	first, last, length, min, mean, median, max, slope	8
Bicarbonate	Intra-operative bicarbonate	Time-varying	Intra-operative lab	[0, 55]	post-CPB	first, last, length, min, mean, median, max, slope	8
Glucose	Intra-operative glucose	Time-varying	Intra-operative lab	[0, 600]	pre-CPB	first, last, length, min, mean, median, max, slope	8
Glucose	Intra-operative glucose	Time-varying	Intra-operative lab	[0, 600]	intra-CPB	first, last, length, min, mean, median, max, slope	8
Glucose	Intra-operative glucose	Time-varying	Intra-operative lab	[0, 600]	post-CPB	first, last, length, min, mean, median, max, slope	8
Hemoglobin	Intra-operative hemoglobin	Time-varying	Intra-operative lab	[0, 30]	pre-CPB	first, last, length, min, mean, median, max, slope	8
Hemoglobin	Intra-operative hemoglobin	Time-varying	Intra-operative lab	[0, 30]	intra-CPB	first, last, length, min, mean, median, max, slope	8
Hemoglobin	Intra-operative hemoglobin	Time-varying	Intra-operative lab	[0, 30]	post-CPB	first, last, length, min, mean, median, max, slope	8
pCO2	Intra-operative pCO2	Time-varying	Intra-operative lab	[0, 200]	pre-CPB	first, last, length, min, mean, median, max, slope	8
pCO2	Intra-operative pCO2	Time-varying	Intra-operative lab	[0, 200]	intra-CPB	first, last, length, min, mean,	8

						median, max, slope	
pCO2	Intra-operative pCO2	Time-varying	Intra-operative lab	[0, 200]	post-CPB	first, last, length, min, mean, median, max, slope	8
pH	Intra-operative pH	Time-varying	Intra-operative lab	[6.7, 8]	pre-CPB	first, last, length, min, mean, median, max, slope	8
pH	Intra-operative pH	Time-varying	Intra-operative lab	[6.7, 8]	intra-CPB	first, last, length, min, mean, median, max, slope	8
pH	Intra-operative pH	Time-varying	Intra-operative lab	[6.7, 8]	post-CPB	first, last, length, min, mean, median, max, slope	8
Potassium	Intra-operative potassium	Time-varying	Intra-operative lab	[0, 50]	pre-CPB	first, last, length, min, mean, median, max, slope	8
Potassium	Intra-operative potassium	Time-varying	Intra-operative lab	[0, 50]	intra-CPB	first, last, length, min, mean, median, max, slope	8
Potassium	Intra-operative potassium	Time-varying	Intra-operative lab	[0, 50]	post-CPB	first, last, length, min, mean, median, max, slope	8
Sodium	Intra-operative sodium	Time-varying	Intra-operative lab	[90, 190]	pre-CPB	first, last, length, min, mean, median, max, slope	8
Sodium	Intra-operative sodium	Time-varying	Intra-operative lab	[90, 190]	intra-CPB	first, last, length, min, mean, median, max, slope	8
Sodium	Intra-operative sodium	Time-varying	Intra-operative lab	[90, 190]	post-CPB	first, last, length, min, mean, median, max, slope	8
Albuterol	Intra-operative administration of albuterol	Time-varying	Intra-operative medication		pre-CPB	length	1
Albuterol	Intra-operative administration of albuterol	Time-varying	Intra-operative medication		intra-CPB	length	1
Albuterol	Intra-operative administration of albuterol	Time-varying	Intra-operative medication		post-CPB	length	1
Angiotension II	Intra-operative administration of angiotension II	Time-varying	Intra-operative medication		pre-CPB	length	1
Angiotension II	Intra-operative administration of angiotension II	Time-varying	Intra-operative medication		intra-CPB	length	1
Angiotension II	Intra-operative administration of angiotension II	Time-varying	Intra-operative medication		post-CPB	length	1
Dobutamine	Intra-operative administration of dobutamine	Time-varying	Intra-operative medication		pre-CPB	length	1
Dobutamine	Intra-operative administration of dobutamine	Time-varying	Intra-operative medication		intra-CPB	length	1
Dobutamine	Intra-operative administration of dobutamine	Time-varying	Intra-operative medication		post-CPB	length	1
Dopamine	Intra-operative administration of dopamine	Time-varying	Intra-operative medication		pre-CPB	length	1

Dopamine	Intra-operative administration of dopamine	Time-varying	Intra-operative medication		intra-CPB	length	1
Dopamine	Intra-operative administration of dopamine	Time-varying	Intra-operative medication		post-CPB	length	1
Ephedrine	Intra-operative administration of ephedrine	Time-varying	Intra-operative medication		pre-CPB	length	1
Ephedrine	Intra-operative administration of ephedrine	Time-varying	Intra-operative medication		intra-CPB	length	1
Ephedrine	Intra-operative administration of ephedrine	Time-varying	Intra-operative medication		post-CPB	length	1
Epinephrine	Intra-operative administration of epinephrine	Time-varying	Intra-operative medication		pre-CPB	length	1
Epinephrine	Intra-operative administration of epinephrine	Time-varying	Intra-operative medication		intra-CPB	length	1
Epinephrine	Intra-operative administration of epinephrine	Time-varying	Intra-operative medication		post-CPB	length	1
Milrinone	Intra-operative administration of milrinone	Time-varying	Intra-operative medication		pre-CPB	length	1
Milrinone	Intra-operative administration of milrinone	Time-varying	Intra-operative medication		intra-CPB	length	1
Milrinone	Intra-operative administration of milrinone	Time-varying	Intra-operative medication		post-CPB	length	1
Norepinephrine	Intra-operative administration of norepinephrine	Time-varying	Intra-operative medication		pre-CPB	length	1
Norepinephrine	Intra-operative administration of norepinephrine	Time-varying	Intra-operative medication		intra-CPB	length	1
Norepinephrine	Intra-operative administration of norepinephrine	Time-varying	Intra-operative medication		post-CPB	length	1
Phenylephrine	Intra-operative administration of phenylephrine	Time-varying	Intra-operative medication		pre-CPB	length	1
Phenylephrine	Intra-operative administration of phenylephrine	Time-varying	Intra-operative medication		intra-CPB	length	1
Phenylephrine	Intra-operative administration of phenylephrine	Time-varying	Intra-operative medication		post-CPB	length	1
Vasopressin	Intra-operative administration of vasopressin	Time-varying	Intra-operative medication		pre-CPB	length	1
Vasopressin	Intra-operative administration of vasopressin	Time-varying	Intra-operative medication		intra-CPB	length	1
Vasopressin	Intra-operative administration of vasopressin	Time-varying	Intra-operative medication		post-CPB	length	1
HR	Intra-operative heart rate	Time-varying	Intra-operative physiology	[30, 180]	pre-CPB	first, last, length, min, mean, median, max, slope	8
HR	Intra-operative heart rate	Time-varying	Intra-operative physiology	[30, 180]	intra-CPB	first, last, length, min, mean, median, max, slope	8
HR	Intra-operative heart rate	Time-varying	Intra-operative physiology	[30, 180]	post-CPB	first, last, length, min, mean, median, max, slope	8
SpO2	Intra-operative SpO2	Time-varying	Intra-operative physiology	[60, 100]	pre-CPB	first, last, length, min, mean, median, max, slope	8

SpO2	Intra-operative SpO2	Time-varying	Intra-operative physiology	[60, 100]	intra-CPB	first, last, length, min, mean, median, max, slope	8
SpO2	Intra-operative SpO2	Time-varying	Intra-operative physiology	[60, 100]	post-CPB	first, last, length, min, mean, median, max, slope	8
Total							426

Supplemental Table 4.2 Number of cases at each center by data partition

Center	Training	Temporal Validation	External Validation	University-affiliated/Academic
1	2,473	1,956	-	Yes
4	2,139	972	-	Yes
5	1,811	0	-	Yes
7	1,105	266	-	Yes
10	3,345	1,403	-	Yes
14	0	744	-	Yes
16	16,251	3,469	-	Yes
19	2,140	825	-	Yes
23	700	326	-	Yes
32	-	-	238	No
35	31	521	-	Yes
37	-	-	989	Yes
38	829	1	-	Yes
40	728	1	-	No
46	-	-	259	No
47	168	58	-	Yes
58	1,356	918	-	Yes
65	261	0	-	Yes
66	1,769	560	-	Yes
68	1,721	1,374	-	Yes
70	303	205	-	Yes
76	57	94	-	Yes
78	0	370	-	Yes
83	280	88	-	Yes
84	6,318	1,941	-	Yes
86	-	-	2,622	Yes
89	90	724	-	Yes
91	36	275	-	Yes
92	0	20	-	No
101	0	548	-	Yes
102	15	473	-	Yes
Total	43,926	18,132	4,108	

Supplemental Table 4.3 Extended patient and surgical characteristics.

Characteristic		Overall (N = 66,166)	Training (N = 43,926)	Temporal Validation (N = 18,132)	External Validation (N = 4,108)
Preoperative Patient Characteristics					
Age (years)		62.0 (13.5)	61.9 (13.6)	61.7 (13.3)	64.4 (12.3)
Sex					
	Female	20,921 (31.6%)	14,062 (32.0%)	5,726 (31.6%)	1,133 (27.6%)
	Male	45,245 (68.4%)	29,864 (68.0%)	12,406 (68.4%)	2,975 (72.4%)
Race / Ethnicity					
	White not of hispanic origin	52,643 (79.6%)	34,822 (79.3%)	14,335 (79.1%)	3,486 (84.9%)
	Black not of hispanic origin	4,304 (6.5%)	2,720 (6.2%)	1,494 (8.2%)	90 (2.2%)
	Asian or Pacific Islander	2,068 (3.1%)	1,215 (2.8%)	624 (3.4%)	229 (5.6%)
	Bi or Multi Racial	569 (0.9%)	413 (0.9%)	156 (0.9%)	0 (0.0%)
	American Indian or Alaska Native	180 (0.3%)	102 (0.2%)	66 (0.4%)	12 (0.3%)
	Hispanic white	514 (0.8%)	267 (0.6%)	206 (1.1%)	41 (1.0%)
	Hispanic black	38 (0.1%)	18 (0.0%)	19 (0.1%)	1 (0.0%)
	Middle Eastern	38 (0.1%)	38 (0.1%)	0 (0.0%)	0 (0.0%)
	Missing	5,812 (8.8%)	4,331 (9.9%)	1,232 (6.8%)	249 (6.1%)
Height (cm)		171.9 (14.1)	172.5 (10.9)	170.2 (19.9)	172.4 (10.4)
	Missing	3,424 (5.2%)	2,877 (6.5%)	524 (2.9%)	23 (0.6%)
Weight (kg)		87.0 (20.8)	86.7 (20.7)	87.7 (21.2)	87.2 (21.0)
	Missing	1,795 (2.7%)	1,055 (2.4%)	734 (4.0%)	6 (0.1%)
Body Mass Index (kg/m^2)		29.1 (6.2)	29.0 (6.2)	29.2 (6.3)	29.3 (6.3)

	Missing	4,672 (7.1%)	3,071 (7.0%)	1,572 (8.7%)	29 (0.7%)
Smoking Classification					
	Non-Smoker	3,450 (21.3%)	2,074 (18.6%)	1,106 (29.1%)	270 (21.6%)
	Smoker	5,232 (32.3%)	3,705 (33.3%)	1,198 (31.5%)	329 (26.3%)
	Former Smoker	7,017 (43.3%)	5,074 (45.6%)	1,301 (34.2%)	642 (51.3%)
	Conflicting Documentation	489 (3.0%)	283 (2.5%)	196 (5.2%)	10 (0.8%)
ASA Physical Status Classification					
	ASA Class 1	59 (0.1%)	54 (0.1%)	3 (0.0%)	2 (0.0%)
	ASA Class 2	397 (0.6%)	267 (0.6%)	72 (0.4%)	58 (1.4%)
	ASA Class 3	13,838 (20.9%)	8,806 (20.0%)	3,130 (17.3%)	1,902 (46.3%)
	ASA Class 4	50,982 (77.1%)	34,286 (78.1%)	14,614 (80.6%)	2,082 (50.7%)
	ASA Class 5	890 (1.3%)	513 (1.2%)	313 (1.7%)	64 (1.6%)
Preoperative Laboratory Values					
Platelet Count, (K/mL)		218.6 (72.2)	216.9 (70.2)	221.9 (77.2)	223.0 (70.4)
	Missing	1,122 (1.7%)	134 (0.3%)	984 (5.4%)	4 (0.1%)
White Blood Cell Count (per mL)		7.6 (3.1)	7.5 (3.1)	7.7 (3.1)	7.9 (3.4)
	Missing	1,077 (1.6%)	1,058 (2.4%)	17 (0.1%)	2 (0.0%)
Sodium (mEq/L)		138.8 (3.2)	139.1 (3.2)	138.3 (3.1)	138.5 (3.1)
Potassium (mEq/L)		4.2 (0.4)	4.2 (0.4)	4.2 (0.4)	4.1 (0.4)
Glucose (g/dL)		115.8 (39.3)	115.1 (39.6)	116.5 (38.7)	119.0 (38.0)
Hemoglobin (g/dL)		13.3 (2.0)	13.3 (2.0)	13.2 (2.1)	13.3 (1.9)
Bicarbonate (mmol/L)		25.6 (3.1)	25.8 (3.1)	25.3 (3.3)	24.8 (2.8)
Creatinine-Related Variables					
Preoperative Baseline Serum Creatinine, g/dL		1.0 (0.3)	1.0 (0.3)	0.9 (0.3)	0.9 (0.3)

Preoperative Most Recent Serum Creatinine, g/dL		1.0 (0.5)	1.0 (0.6)	1.0 (0.3)	1.0 (0.3)
Preoperative Serum Creatinine Ratio (Most Recent/Baseline)		1.1 (0.5)	1.1 (0.6)	1.1 (0.2)	1.1 (0.2)
Preoperative Serum Creatinine Difference (Most Recent - Baseline)		0.1 (0.5)	0.1 (0.5)	0.1 (0.1)	0.1 (0.1)
First Post-operative Serum Creatinine Within 24h		1.0 (0.3)	1.0 (0.3)	1.0 (0.3)	0.9 (0.3)
	Missing	286 (0.4%)	222 (0.5%)	60 (0.3%)	4 (0.1%)
Preoperative AKI					
	No Preoperative AKI	62,831 (95.0%)	41,831 (95.2%)	17,081 (94.2%)	3,919 (95.4%)
	Preoperative AKI-1	3,066 (4.6%)	1,934 (4.4%)	967 (5.3%)	165 (4.0%)
	Preoperative AKI-2	218 (0.3%)	136 (0.3%)	66 (0.4%)	16 (0.4%)
	Preoperative AKI-3	51 (0.1%)	25 (0.1%)	18 (0.1%)	8 (0.2%)
Preoperative Patient Comorbidities (Elixhauser)					
AIDS/HIV		195 (0.3%)	110 (0.3%)	57 (0.3%)	28 (0.7%)
Alcohol Abuse		619 (0.9%)	497 (1.1%)	94 (0.5%)	28 (0.7%)
Blood Loss Anemia		1,532 (2.3%)	1,053 (2.4%)	398 (2.2%)	81 (2.0%)
Cardiac Arrhythmia		42,934 (64.9%)	27,245 (62.0%)	12,912 (71.2%)	2,777 (67.6%)
Chronic Pulmonary Disease		15,116 (22.8%)	10,143 (23.1%)	4,008 (22.1%)	965 (23.5%)
Coagulopathy		25,473 (38.5%)	15,989 (36.4%)	8,898 (49.1%)	586 (14.3%)
Congestive Heart Failure		31,495 (47.6%)	19,952 (45.4%)	9,910 (54.7%)	1,633 (39.8%)
Deficiency Anemia		3,142 (4.7%)	1,925 (4.4%)	1,040 (5.7%)	177 (4.3%)
Depression		9,898 (15.0%)	6,071 (13.8%)	3,160 (17.4%)	667 (16.2%)
Diabetes with Complications		6,719 (10.2%)	4,158 (9.5%)	2,159 (11.9%)	402 (9.8%)
Diabetes without Complications					
	No	54,197 (81.9%)	36,329 (82.7%)	14,946 (82.4%)	2,922 (71.1%)
	Yes	11,851 (17.9%)	7,503 (17.1%)	3,162 (17.4%)	1,186 (28.9%)
	Missing	118 (0.2%)	94 (0.2%)	24 (0.1%)	0 (0.0%)
Drug Abuse		2,672 (4.0%)	1,559 (3.5%)	888 (4.9%)	225 (5.5%)

Fluid and Electrolyte Disorders		39,655 (59.9%)	25,740 (58.6%)	12,602 (69.5%)	1,313 (32.0%)
Hypertension		24,791 (37.5%)	14,337 (32.6%)	8,981 (49.5%)	1,473 (35.9%)
Hypothyroidism		9,223 (13.9%)	6,235 (14.2%)	2,507 (13.8%)	481 (11.7%)
Liver Disease		4,758 (7.2%)	2,850 (6.5%)	1,609 (8.9%)	299 (7.3%)
Lymphoma		496 (0.7%)	335 (0.8%)	124 (0.7%)	37 (0.9%)
Metastatic Cancer		346 (0.5%)	218 (0.5%)	105 (0.6%)	23 (0.6%)
Obesity		16,562 (25.0%)	10,161 (23.1%)	5,387 (29.7%)	1,014 (24.7%)
Other Neurological Disorders		5,337 (8.1%)	3,181 (7.2%)	1,790 (9.9%)	366 (8.9%)
Paralysis		1,267 (1.9%)	788 (1.8%)	394 (2.2%)	85 (2.1%)
Peptic Ulcer Disease excluding Bleeding		628 (0.9%)	413 (0.9%)	183 (1.0%)	32 (0.8%)
Peripheral Vascular Disorders		24,588 (37.2%)	15,781 (35.9%)	7,557 (41.7%)	1,250 (30.4%)
Psychoses		470 (0.7%)	303 (0.7%)	137 (0.8%)	30 (0.7%)
Pulmonary Circulation Disorders		12,034 (18.2%)	7,682 (17.5%)	3,632 (20.0%)	720 (17.5%)
Rheumatoid Arthritis / Collagen Vascular Disease					
	No	63,888 (96.6%)	42,446 (96.6%)	17,463 (96.3%)	3,979 (96.9%)
	Yes	2,160 (3.3%)	1,386 (3.2%)	645 (3.6%)	129 (3.1%)
	Missing	118 (0.2%)	94 (0.2%)	24 (0.1%)	0 (0.0%)
Solid Tumor without Metastatis					
	No	64,437 (97.4%)	42,830 (97.5%)	17,616 (97.2%)	3,991 (97.2%)
	Yes	1,611 (2.4%)	1,002 (2.3%)	492 (2.7%)	117 (2.8%)
	Missing	118 (0.2%)	94 (0.2%)	24 (0.1%)	0 (0.0%)
Valvular Disease					
	Yes	46,092 (69.7%)	31,145 (70.9%)	12,513 (69.0%)	2,434 (59.3%)
	No	19,956 (30.2%)	12,687 (28.9%)	5,595 (30.9%)	1,674 (40.7%)
	Missing	118 (0.2%)	94 (0.2%)	24 (0.1%)	0 (0.0%)
Weight Loss					

	No	60,259 (91.1%)	40,069 (91.2%)	16,286 (89.8%)	3,904 (95.0%)
	Yes	5,789 (8.7%)	3,763 (8.6%)	1,822 (10.0%)	204 (5.0%)
	Missing	118 (0.2%)	94 (0.2%)	24 (0.1%)	0 (0.0%)
Surgical Characteristics - Procedure Type					
Valve Only		21,670 (32.8%)	15,371 (35.0%)	5,232 (28.9%)	1,067 (26.0%)
Coronary Artery Bypass Only		18,573 (28.1%)	11,350 (25.8%)	5,356 (29.5%)	1,867 (45.4%)
Aortic		9,316 (14.1%)	6,114 (13.9%)	2,852 (15.7%)	350 (8.5%)
Valve + Coronary Artery Bypass Only		6,455 (9.8%)	4,433 (10.1%)	1,492 (8.2%)	530 (12.9%)
Myectomy		2,156 (3.3%)	1,588 (3.6%)	541 (3.0%)	27 (0.7%)
Ventricular Assist Device		1,680 (2.5%)	1,161 (2.6%)	463 (2.6%)	56 (1.4%)
Heart Transplant		1,669 (2.5%)	953 (2.2%)	664 (3.7%)	52 (1.3%)
Pulmonary Thromboendarterectomy		385 (0.6%)	242 (0.6%)	143 (0.8%)	0 (0.0%)
Other		4,264 (6.4%)	2,716 (6.2%)	1,389 (7.7%)	159 (3.9%)
Additional Surgical Characteristics					
Anesthesia Duration (min)		419 (133)	417 (133)	434 (134)	371 (121)
Cardiopulmonary Bypass Duration (min)		140 (88)	133 (85)	157 (99)	137 (65)
Circulatory Arrest Used		1,267 (1.9%)	790 (1.8%)	464 (2.6%)	13 (0.3%)
Intra-Aortic Balloon Pump Used		843 (1.3%)	452 (1.0%)	367 (2.0%)	24 (0.6%)
Other Mechanical Support Used following CPB		226 (0.3%)	106 (0.2%)	118 (0.7%)	2 (0.0%)
Emergent					
	No	61,605 (93.1%)	41,313 (94.1%)	16,465 (90.8%)	3,827 (93.2%)
	Yes	4,059 (6.1%)	2,613 (5.9%)	1,165 (6.4%)	281 (6.8%)
	Missing	502 (0.8%)	0 (0.0%)	502 (2.8%)	0 (0.0%)
Anesthesia Technique					
	General - ETT	65,097 (98.4%)	43,225 (98.4%)	17,771 (98.0%)	4,101 (99.8%)
	General - LMA followed by ETT	1,069 (1.6%)	701 (1.6%)	361 (2.0%)	7 (0.2%)

Case Scheduling / Staffing Characteristics					
Weekend					
	Weekday	64,317 (97.2%)	42,829 (97.5%)	17,498 (96.5%)	3,990 (97.1%)
	Weekend	1,849 (2.8%)	1,097 (2.5%)	634 (3.5%)	118 (2.9%)
Holiday		240 (0.4%)	144 (0.3%)	79 (0.4%)	17 (0.4%)
Anesthesiology Resident Present *		44,045 (66.6%)	28,972 (66.0%)	11,801 (65.1%)	3,272 (79.6%)
Nurse Anesthetist Present *		13,117 (19.8%)	9,668 (22.0%)	3,060 (16.9%)	389 (9.5%)
Anesthesiology Attending Only		9,085 (13.7%)	5,320 (12.1%)	3,313 (18.3%)	452 (11.0%)
Institutional Characteristics					
Academic Hospital		64,920 (98.1%)	43,198 (98.3%)	18,111 (99.9%)	3,611 (87.9%)
Community Hospital		1,246 (1.9%)	728 (1.7%)	21 (0.1%)	497 (12.1%)
Outcome Characteristics					
CSA-AKI Stage					
	No CSA-AKI	49,264 (74.5%)	33,019 (75.2%)	12,901 (71.2%)	3,344 (81.4%)
	CSA-AKI-1	11,759 (17.8%)	7,771 (17.7%)	3,449 (19.0%)	539 (13.1%)
	CSA-AKI-2	3,457 (5.2%)	2,161 (4.9%)	1,142 (6.3%)	154 (3.7%)
	CSA-AKI-3	1,686 (2.5%)	975 (2.2%)	640 (3.5%)	71 (1.7%)

** Non-mutually exclusive*

Statistics presented as mean (SD) for numeric variables; N(%) for categorical variables. AIDS/HIV = acquired immunodeficiency syndrome / human immunodeficiency virus; AKI = acute kidney injury; ASA = American Society of Anesthesiologists; CPB = cardiopulmonary bypass; CSA-AKI = cardiac surgery-associated acute kidney injury; ETT = endotracheal tube; LMA = laryngeal mask airway

Supplemental Table 4.4 Temporal validation AUCs at individual centers.

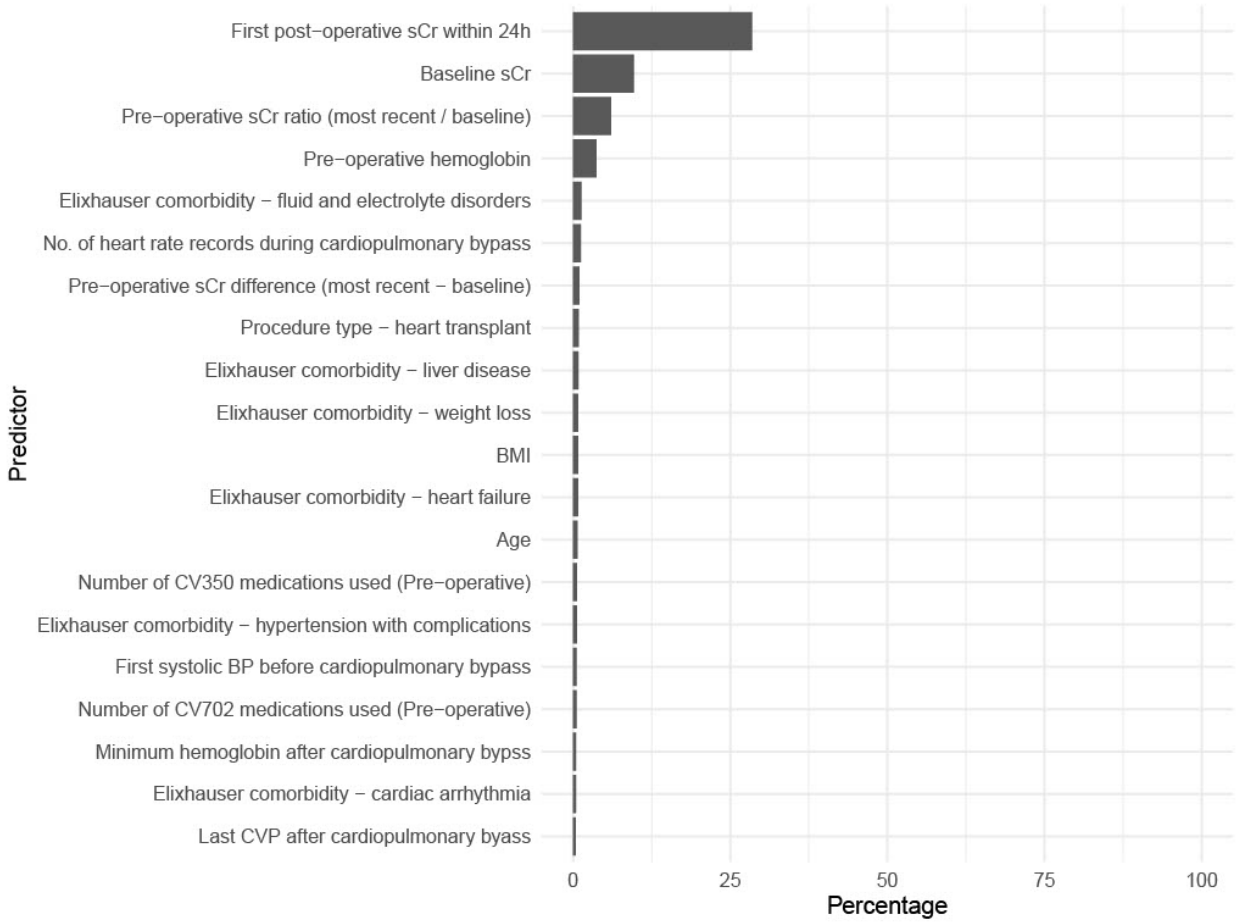
Outcome	Model	14	92	101	78	102	35	91	76	89	47	83	70	23	7	58	68	66	4	19	1	10	84	16
No. of training cases		0	0	0	0	0	31	36	57	90	168	280	303	700	1,105	1,356	1,721	1,769	2,139	2,140	2,473	3,345	6,318	16,251
No. of validation cases		744	20	548	370	473	521	275	94	724	58	88	205	326	266	918	1,374	560	972	825	1,956	1,403	1,941	3,469
AKI-1+	**Base**	NA	NA	NA	NA	NA	0.5412	0.444	0.649	0.6303	0.6636	0.6944	0.6134	0.7497	0.7825	0.7402	0.8268	0.7503	0.7576	0.7708	0.7771	0.8778	0.804	0.8857
	FSL	0.7861	0.726	0.8106	0.8571	0.8259	0.7579	0.8113	0.836	0.7565	0.7545	0.7805	0.7285	0.8095	0.8061	0.7935	0.8346	0.7648	0.7933	0.7954	0.8039	0.8784	0.8208	0.8797
	FSL (with metadata)	0.7881	0.667	0.7986	0.8543	0.8311	0.7581	0.8252	0.835	0.7602	0.8727	0.8246	0.7403	0.7912	0.8262	0.791	0.8428	0.7663	0.8007	0.7956	0.8068	0.8764	0.8185	0.8805
	Pooled	0.8377	0.667	0.8369	0.871	0.8459	0.8057	0.8347	0.87	0.8068	0.8818	0.8328	0.7573	0.8552	0.8407	0.8186	0.8703	0.7945	0.8282	0.8245	0.8347	0.9045	0.8378	0.8982
AKI-2+	**Base**	NA	NA	NA	NA	NA	0.5876	0.6315	0.527	0.6678	NA	0.7494	0.6377	0.7658	0.7923	0.8193	0.8805	0.7313	0.8104	0.7927	0.8212	0.9114	0.8396	0.9071
	FSL	0.8537	NA	0.8599	0.8722	0.8568	0.819	0.7682	0.691	0.8701	NA	0.7711	0.8262	0.842	0.908	0.8486	0.8803	0.727	0.8377	0.8225	0.8462	0.905	0.8339	0.8965
	FSL (with metadata)	0.8464	NA	0.8737	0.873	0.8685	0.8218	0.7973	0.722	0.8742	NA	0.7855	0.7884	0.8466	0.9104	0.8509	0.8816	0.744	0.8497	0.8154	0.8516	0.9076	0.8393	0.8944
	Pooled	0.8828	NA	0.8813	0.9022	0.8899	0.864	0.7414	0.722	0.8967	NA	0.841	0.8683	0.8581	0.9139	0.8774	0.9119	0.7917	0.8527	0.8608	0.8758	0.9298	0.8754	0.9219
AKI-3+	**Base**	NA	NA	NA	NA	NA	0.6774	0.7605	0.6	0.777	NA	0.8046	0.6851	0.7776	0.7602	0.8263	0.8696	0.7499	0.8393	0.8121	0.906	0.9243	0.8887	0.8879
	FSL	0.9282	1	0.8326	0.8512	0.8822	0.89	0.9272	0.874	0.924	NA	0.7471	0.769	0.8305	0.9425	0.8647	0.8911	0.7701	0.881	0.8443	0.906	0.9077	0.8837	0.9049
	FSL (with metadata)	0.9581	1	0.8271	0.8513	0.902	0.8729	0.9494	0.885	0.944	NA	0.8506	0.7807	0.8374	0.9333	0.8709	0.8778	0.8044	0.8859	0.8469	0.9009	0.92	0.889	0.9087
	Pooled	0.9796	1	0.791	0.8568	0.9164	0.8276	0.8926	0.874	0.9281	NA	0.7241	0.9058	0.848	0.9502	0.8499	0.9333	0.7923	0.9156	0.8678	0.9302	0.9479	0.9268	0.928

Supplemental Table 4.5 Learning curve analysis results for predicting AKI 1+.

No. of sites	Temporal AUC			External AUC		
	FSL mean (SD)	Pooled mean (SD)	Mean Difference (Pooled - FSL)	FSL mean (SD)	Pooled mean (SD)	Mean Difference (Pooled - FSL)
1	0.6840 (0.0533)	0.7283 (0.0773)	0.0443	0.6928 (0.0690)	0.7419 (0.0930)	0.0491
2	0.7196 (0.0500)	0.7800 (0.0476)	0.0604	0.7323 (0.0643)	0.7967 (0.0633)	0.0644
3	0.7425 (0.0440)	0.8032 (0.0279)	0.0607	0.7615 (0.0557)	0.8247 (0.0391)	0.0632
4	0.7627 (0.0348)	0.8153 (0.0194)	0.0526	0.7834 (0.0450)	0.8393 (0.0276)	0.0559
5	0.7771 (0.0236)	0.8240 (0.0147)	0.0469	0.7999 (0.0325)	0.8487 (0.0218)	0.0488
6	0.7849 (0.0193)	0.8286 (0.0128)	0.0437	0.8084 (0.0285)	0.8531 (0.0226)	0.0447
7	0.7917 (0.0156)	0.8332 (0.0103)	0.0415	0.8161 (0.0230)	0.8592 (0.0186)	0.0431
8	0.7974 (0.0127)	0.8361 (0.0091)	0.0387	0.8229 (0.0194)	0.8634 (0.0155)	0.0405
9	0.8009 (0.0116)	0.8388 (0.0077)	0.0379	0.8282 (0.0175)	0.8653 (0.0123)	0.0371
10	0.8048 (0.0102)	0.8409 (0.0073)	0.0361	0.8321 (0.0150)	0.8680 (0.0114)	0.0359
11	0.8082 (0.0094)	0.8430 (0.0066)	0.0348	0.8365 (0.0138)	0.8699 (0.0095)	0.0334
12	0.8099 (0.0088)	0.8445 (0.0058)	0.0346	0.8393 (0.0132)	0.8714 (0.0084)	0.0321
13	0.8124 (0.0083)	0.8461 (0.0053)	0.0337	0.8426 (0.0112)	0.8729 (0.0086)	0.0303
14	0.8139 (0.0086)	0.8473 (0.0050)	0.0334	0.8437 (0.0109)	0.8746 (0.0071)	0.0309
15	0.8159 (0.0078)	0.8484 (0.0044)	0.0318	0.8462 (0.0108)	0.8756 (0.0072)	0.0294
16	0.8178 (0.0074)	0.8496 (0.0041)	0.0318	0.8490 (0.0098)	0.8773 (0.0062)	0.0283
17	0.8197 (0.0070)	0.8507 (0.0038)	0.031	0.8516 (0.0082)	0.8781 (0.0057)	0.0265
18	0.8211 (0.0061)	0.8514 (0.0035)	0.0303	0.8536 (0.0084)	0.8789 (0.0049)	0.0253
19	0.8222 (0.0052)	0.8521 (0.0030)	0.0299	0.8554 (0.0065)	0.8800 (0.0042)	0.0246
20	0.8237 (0.0044)	0.8530 (0.0029)	0.0293	0.8576 (0.0049)	0.8805 (0.0035)	0.0229
21	0.8246 (0.0032)	0.8539 (0.0022)	0.0293	0.8592 (0.0040)	0.8812 (0.0030)	0.022
22	0.8259 (0.0017)	0.8553 (0.0012)	0.0294	0.8610 (0.0033)	0.8812 (0.0017)	0.0202
23	0.8261 (0)	0.8557 (0)	0.0296	0.8647 (0)	0.8824 (0)	0.0177

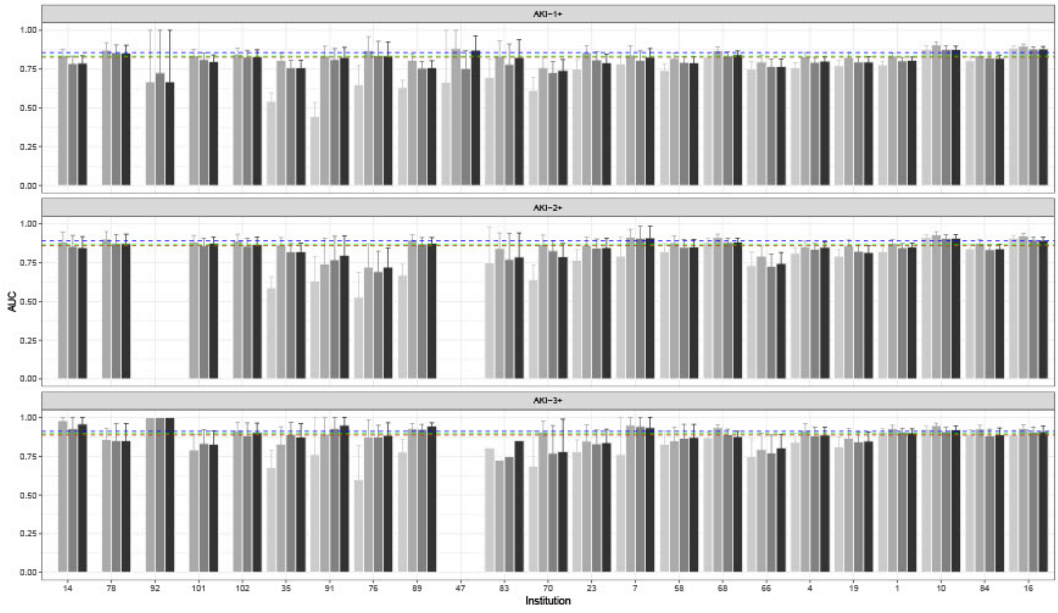
AUC = receiver operating characteristic area under curve; FSL = federated stacked learning

4.6.3 Supplemental Figures



Supplemental Figure 4.1 Feature importance plot of the pooled model.

Top 20 important features of the pooled model. Predictors are ranked by their relative importance and expressed as a percentage.



Outcome	Model	Overall	14	78	92	101	102	35	91	76	89	47	83	70	23	7	58	68	66	4	19	1	10	84	16
			Nt = 0	Nt = 0	Nt = 0	Nt = 0	Nt = 15	Nt = 31	Nt = 36	Nt = 57	Nt = 90	Nt = 168	Nt = 280	Nt = 303	Nt = 700	Nt = 1105	Nt = 1358	Nt = 1721	Nt = 1769	Nt = 2139	Nt = 2140	Nt = 2473	Nt = 3345	Nt = 6318	Nt = 16251
			Nv = 744	Nv = 370	Nv = 20	Nv = 548	Nv = 473	Nv = 521	Nv = 275	Nv = 94	Nv = 724	Nv = 58	Nv = 88	Nv = 205	Nv = 328	Nv = 266	Nv = 918	Nv = 1374	Nv = 580	Nv = 972	Nv = 825	Nv = 1956	Nv = 1403	Nv = 1941	Nv = 3469
AKI−1+	Base	NA	NA	NA	NA	NA	NA	0.5412	0.4440	0.6492	0.6303	0.6636	0.6944	0.6134	0.7497	0.7825	0.7402	0.8268	0.7503	0.7576	0.7708	0.7771	0.8778	0.8040	0.8857
	Pooled	0.8557	0.8377	0.8710	0.6667	0.8369	0.8459	0.8057	0.8347	0.8702	0.8068	0.8818	0.8328	0.7573	0.8552	0.8407	0.8186	0.8703	0.7945	0.8282	0.8245	0.8347	0.9045	0.8378	0.8982
	Federated	0.8261	0.7861	0.8571	0.7255	0.8106	0.8259	0.7579	0.8113	0.8357	0.7565	0.7545	0.7805	0.7285	0.8095	0.8061	0.7935	0.8346	0.7648	0.7933	0.7954	0.8039	0.8784	0.8208	0.8797
	Federated with metadata	0.8317	0.7881	0.8543	0.6667	0.7986	0.8311	0.7581	0.8252	0.8351	0.7602	0.8727	0.8246	0.7403	0.7912	0.8282	0.7910	0.8428	0.7663	0.8007	0.7958	0.8088	0.8764	0.8185	0.8805
AKI−2+	Base	NA	NA	NA	NA	NA	NA	0.5578	0.6315	0.5271	0.6678	NA	0.7494	0.6377	0.7658	0.7923	0.8193	0.8805	0.7313	0.8104	0.7927	0.8212	0.9114	0.8396	0.9071
	Pooled	0.8569	0.8828	0.9022	NA	0.8813	0.8899	0.8640	0.7414	0.7216	0.8967	NA	0.8410	0.8683	0.8581	0.9139	0.8774	0.9119	0.7917	0.8527	0.8808	0.8758	0.9298	0.8754	0.9219
	Federated	0.8510	0.8537	0.8722	NA	0.8599	0.8588	0.8190	0.7682	0.6910	0.8701	NA	0.7711	0.8262	0.8420	0.9080	0.8486	0.8803	0.7270	0.8377	0.8225	0.8462	0.9050	0.8339	0.8985
	Federated with metadata	0.8551	0.8464	0.8730	NA	0.8737	0.8685	0.8218	0.7973	0.7216	0.8742	NA	0.7855	0.7884	0.8466	0.9104	0.8509	0.8818	0.7440	0.8497	0.8154	0.8516	0.9076	0.8393	0.8944
AKI−3+	Base	NA	NA	NA	NA	NA	NA	0.6774	0.7605	0.6000	0.7770	NA	0.8046	0.6851	0.7776	0.7602	0.8283	0.8696	0.7499	0.8393	0.8121	0.9080	0.9243	0.8887	0.8879
	Pooled	0.9108	0.9796	0.8568	1	0.7910	0.9164	0.8276	0.8626	0.8742	0.9281	NA	0.7241	0.9058	0.8480	0.9502	0.8499	0.9333	0.7923	0.9156	0.8678	0.9302	0.9479	0.9288	0.9280
	Federated	0.8873	0.9282	0.8512	1	0.8326	0.8822	0.8900	0.9272	0.8742	0.9240	NA	0.7471	0.7690	0.8305	0.9425	0.8647	0.8911	0.7701	0.8810	0.8443	0.9080	0.9077	0.8837	0.9049
	Federated with metadata	0.8938	0.9581	0.8513	1	0.8271	0.9020	0.8729	0.9494	0.8954	0.9440	NA	0.8506	0.7807	0.8374	0.9333	0.8709	0.8778	0.8044	0.8859	0.8469	0.9009	0.9200	0.8890	0.9087

Supplemental Figure 4.2 Comparison of model performance (AUC) at each center among single-center model and multicenter models, for all AKI severities.

The bar plot at the top panel demonstrates visual comparison of AUCs of four examined models (base, FSL, FSL with metadata, and pooled) when tested in temporal validation set at each individual centers for all AKI severities. The dashed lines (blue: pooled, red: FSL, green: FSL with metadata) indicate the model performance in aggregate. The numeric values of model performance are shown in the table at the bottom panel.

Chapter 5 Conclusion

5.1 Summary

The overarching goal of this dissertation was to develop and evaluate machine learning models for AKI. Unlike many model development researches that push the limits of model performance, my research takes barriers in clinical model implementation into consideration and strives to develop models that are transportable, clinically applicable and scalable, while maintaining optimal performance.

In Chapter II, I evaluated the transportability of a reproduced version of a state-of-the-art AKI model across health systems. The AKI model originally developed by DeepMind for the VA health system showed high performance in predicting AKI within 48 hours. However, its generalizability faced challenges due to being trained in a predominantly male population. In this study, I approximated the DeepMind's GBDT AKI model and assessed its performance in a more sex-balanced patient population at the UM. Identifying suboptimal discrimination and calibration in females, I updated the model through continued training at UM to address this sex-related disparities. Furthermore, I investigated the potential reasons for this model discrepancy by sex and showed that it is complex and cannot be simply explained by a low sample size or difference in patient characteristics. This study contributes valuable evidence highlighting the existence of sex and gender inequalities in healthcare machine learning models and explores promising ways for mitigating such challenges through local fine-tuning of models.

In Chapter III, I investigated the pattern of urine output (UO) documentation in the UM EHR system and assessed the clinical applicability of UO as a predictor in an AKI risk prediction model. Utilizing a five-year inpatient cohort at the UM, I found UO documentation to be generally of high frequency and quality. I also identified three different phenotypes of UO documentation for non-ICU patients, revealing variations in UO monitoring and documentation across different hospital stays. Additionally, I evaluated the utility of incorporating UO in AKI risk prediction models, finding that while UO is valuable, its additive contribution is limited when integrated into an otherwise comprehensive AKI prediction model. This study underscores the challenges associated with UO documentation and emphasizes the need for ongoing efforts to enhance its consistency, providing valuable insights for refining AKI prediction strategies in diverse clinical contexts.

In Chapter IV, I introduced a new federated learning framework, federated stacked learning (FSL), designed to enhance the scalability of AKI models for potential multicenter modeling purposes. Focusing on the prediction of cardiac surgery-associated AKI within a national perioperative research network of 31 academic and community hospitals, I compared the performance of single-center models with pooled-data and the newly proposed FSL approaches. Contrary to conventional assumptions, the findings reveal that single-center models do not surpass multicenter approaches, highlighting the substantial benefits of multi-center AI models for individual centers, particularly smaller ones. While pooled models demonstrated the highest overall performance, the FSL approach achieved comparable performance to pooled models, making it an ideal solution, especially when patient-level data sharing is challenging. The efficiency of FSL remains consistent even with additional centers, making it a practical choice for implementation within research networks and improving model scalability. This study

underscores the significance of collaborative research networks, emphasizing the varied impacts of different modeling strategies on hospitals of different sizes within the network and stressing the importance of participation in collaborative efforts for both large and small health systems.

5.2 Future Directions

In the rapidly growing field of machine learning for healthcare, this dissertation has investigated critical dimensions of applying machine learning models for AKI. While several important areas have been studied and discussed, there remain untapped avenues that warrant exploration in future work.

One direction that future research should focus on is developing methods to enhance the interpretability and explainability of AKI machine learning models. While this dissertation presented GBDT models and provided features importance plots for model interpretability, a more in-depth exploration of interpretability and explainability is crucial. For instance, the novel Federated Stacked Learning (FSL) framework proposed in Chapter IV could benefit from detailed interpretation methods to explain how the algorithm leads to predictions based on the weights assigned to each center. Improving model transparency through interpretability and providing clinically meaningful explanations can foster trust among stakeholders.

The dissertation demonstrated that the FSL framework is an efficient and effective approach for collaborative efforts in building multi-center AKI models. However, future work should involve benchmarking existing federated learning algorithms in the same clinical scenario and comparing their performance and communication efficiency with the FSL. This benchmarking and comparison will contribute to a better understanding of the practicality of the FSL algorithm and identify areas for potential improvement.

Ideally, the developed and validated AKI models should be implemented in real-world clinical care to improve patient health. Future prospective studies or clinical trials can be designed to investigate the actual treatment effect brought about by using the model. Collecting feedback from stakeholders (e.g. clinicians, patients, etc.), is crucial for understanding end-users' perspectives and making improvements in model implementation and acceptance in clinical practice. Exploring the integration of developed models into real-time clinical decision support systems is essential, and collaboration with healthcare providers can facilitate the implementation of models that offer timely alerts and suggestions for patient care based on AKI risk predictions. Continuous monitoring and governance of the model in use are imperative if the model is implemented for practical use.

These future directions aim to advance the field by addressing aspects of interpretability, benchmarking, and real-world implementation, ensuring that machine learning models for AKI are transparent, practical, and widely applicable in diverse healthcare settings.

5.3 Conclusion

In conclusion, this dissertation represents a comprehensive exploration of critical dimensions in the application of AKI machine learning models. The identification and resolution of sex-related disparities in a leading AKI model underscore the significance of context-specific adjustments, enhancing the model's transportability. Uncovering both the pattern of UO data in the EHR and the challenges of integrating it into established AKI prediction models emphasize the ongoing need for refining UO documentation practices to augment its clinical applicability. The introduction of the novel FSL framework addresses the critical aspect of AKI model scalability. The superiority of multicenter AI models over single-center approaches, along with the practicality of FSL in collaborative research networks, further emphasizes the importance of

scalability in AKI modeling. Collectively, this dissertation work contributes not only valuable insights into specific areas of AKI prediction but also advocates for a pragmatic approach to model development that considers transportability, clinical utility, and scalability. The findings of this dissertation pave the way for future advancements in machine learning applications for AKI, promoting the development of models that are not only accurate but also accessible, generalizable, and adaptable across diverse healthcare settings.

Bibliography

1. Wang HE, Muntner P, Chertow GM, Warnock DG. Acute Kidney Injury and Mortality in Hospitalized Patients. *Am J Nephrol*. 2012;35(4):349-355. doi:10.1159/000337487

2. Al-Jaghbeer M, Dealmeida D, Bilderback A, Ambrosino R, Kellum JA. Clinical Decision Support for In-Hospital AKI. *J Am Soc Nephrol*. 2018;29(2):654-660. doi:10.1681/ASN.2017070765

3. Susantitaphong P, Cruz DN, Cerda J, et al. World Incidence of AKI: A Meta-Analysis. *Clin J Am Soc Nephrol*. 2013;8(9):1482-1493. doi:10.2215/CJN.00710113

4. Nisula S, Kaukonen KM, Vaara ST, et al. Incidence, risk factors and 90-day mortality of patients with acute kidney injury in Finnish intensive care units: the FINNAKI study. *Intensive Care Med*. 2013;39(3):420-428. doi:10.1007/s00134-012-2796-5

5. Srisawat N, Sileanu FE, Murugan R, et al. Variation in Risk and Mortality of Acute Kidney Injury in Critically Ill Patients: A Multicenter Study. *Am J Nephrol*. 2015;41(1):81-89. doi:10.1159/000371748

6. Hoste EAJ, Bagshaw SM, Bellomo R, et al. Epidemiology of acute kidney injury in critically ill patients: the multinational AKI-EPI study. *Intensive Care Med*. 2015;41(8):1411-1423. doi:10.1007/s00134-015-3934-7

7. Hirsch JS, Ng JH, Ross DW, et al. Acute kidney injury in patients hospitalized with COVID-19. *Kidney Int*. 2020;98(1):209-218. doi:10.1016/j.kint.2020.05.006

8. Chan L, Chaudhary K, Saha A, et al. AKI in Hospitalized Patients with COVID-19. *J Am Soc Nephrol*. 2021;32(1):151. doi:10.1681/ASN.2020050615

9. Abebe A, Kumela K, Belay M, Kebede B, Wobie Y. Mortality and predictors of acute kidney injury in adults: a hospital-based prospective observational study. *Sci Rep*. 2021;11(1):15672. doi:10.1038/s41598-021-94946-3

10. Coca SG, Singanamala S, Parikh CR. Chronic kidney disease after acute kidney injury: a systematic review and meta-analysis. *Kidney Int*. 2012;81(5):442-448. doi:10.1038/ki.2011.379

11. Odutayo A, Wong CX, Farkouh M, et al. AKI and Long-Term Risk for Cardiovascular Events and Mortality. *J Am Soc Nephrol*. 2017;28(1):377. doi:10.1681/ASN.2016010105

12. Villeneuve PM, Clark EG, Sikora L, Sood MM, Bagshaw SM. Health-related quality-of-life among survivors of acute kidney injury in the intensive care unit: a systematic review. *Intensive Care Med*. 2016;42(2):137-146. doi:10.1007/s00134-015-4151-0

13. Silver SA, Long J, Zheng Y, Chertow GM. Cost of Acute Kidney Injury in Hospitalized Patients. *J Hosp Med*. 2017;12(2):70-76. doi:10.12788/jhm.2683

14. MacLeod A. NCEPOD report on acute kidney injury—must do better. *The Lancet*. 2009;374(9699):1405-1406. doi:10.1016/S0140-6736(09)61843-2

15. Meersch M, Schmidt C, Hoffmeier A, et al. Prevention of cardiac surgery-associated AKI by implementing the KDIGO guidelines in high risk patients identified by biomarkers: the PrevAKI randomized controlled trial. *Intensive Care Med*. 2017;43(11):1551-1561. doi:10.1007/s00134-016-4670-3

16. Khwaja A. KDIGO Clinical Practice Guidelines for Acute Kidney Injury. *Nephron Clin Pract*. 2012;120(4):c179-c184. doi:10.1159/000339789

17. Luo X, Jiang L, Du B, et al. A comparison of different diagnostic criteria of acute kidney injury in critically ill patients. *Crit Care*. 2014;18(4):R144. doi:10.1186/cc13977

18. Kellum JA, Lameire N, for the KDIGO AKI Guideline Work Group. Diagnosis, evaluation, and management of acute kidney injury: a KDIGO summary (Part 1). *Crit Care*. 2013;17(1):204. doi:10.1186/cc11454

19. Bellomo R, Ronco C, Kellum JA, Mehta RL, Palevsky P, the ADQI workgroup. Acute renal failure – definition, outcome measures, animal models, fluid therapy and information technology needs: the Second International Consensus Conference of the Acute Dialysis Quality Initiative (ADQI) Group. *Crit Care*. 2004;8(4):R204. doi:10.1186/cc2872

20. Mehta RL, Kellum JA, Shah SV, et al. Acute Kidney Injury Network: report of an initiative to improve outcomes in acute kidney injury. *Crit Care*. 2007;11(2):R31. doi:10.1186/cc5713

21. Chertow GM, Burdick E, Honour M, Bonventre JV, Bates DW. Acute Kidney Injury, Mortality, Length of Stay, and Costs in Hospitalized Patients. *J Am Soc Nephrol*. 2005;16(11):3365-3370. doi:10.1681/ASN.2004090740

22. Sutherland SM, Chawla LS, Kane-Gill SL, et al. Utilizing Electronic Health Records to Predict Acute Kidney Injury Risk and Outcomes: Workgroup Statements from the 15th ADQI Consensus Conference: *Can J Kidney Health Dis*. Published online February 26, 2016. doi:10.1186/s40697-016-0099-4

23. Cronin RM, VanHouten JP, Siew ED, et al. National Veterans Health Administration inpatient risk stratification models for hospital-acquired acute kidney injury. *J Am Med Inform Assoc*. 2015;22(5):1054-1071. doi:10.1093/jamia/ocv051

24. Koyner JL, Adhikari R, Edelson DP, Churpek MM. Development of a Multicenter Ward–Based AKI Prediction Model. *Clin J Am Soc Nephrol*. 2016;11(11):1935-1943. doi:10.2215/CJN.00280116

25. 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(6):1052-1061. doi:10.1093/jamia/ocx030

26. Cheng P, Waitman LR, Hu Y, Liu M. Predicting Inpatient Acute Kidney Injury over Different Time Horizons: How Early and Accurate? *AMIA Annu Symp Proc*. 2018;2017:565-574.

27. Huang C, Murugiah K, Mahajan S, et al. Enhancing the prediction of acute kidney injury risk after percutaneous coronary intervention using machine learning techniques: A retrospective cohort study. *PLoS Med*. 2018;15(11):e1002703. doi:10.1371/journal.pmed.1002703

28. Koyner JL, Carey KA, Edelson DP, Churpek MM. The Development of a Machine Learning Inpatient Acute Kidney Injury Prediction Model*. *Crit Care Med*. 2018;46(7):1070-1077. doi:10.1097/CCM.0000000000003123

29. Mohamadlou H, Lynn-Palevsky A, Barton C, et al. Prediction of Acute Kidney Injury With a Machine Learning Algorithm Using Electronic Health Record Data. *Can J Kidney Health Dis*. 2018;5:2054358118776326. doi:10.1177/2054358118776326

30. He J, Hu Y, Zhang X, Wu L, Waitman LR, Liu M. Multi-perspective predictive modeling for acute kidney injury in general hospital populations using electronic medical records. *JAMIA Open*. 2019;2(1):115-122. doi:10.1093/jamiaopen/ooy043

31. 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. doi:10.1038/s41586-019-1390-1

32. Zimmerman LP, Reyfman PA, Smith ADR, et al. Early prediction of acute kidney injury following ICU admission using a multivariate panel of physiological measurements. *BMC Med Inform Decis Mak*. 2019;19(1):16. doi:10.1186/s12911-019-0733-z

33. Demirjian S, Bashour CA, Shaw A, et al. Predictive Accuracy of a Perioperative Laboratory Test–Based Prediction Model for Moderate to Severe Acute Kidney Injury After Cardiac Surgery. *JAMA*. 2022;327(10):956-964. doi:10.1001/jama.2022.1751

34. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Med Res Methodol*. 2014;14(1):137. doi:10.1186/1471-2288-14-137

35. Robbins R. AI systems are worse at diagnosing disease when training data is skewed by sex. STAT. Published May 25, 2020. Accessed September 23, 2023. https://www.statnews.com/2020/05/25/ai-systems-training-data-sex-bias/

36. Rank N, Pfahringer B, Kempfert J, et al. Deep-learning-based real-time prediction of acute kidney injury outperforms human predictive performance. *Npj Digit Med*. 2020;3(1):1-12. doi:10.1038/s41746-020-00346-8

37. 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. doi:10.1371/journal.pmed.1002861

38. Hodgson LE, Sarnowski A, Roderick PJ, Dimitrov BD, Venn RM, Forni LG. Systematic review of prognostic prediction models for acute kidney injury (AKI) in general hospital populations. *BMJ Open*. 2017;7(9):e016591. doi:10.1136/bmjopen-2017-016591

39. El Emam K. Methods for the de-identification of electronic health records for genomic research. *Genome Med*. 2011;3(4):25. doi:10.1186/gm239

40. Platt J, Kardia S. Public Trust in Health Information Sharing: Implications for Biobanking and Electronic Health Record Systems. *J Pers Med*. 2015;5(1):3-21. doi:10.3390/jpm5010003

41. Platt JE, Jacobson PD, Kardia SLR. Public Trust in Health Information Sharing: A Measure of System Trust. *Health Serv Res*. 2018;53(2):824-845. doi:10.1111/1475-6773.12654

42. Gulamali FF, Nadkarni GN. Federated Learning in Risk Prediction: A Primer and Application to COVID-19-Associated Acute Kidney Injury. *Nephron*. 2023;147(1):52-56. doi:10.1159/000525645

43. Rajendran S, Xu Z, Pan W, Ghosh A, Wang F. Data heterogeneity in federated learning with Electronic Health Records: Case studies of risk prediction for acute kidney injury and sepsis diseases in critical care. *PLOS Digit Health*. 2023;2(3):e0000117. doi:10.1371/journal.pdig.0000117

44. Hoste EAJ, Kellum JA, Selby NM, et al. Global epidemiology and outcomes of acute kidney injury. *Nat Rev Nephrol*. 2018;14(10):607-625. doi:10.1038/s41581-018-0052-0

45. Wilson FP, Shashaty M, Testani J, et al. Automated, electronic alerts for acute kidney injury: a single-blind, parallel-group, randomised controlled trial. *The Lancet*. 2015;385(9981):1966-1974. doi:10.1016/S0140-6736(15)60266-5

46. McCradden MD, Stephenson EA, Anderson JA. Clinical research underlies ethical integration of healthcare artificial intelligence. *Nat Med*. 2020;26(9):1325-1326. doi:10.1038/s41591-020-1035-9

47. Tomašev N, Harris N, Baur S, et al. Use of deep learning to develop continuous-risk models for adverse event prediction from electronic health records. *Nat Protoc*. 2021;16(6):2765-2787. doi:10.1038/s41596-021-00513-5

48. Google. EHR modeling framework. Published online 2021. Accessed September 23, 2023. https://github.com/google/ehr-predictions

49. Haibe-Kains B, Adam GA, Hosny A, et al. Transparency and reproducibility in artificial intelligence. *Nature*. 2020;586(7829):E14-E16. doi:10.1038/s41586-020-2766-y

50. McDermott MBA, Wang S, Marinsek N, Ranganath R, Foschini L, Ghassemi M. Reproducibility in machine learning for health research: Still a ways to go. *Sci Transl Med*. 2021;13(586):eabb1655. doi:10.1126/scitranslmed.abb1655

51. Stupple A, Singerman D, Celi LA. The reproducibility crisis in the age of digital medicine. *Npj Digit Med*. 2019;2(1):1-3. doi:10.1038/s41746-019-0079-z

52. Carter RE, Attia ZI, Lopez-Jimenez F, Friedman PA. Pragmatic considerations for fostering reproducible research in artificial intelligence. *Npj Digit Med*. 2019;2(1):1-3. doi:10.1038/s41746-019-0120-2

53. Singh K, Beam AL, Nallamothu BK. Machine Learning in Clinical Journals: Moving from Inscrutable to Informative. *Circ Cardiovasc Qual Outcomes*. 2020;13(10):e007491. doi:10.1161/CIRCOUTCOMES.120.007491

54. Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proc Natl Acad Sci*. 2020;117(23):12592-12594. doi:10.1073/pnas.1919012117

55. Singh K. ML4LHS/va-aki-model: Initial release. Published online September 30, 2022. doi:10.5281/zenodo.7129945

56. Sundararajan V, Henderson T, Perry C, Muggivan A, Quan H, Ghali WA. New ICD-10 version of the Charlson comorbidity index predicted in-hospital mortality. *J Clin Epidemiol*. 2004;57(12):1288-1294. doi:10.1016/j.jclinepi.2004.03.012

57. Hand DJ, Till RJ. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Mach Learn*. 2001;45(2):171-186. doi:10.1023/A:1010920819831

58. 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. doi:10.2307/2531595

59. Morris N. tboot: Tilted bootstrap. Published online 2020. https://github.com/njm18/tboot

60. Friedman JH. Greedy function approximation: A gradient boosting machine. *Ann Stat*. 2001;29(5):1189-1232. doi:10.1214/aos/1013203451

61. R Core Team. R: A language and environment for statistical computing. Published online 2022. https://www.R-project.org/

62. Singh K, Meyer SR. ML4LHS/gpmodels: Initial release. Published online October 7, 2022. doi:10.5281/zenodo.7158501

63. Fryda T, LeDell E, Gill N, et al. h2o: R Interface for the “H2O” Scalable Machine Learning Platform. Published online 2022. https://cran.r-project.org/web/packages/h2o/index.html

64. Pafka S. GBM Performance. Published online 2021. Accessed September 23, 2023. https://github.com/szilard/GBM-perf

65. World Health Organization. International Classification of Diseases (ICD). Published 2022. Accessed September 23, 2023. https://www.who.int/standards/classifications/classification-of-diseases

66. Eknoyan G. Emergence of the Concept of Acute Renal Failure. *Am J Nephrol*. 2002;22(2-3):225-230. doi:10.1159/000063766

67. Macedo E, Malhotra R, Bouchard J, Wynn SK, Mehta RL. Oliguria is an early predictor of higher mortality in critically ill patients. *Kidney Int*. 2011;80(7):760-767. doi:10.1038/ki.2011.150

68. Md Ralib A, Pickering JW, Shaw GM, Endre ZH. The urine output definition of acute kidney injury is too liberal. *Crit Care*. 2013;17(3):R112. doi:10.1186/cc12784

69. Kellum JA, Sileanu FE, Murugan R, Lucko N, Shaw AD, Clermont G. Classifying AKI by Urine Output versus Serum Creatinine Level. *J Am Soc Nephrol*. 2015;26(9):2231. doi:10.1681/ASN.2014070724

70. Prowle JR. Measurement of AKI biomarkers in the ICU: still striving for appropriate clinical indications. *Intensive Care Med*. 2015;41(3):541-543. doi:10.1007/s00134-015-3662-z

71. Vaara ST, Parviainen I, Pettilä V, et al. Association of oliguria with the development of acute kidney injury in the critically ill. *Kidney Int*. 2016;89(1):200-208. doi:10.1038/ki.2015.269

72. McIlroy DR, Argenziano M, Farkas D, Umann T, Sladen RN. Incorporating Oliguria Into the Diagnostic Criteria for Acute Kidney Injury After On-Pump Cardiac Surgery: Impact on Incidence and Outcomes. *J Cardiothorac Vasc Anesth*. 2013;27(6):1145-1152. doi:10.1053/j.jvca.2012.12.017

73. Tarvasmäki T, Haapio M, Mebazaa A, et al. Acute kidney injury in cardiogenic shock: definitions, incidence, haemodynamic alterations, and mortality. *Eur J Heart Fail*. 2018;20(3):572-581. doi:10.1002/ejhf.958

74. Song X, Yu ASL, Kellum JA, et al. Cross-site transportability of an explainable artificial intelligence model for acute kidney injury prediction. *Nat Commun*. 2020;11(1):5668. doi:10.1038/s41467-020-19551-w

75. Alfieri F, Ancona A, Tripepi G, et al. Continuous and early prediction of future moderate and severe Acute Kidney Injury in critically ill patients: Development and multi-centric, multi-national external validation of a machine-learning model. *PLOS ONE*. 2023;18(7):e0287398. doi:10.1371/journal.pone.0287398

76. Alfieri F, Ancona A, Tripepi G, et al. A deep-learning model to continuously predict severe acute kidney injury based on urine output changes in critically ill patients. *J Nephrol*. 2021;34(6):1875-1886. doi:10.1007/s40620-021-01046-6

77. Zhao BC, Lei SH, Yang X, et al. Assessment of prognostic value of intraoperative oliguria for postoperative acute kidney injury: a retrospective cohort study. *Br J Anaesth*. 2021;126(4):799-807. doi:10.1016/j.bja.2020.11.018

78. Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. KDD'96. AAAI Press; 1996:226-231.

79. Fryda T, LeDell E, Gill N, et al. h2o: R Interface for the "H2O" Scalable Machine Learning Platform. Published online 2022. https://cran.r-project.org/web/packages/h2o/index.html

80. Vanmassenhove J, Steen J, Vansteelandt S, et al. The importance of the urinary output criterion for the detection and prognostic meaning of AKI. *Sci Rep*. 2021;11(1):11089. doi:10.1038/s41598-021-90646-0

81. Solomon AW, Kirwan CJ, Alexander NDE, et al. Urine output on an intensive care unit: case-control study. *BMJ*. 2010;341:c6761. doi:10.1136/bmj.c6761

82. Jin K, Murugan R, Sileanu FE, et al. Intensive Monitoring of Urine Output Is Associated With Increased Detection of Acute Kidney Injury and Improved Outcomes. *CHEST*. 2017;152(5):972-979. doi:10.1016/j.chest.2017.05.011

83. Coalition for Health AI. Blueprint for Trustworthy AI Implementation Guidance and Assurance for Healthcare. Accessed October 23, 2023. https://www.coalitionforhealthai.org/papers/Blueprint%20for%20Trustworthy%20AI.pdf

84. Bedoya AD, Economou-Zavlanos NJ, Goldstein BA, et al. A framework for the oversight and local deployment of safe and high-quality prediction models. *J Am Med Inform Assoc*. 2022;29(9):1631-1636. doi:10.1093/jamia/ocac078

85. Van Calster B, Steyerberg EW, Wynants L, van Smeden M. There is no such thing as a validated prediction model. *BMC Med*. 2023;21(1):70. doi:10.1186/s12916-023-02779-w

86. Cao J, Zhang X, Shahinian V, et al. Generalizability of an acute kidney injury prediction model across health systems | Nature Machine Intelligence. *Nat Mach Intell*. 2022;4(12):1121-1129.

87. Justice AC, Covinsky KE, Berlin JA. Assessing the Generalizability of Prognostic Information. *Ann Intern Med*. 1999;130(6):515-524. doi:10.7326/0003-4819-130-6-199903160-00016

88. Finlayson SG, Subbaswamy A, Singh K, et al. The Clinician and Dataset Shift in Artificial Intelligence. *N Engl J Med*. 2021;385(3):283-286. doi:10.1056/NEJMc2104626

89. Corrales Compagnucci M, Wilson ML, Fenwick M, Forgó N, Bärnighausen T, eds. *AI in eHealth: Human Autonomy, Data Governance and Privacy in Healthcare*. 1st ed. Cambridge University Press; 2022. doi:10.1017/9781108921923

90. Spector-Bagdady K, De Vries RG, Gornick MG, Shuman AG, Kardia S, Platt J. Encouraging Participation And Transparency In Biobank Research. *Health Aff (Millwood)*. 2018;37(8):1313-1320. doi:10.1377/hlthaff.2018.0159

91. Rieke N, Hancox J, Li W, et al. The future of digital health with federated learning. *Npj Digit Med*. 2020;3(1):1-7. doi:10.1038/s41746-020-00323-1

92. Brisimi TS, Chen R, Mela T, Olshevsky A, Paschalidis ICh, Shi W. Federated learning of predictive models from federated Electronic Health Records. *Int J Med Inf*. 2018;112:59-67. doi:10.1016/j.ijmedinf.2018.01.007

93. Huang L, Shea AL, Qian H, Masurkar A, Deng H, Liu D. Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. *J Biomed Inform*. 2019;99:103291. doi:10.1016/j.jbi.2019.103291

94. Loftus TJ, Ruppert MM, Shickel B, et al. Federated learning for preserving data privacy in collaborative healthcare research. *Digit Health*. 2022;8:20552076221134455. doi:10.1177/20552076221134455

95. Bai X, Wang H, Ma L, et al. Advancing COVID-19 diagnosis with privacy-preserving collaboration in artificial intelligence. *Nat Mach Intell*. 2021;3(12):1081-1089. doi:10.1038/s42256-021-00421-z

96. Dayan I, Roth HR, Zhong A, et al. Federated learning for predicting clinical outcomes in patients with COVID-19. *Nat Med*. 2021;27(10):1735-1743. doi:10.1038/s41591-021-01506-3

97. Roth HR, Chang K, Singh P, et al. Federated Learning for Breast Density Classification: A Real-World Implementation. In: Albarqouni S, Bakas S, Kamnitsas K, et al., eds. *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*. Lecture Notes in Computer Science. Springer International Publishing; 2020:181-191. doi:10.1007/978-3-030-60548-3_18

98. Pati S, Baid U, Edwards B, et al. Federated learning enables big data for rare cancer boundary detection. *Nat Commun*. 2022;13(1):7346. doi:10.1038/s41467-022-33407-5

99. Youssef A, Pencina M, Thakur A, Zhu T, Clifton D, Shah NH. External validation of AI models in health should be replaced with recurring local validation. *Nat Med*. Published online October 18, 2023:1-2. doi:10.1038/s41591-023-02540-z

100. Colquhoun DA, Shanks AM, Kapeles SR, et al. Considerations for Integration of Perioperative Electronic Health Records Across Institutions for Research and Quality Improvement: The Approach Taken by the Multicenter Perioperative Outcomes Group. *Anesth Analg*. 2020;130(5):1133-1146. doi:10.1213/ANE.0000000000004489

101. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Med*. 2015;13(1):1. doi:10.1186/s12916-014-0241-z

102. Sun E, Mello MM, Rishel CA, et al. Association of Overlapping Surgery With Perioperative Outcomes. *JAMA*. 2019;321(8):762-772. doi:10.1001/jama.2019.0711

103. Sun EC, Mello MM, Vaughn MT, et al. Assessment of Perioperative Outcomes Among Surgeons Who Operated the Night Before. *JAMA Intern Med*. 2022;182(7):720-728. doi:10.1001/jamainternmed.2022.1563

104. Mathis MR, Naik BI, Freundlich RE, et al. Preoperative Risk and the Association between Hypotension and Postoperative Acute Kidney Injury. *Anesthesiology*. 2020;132(3):461-475. doi:10.1097/ALN.0000000000003063

105. Inker LA, Eneanya ND, Coresh J, et al. New Creatinine- and Cystatin C–Based Equations to Estimate GFR without Race. *N Engl J Med*. 2021;385(19):1737-1749. doi:10.1056/NEJMoa2102953

106. Pirracchio R, Mavrothalassitis O, Mathis M, Kheterpal S, Legrand M. Response of US hospitals to elective surgical cases in the COVID-19 pandemic. *Br J Anaesth*. 2021;126(1):e46-e48. doi:10.1016/j.bja.2020.10.013

107. Ross C. Epic overhauls popular sepsis algorithm criticized for faulty alarms. STAT. Published October 3, 2022. Accessed November 1, 2023. https://www.statnews.com/2022/10/03/epic-sepsis-algorithm-revamp-training/

108. Laan MJ van der, Polley EC, Hubbard AE. Super Learner. *Stat Appl Genet Mol Biol*. 2007;6(1). doi:10.2202/1544-6115.1309

109. Breiman L. Stacked regressions. *Mach Learn*. 1996;24(1):49-64. doi:10.1007/BF00117832

110. ggbreak: set axis breaks for 'ggplot2.' Published online October 4, 2023. Accessed November 1, 2023. https://github.com/YuLab-SMU/ggbreak

111. Xu S, Chen M, Feng T, Zhan L, Zhou L, Yu G. Use ggbreak to Effectively Utilize Plotting Space to Deal With Large Datasets and Outliers. *Front Genet*. 2021;12. Accessed November 1, 2023. https://www.frontiersin.org/articles/10.3389/fgene.2021.774846

112. Van Calster B, McLernon DJ, van Smeden M, et al. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17(1):230. doi:10.1186/s12916-019-1466-7