Development and validation of models to predict the risk of major cardiovascular events and death for people with kidney failure having non-cardiac surgery

by

Gurpreet S. Pabla

A thesis submitted to the Faculty of Graduate Studies of the University of Manitoba
in partial fulfillment of the requirements of the degree of

MASTER OF SCIENCE

Department of Community Health Sciences
Max Rady College of Medicine
Rady Faculty of Health Sciences
University of Manitoba
Winnipeg

Abstract

Introduction: Patients with kidney failure undergoing non-cardiac surgery have a significantly higher risk of adverse cardiovascular events and mortality than the general population. Existing risk prediction tools have limited accuracy in estimating the risk of major post-operative events for people with kidney failure. To close this gap, Harrison et al. developed three risk prediction models to predict the risk of major post-operative events in individuals with kidney failure (AKDN models). These models, however, require external validation to ensure their reliability and accuracy in different populations.

Objective: The objectives of this research were to 1) externally validate the established AKDN models and 2) develop and validate a new risk prediction model using machine learning methods that can predict the risk of major cardiovascular events and mortality in patients with kidney failure within 30 days of undergoing outpatient or inpatient non-cardiac surgery in the Manitoba population.

Methods: Data was sourced from Manitoba Health. The cohort included adults (≥ 18 years) with pre-existing kidney failure (estimated glomerular filtration rate [eGFR] < 15 mL/min/1.73m^2 or receipt of maintenance dialysis) undergoing non-cardiac surgery procedures between April 1, 2007, and December 31, 2019. The primary outcome of this study was a composite acute myocardial infarction (AMI), cardiac arrest, ventricular arrhythmia, and all-cause mortality. The performance of the AKDN models was evaluated through Area Under the Receiver Operating Characteristic Curve (AUC-ROC), Area Under the Precision Recall Curve (AUC-PR), calibration, Brier score and other metrices on Manitoba data using two approaches:

1. Model deployment: The coefficients from AKDN models were extracted and then used to make predictions on Manitoba data.
2. Model refitting: We used logistic regression to re-estimate the coefficients of the models using Manitoba data while maintaining the same variable structure as used in the AKDN models.

To develop a machine learning model, data was split into 70% for training, 15% for validation, and 15% for testing. The training set was used to tune the hyperparameters and train the models; the validation dataset was used for feature selection and evaluate model performance, while the testing set evaluated the model's final performance. We used XGBoost and random forest, selecting a model with reasonable and balanced AUC-ROC and AUC-PR. The final model was externally validated using Alberta data.

Results: We identified 12,082 surgeries performed in 4,175 participants and observed 569 outcomes (5%). Models discriminated well, with AUC-ROC ranging from 0.821 for model 1 to 0.874 for model 3. The calibration slopes for models 1, 2, and 3 were 1.32, 1.40, and 1.24, respectively. Model refitting improved discrimination and calibration, with AUC-ROC ranging from 0.830 (model 1) to 0.861 (model 3). Calibration slopes were 1 across all the refit models. Brier scores were consistently low for all the models in both approaches. Using the machine learning approach, the final model (XGBoost) included surgery type, surgery setting (emergency inpatient, outpatient), history of myocardial infarction, albumin, and hemoglobin levels. It had an estimated AUC-ROC of 0.861, an AUC-PR of 0.304, and a Brier score of 0.038 in the testing cohort. External testing in Alberta showed similar performance. Calibration plots demonstrated excellent calibration, except for under-estimation at the highest predicted risks.

Conclusion: This external validation confirms the discrimination and calibration of the AKDN models in an independent population. Our new machine learning models also demonstrated good performance, and were more parsimonious than existing AKDN models. Future work should focus on implementation of these tools to test the impact of risk-guided approaches to perioperative care.

Acknowledgments

I extend my deep gratitude to Dr. Navdeep Tangri for guidance and support throughout the course of my thesis. I would also like to express my heartfelt thanks to the committee members, Dr. Tyrone Gorden Harrison and Dr. Lisa Lix, for their time and constructive feedback.

I also thank Reid Whitlock and Thomas Ferguson for their guidance in preparing the cohort. Additionally, I am thankful to Dr. Tyrone Gorden Harrison and Emir Sevinc for externally validating the models on Alberta data.

The authors acknowledge the Manitoba Centre for Health Policy for use of data contained in the Manitoba Population Research Data Repository under project # P2024-23. The results and conclusions are those of the authors and no official endorsement by the Manitoba Centre for Health Policy, Manitoba Health, or other data providers is intended or should be inferred. Data used in this study are from the Manitoba Population Research Data Repository housed at the Manitoba Centre for Health Policy, University of Manitoba and were derived from data provided by Manitoba Health.

Table of Contents

List of tables

List of Figures

Table of Abbreviations

Abbreviation	Full Name
ACS	American College of Surgery
AKDN	Alberta Kidney Disease Network
AMI	Acute Myocardial Infarction
ASA	American Society of Anesthesiologists
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
AUC-PR	Area Under the Precision Recall Curve
AVF	Arteriovenous Fistula
BNP/NT-proBNP	Brain Natriuretic Peptide and N-Terminal pro B-type Natriuretic Peptide
BUPA	British United Provident Association
CALIBRA	CArdiovascular, LIterature-Based, Risk Algorithm
CCI	Canadian Classification of Health Interventions
CEPOD	Confidential Enquiry into Perioperative Deaths
CI	Confidence Interval
CKD	Chronic Kidney Disease
CKD G5	Chronic Kidney Disease Category 5
CKD G5D	Chronic Kidney Disease Stage 5 on Dialysis
CKD G5ND	Chronic Kidney Disease Stage 5 Not on Dialysis
CKD-EPI	Chronic Kidney Disease Epidemiology Collaboration equation
C-statistic	Concordance Statistic
CV	Cardiovascular
CVD	Cardiovascular disease
DCA	Decision Curve Analysis
eGFR	Estimated Glomerular Filtration Rate
EGS	Emergency General Surgeries
GFR	Glomerular Filtration Rate
HAD	Hospital Abstracts Database
HDL	High-Density Lipoprotein
HIV	Human immunodeficiency viruses
ICD	International Classification of Diseases
ICD-9-CM	Ninth Revision, Clinical Modification of ICD
ICD-10-CA	Canadian adaptation of the ICD-10

IQR	Interquartile Range
IQRs	Interquartile Ranges
KDIGO	Kidney Disease: Improving Global Outcomes
LDL	Low-Density Lipoprotein
MACE	Major Adverse Cardiovascular Events
MCHP	Manitoba Centre for Health Policy
MDI	Mean Decrease in Impurity
MH	Manitoba Health
MHPR	Manitoba Health Population Registry
MICA	Myocardial Infarction or Cardiac Arrest
MSD	Medical Services Database
NARI	Net Absolute Reclassification Index
NBC	Naive-Bayes Classifier
NRI	Net Reclassification Improvement
NPV	Negative Predictive Value
NSQIP	National Surgical Quality Improvement Program
PD	Peritoneal Dialysis
PHIN	Personal Health Identification Number
PHRPC	Provincial Health Research Privacy Committee
P-POSSUM	Portsmouth Physiological and Operative Severity Score for the enUmeration of Mortality and Morbidity
PPV	Positive Predictive Value
RCRI	Revised Cardiac Risk Index
SAS	Statistical Analysis System
SHDS	Shared Health Diagnostic Services
SD	Standard Deviation
SRS	Surgical Risk Scale
UK	United Kingdom
US	United States
VIMP	Variable Importance
XGBoost	eXtreme Gradient Boosting

Introduction:

Chronic Kidney Disease and the Risk of Perioperative Events

Chronic Kidney Disease (CKD) is a condition when the kidneys slowly stop working over time. The kidneys play a crucial role in filtering blood, removing waste products, and maintaining the body's fluid and electrolyte balance. When the kidneys do not function effectively, harmful levels of waste can accumulate in the body, leading to various health conditions. According to Kidney Disease: Improving Global Outcomes (KDIGO), CKD is defined by changes in categories of glomerular filtration rate (GFR) and albuminuria (Figure 1), or abnormalities in kidney structure.[1]

Figure 1: CKD definition based on GFR and albuminuria category.[1]

Prognosis of CKD by GFR and albuminuria category

Prognosis of CKD by GFR and Albuminuria Categories: KDIGO 2012				Persistent albuminuria categories Description and range		
				A1	A2	A3
				Normal to mildly increased	Moderately increased	Severely increased
				<30 mg/g <3 mg/mmol	30-300 mg/g 3-30 mg/mmol	>300 mg/g >30 mg/mmol
GFR categories (ml/min/ 1.73 m^2) Description and range	G1	Normal or high	≥90			
	G2	Mildly decreased	60-89			
	G3a	Mildly to moderately decreased	45-59			
	G3b	Moderately to severely decreased	30-44			
	G4	Severely decreased	15-29			
	G5	Kidney failure	<15			

Green: low risk (if no other markers of kidney disease, no CKD); Yellow: moderately increased risk; Orange: high risk; Red, very high risk.

CKD is a growing problem in Canada and globally[2], contributing to 1.2 million deaths worldwide. Between 1990 and 2017, the mortality rate from CKD increased by 42%.[3] The prevalence rate of end stage kidney disease among Manitobans is 1530 per million population, which is higher than the average Canadian rate (1193 per million population).[4] The prevalence of kidney failure (CKD G5) has been increasing over several decades[3] with over 50,000 Canadians affected in 2020.[2] Most patients with kidney failure experience severe symptoms and complications and thus require ongoing dialysis treatment or a kidney transplant.

The Lancet Commission on Global Surgery estimated around 313 million surgeries worldwide yearly.[5] Most of the surgeries occur in developed or high-income countries, and 6% in low-income countries. In Canada, about 2.8 million surgeries were performed in 2019.[6] During the same period, there were about 101,225 surgeries performed in Manitoba. The number of surgeries decreased in Manitoba and Canada during the pandemic.[7] Having a surgery has an underlying risk associated with it. A recent report from the Lancet Commission on Global Surgery suggested that at least 4.2 million die worldwide annually within 30 days of surgery which makes it the third leading cause of mortality worldwide, after ischaemic heart disease and stroke.[8]

A systematic review and meta-analysis showed an increased odds ratio of post-operative mortality in patients on dialysis compared to those with normal kidney function.[9] The odds ratio ranged from 3.96 (3.23-4.87) after vascular surgery to 10.76 (7.30-15.86) after orthopaedic surgery.[9] In addition, the incidence of major adverse cardiovascular events at baseline (without surgery) in adults with kidney failure is 16 times higher than in adults with normal Estimated Glomerular Filtration Rate (eGFR).[10] Surgical procedures, whether outpatient or inpatient, can further exacerbate these risks by placing additional stress on the body.[9,11,12]

A large cohort study carried out in the US utilizing the National Surgical Quality Improvement Program (NSQIP) database found that dialysis patients were more likely to

develop post-operative complications. Dialysis patients were 3.92 times more likely to have a post-operative myocardial infarction (2.08, 7.39), and 9.41 times more likely to die within 30 days post surgery (8.05, 11.01).[13] The numbers were attenuated after adjusting for a few covariates but remained statistically significant. Another study done on chronic hemodialysis patients who had abdominal surgery found that morbidity and mortality rates were 39% & 14% respectively.[14] In a colorectal surgery study conducted from 1993 to 2007, the risk of mortality was 4.83 times higher for dialysis recipients compared to patients not on dialysis.[15] So, the existing studies established that individuals with kidney failure have poor outcomes after surgeries.[9,11]

Enhanced care during the perioperative period, which encompasses the preoperative, intraoperative, and postoperative phases of surgery, is crucial for ensuring patient safety and optimizing outcomes. By making informed decisions during this period, healthcare providers can significantly mitigate the risk of post-surgery complications. Risk prediction models are rules or algorithms designed to predict the probability of a specific event for a given individual based on a set of predictors. These models aid in clinical decision-making by forecasting patient outcomes based on certain risk factors. Effective risk prediction tools, when used during the perioperative period, can provide critical information that, when acted upon, has the potential to inform strategies to mitigate adverse outcomes post-surgery, leading to improved patient safety and recovery.

Internal and external validation in risk prediction models:

Developing a risk prediction model requires both internal and external validation to ensure the models are accurate and generalizable.[16] Internal validity refers to the degree to which a study accurately establishes a relationship between the predictor variables and the outcome. It needs a study that is well-designed i.e. it should have clearly defined predictor variables that are precisely measurable, and the outcome should be unambiguously defined and assessed equally among all participants and have sufficient power. The models should be derived using appropriate statistical or machine learning approaches. The selection of an

algorithm should be appropriate for the characteristics of the data. For example, logistic regression is preferred when follow-up period is well defined and there is minimal censoring. However, when the timing of an event is a crucial factor, it may be advantageous to adopt a survival analysis method. Internally validating a prediction model involves comparing the predicted risks to the actual observed outcomes in a population. There are two important elements to assess the model performance: calibration and discrimination. Calibration assesses how well the predicted probabilities agree with the observed outcomes. Calibration measures includes calibration-in-the-large, calibration plot and calibration slope. Discrimination is the ability of the model to distinguish between patients who do and do not experience the outcome and can be assessed by computing AUC-ROC, AUC-PR, Brier score etc.[16-19]

External validation verifies the model's generalizability, as it assesses how well the model performs in patient populations not previously involved in its development. This verification is crucial as it provides evidence that the model can effectively predict outcomes across diverse populations, making it more robust and suitable for broader clinical adoption.[20] A clinical example by Bleeker et al.[21] includes a case study to illustrate the limitations of internal validation. It describes how a prediction model's performance decreased significantly when applied to a data from extra hospitals (validation set), making the model less useful for pediatric care. The study emphasized that internal validation alone is not sufficient to determine the generalizability of a prediction model to future settings. External validation, which involves testing the model on a different population, is necessary to ensure its performance in real-world scenarios. Cohorts for external validation differ fundamentally from development cohort and can include temporal, geographic and external domain validation. Temporal validation means testing the model on data from the same population but from a different time period. Geographic validation means testing the model on data from a different geographic location. Different domains include different hospital or care settings, and this can even be extended to using the model to predict different outcomes or transformations of the initial outcome.[16,17]

Machine learning:

Machine learning is a component of artificial intelligence that allows computers to learn from and make decisions based on data. It uses algorithms to learn patterns in data and make predictions or decisions. Machine learning can be broadly categorized into three main types: supervised learning, unsupervised learning, and reinforcement learning.

1. Supervised Learning: This type involves training an algorithm on a labeled dataset to predict outcomes by recognizing patterns. Supervised learning is often used for classification and regression tasks, such as predicting the risk of an event based on input features.
2. Unsupervised Learning: In unsupervised learning, the algorithm is trained on unlabeled datasets. The goal is to discover inherent patterns or groupings in the data, such as clustering patients based on similarities in their medical records. It's useful for exploratory data analysis or identifying hidden patterns.
3. Reinforcement Learning: This type involves an algorithm learning to make decisions by taking certain actions in an environment to achieve a goal. It learns from the consequences of its actions, rather than from explicit training data, and adjusts its strategy to maximize desired behaviors.

Medical experts are increasingly relying on risk-assessment models to make decisions. It is essential to validate prediction models externally before their widespread adoption. This is because these models often do not perform well outside their initial development cohort. Flawed risk prediction models can lead to misallocation of healthcare resources, diverting attention and interventions from high-risk patients to those who may not need them. Additionally, these models might fail to recognize patients needing critical care, increasing the risk of preventable adverse events in conditions like acute heart failure.[22,23] Many prediction models are available, and only a fraction undergoes external validation. While creating a new model might seem attractive, many of these models remain unused in clinical practice. Validating existing models externally can help optimize their utility in real-world scenarios.[17,20]

Machine learning in healthcare is transforming the way medical professionals and researchers approach diagnosis, treatment planning, and patient care. Machine learning-based methods have been adopted to enhance the efficiency in managing electronic health records, detecting abnormalities in blood samples, as well as in organs and bones through medical imaging and monitoring. Machine learning techniques were also applied to a broad spectrum of clinical operations, ranging from diagnostic tasks like automated skin lesion classification and arrhythmia detection, to predictive tasks such as forecasting 30-day hospital readmissions. Moreover, the application of machine learning has significantly expedited testing and hospital responses during the COVID-19 crisis.[24,25] The application of machine learning in clinical settings is on the rise[26] and have shown success in various medical fields, including predicting disease onset.[27] This is because many machine learning algorithms can naturally capture non-linear relationships between predictors and outcomes.[28] Several studies show that machine learning can improve the performance in clinical outcome prediction[29-32] provided they are properly validated.[33,34] However, a recent systematic review found that machine learning offers no performance benefit over logistic regression in clinical prediction models.[35] However, many studies in this systematic review had a relatively small number of events for each predictor, which is a problem.[35] To the best of our knowledge, and following an extensive review of the current literature, there appear to be no risk prediction models that use machine learning algorithms for the prediction of major cardiovascular events and mortality. This gap underscores the potential impact of our study, which aimed to fill this void by leveraging machine learning algorithms.

Literature review

Models to predict MACE outcomes in the general population after non-cardiac surgery:

The development of a risk prediction model for major CV events and death in patients who undergo non-cardiac surgery is an important area of research. A retrospective study analyzed data on emergency general surgeries (EGS) performed in England between 2004 and 2019 to compare outcomes of patients with kidney failure (either on dialysis or with a kidney transplant) to the general population.[36] The study found that mortality after EGS was higher for patients

with kidney failure, with CV events being the most common cause of death. Dialysis patients had a higher risk for major adverse cardiovascular events (MACEs) after emergency laparotomy in the post-operative interval. The study suggests that kidney failure is a contributing risk factor for mortality after EGS, and the risk may differ between patients on dialysis and with a kidney transplant compared to the general population. There are several risk prediction tools (Risk Stratification Tools for Predicting Morbidity and Mortality in Adult Patients Undergoing Major Surgery) available to predict morbidity and mortality in adult patients undergoing major non-cardiac surgery for the general poulation. North American preoperative risk stratification guidelines recommend some of the tools which are discussed below[37]:

1. P-POSSUM, or Portsmouth Physiological and Operative Severity Score for the enUmeration of Mortality and Morbidity uses physiological score and an operative severity score to calculate risks of mortality and morbidity.[38] It includes a set of physiological variables such as age, cardiac status, respiratory status, blood pressure, and other health parameters and operative variables like operative severity, number of procedures, and mode of surgery. P-POSSUM has been extensively studied and validated in clinical research to assess its accuracy in predicting outcomes for surgical patients, but it does not include specific predictors relevant to patients with CKD undergoing surgery.
2. The Surgical Risk Scale (SRS) is a tool used to assess the risk of surgical procedures and was calculated for each procedure by adding Confidential Enquiry into Perioperative Deaths (CEPOD), American Society of Anesthesiologists (ASA) and British United Provident Association (BUPA) scores.[39] However, it is a risk prediction tool similar to P-POSSUM and does not include predictors for CKD patients.
3. The CArdiovascular, LIterature-Based, Risk Algorithm (CALIBRA) is a literature-based naive-bayes classifier (NBC) designed to predict the risk of cardiovascular hospitalizations among patients with kidney failure.[40] An NBC is an algorithm based on Bayes' theorem. The calculator incorporates 31 predictors, ranging from patient-specific factors like age and sex to CVD risk indicators such as Low-Density Lipoprotein (LDL) and High-Density Lipoprotein (HDL) cholesterol levels. It also accounts for comorbidities and

various other risk factors, including those specifically related to chronic kidney disease (i.e. kidney function measures eGFR, albuminuria or proteinuria, uric acid, C-reactive protein, among others). However, it only considers patients with non-dialysis dependent chronic kidney disease in the model. Also, it does not include surgical variables. This means that any risks associated with surgeries, either related or unrelated to kidney disease, are not directly considered in the risk predictions.

4. The Revised Cardiac Risk Index (RCRI) provides guidelines for perioperative cardiac risk assessment and management for patients undergoing non-cardiac surgery.[41] It includes six independent predictors: high-risk surgical procedure (includes intraperitoneal, intrathoracic, or suprainguinal vascular surgery), patient's history of coronary artery disease, cerebrovascular disease, congestive heart failure, preoperative insulin use, preoperative creatinine > 177 µmol/L. The tool allocates either 0 or 1 point for each of six criteria, and the total score correlates with expected post-operative outcomes. One study found that the chances of experiencing a myocardial infarction, cardiac arrest, or death were 3.9% (95% CI, 2.8% to 5.4%) for those with an RCRI score of 0, 6.0% (95% CI, 4.9% to 7.4%) for those with an RCRI score of one, 10.1% (95% CI, 8.1% to 12.6%) for an RCRI score of 2, and 15.0% (95% CI, 11.1% to 20.0%) for individuals with an RCRI score of 3 or above.[35] Even though RCRI scores are widely used, their value in estimating the risk of major post-operative events for those with kidney failure is limited. The tool includes serum creatinine as a binary variable, and so cannot differentiate between risk for moderate CKD and kidney failure.[42] It does not estimate patient's kidney function (eGFR), which is becoming more popular than serum creatinine among nephrology community as it includes patient's age and sex in the calculation. An external validation study amongst people with kidney failure and undergoing surgery, found this tool overestimated the risk of major cardiac events or death within 30 days of surgery. It suggested that the use of RCRI scores should be carefully considered for patients with kidney failure.[43]
5. NSQIP Myocardial Infarction or Cardiac Arrest (MICA) tool is also widely used to estimate risk of perioperative myocardial infarction or cardiac arrest after surgery.[44] It includes the

variables of American Society of Anesthesiologists (ASA) class, dependent functional status, age, abnormal creatinine (>1.5 mg/dL), and type of surgery. The risk model was validated on a separate dataset and showed high predictive performance, surpassing that of the RCRI. The tool uses a dichotomized variable for creatinine above >1.5 mg/dL, it does not discriminate risk between moderate chronic kidney disease and kidney failure.[42]

6. National Surgical Quality Improvement Program (NSQIP) American College of Surgery (ACS) is an online tool and is a widely used surgical risk calculator to estimate the risk of most operations.[45] It takes into account 21 preoperative factors including demographics, comorbidities, and procedure. The regression models developed for the tool predict eight outcomes within the 30-day post-operative period based on these preoperative risk factors. This calculator is based on data gathered from 1.4 million patients across 400 ACS NSQIP hospitals between 2009 and 2012. It incorporates preoperative dialysis variable; however, it does not distinguish between patients undergoing maintenance dialysis and those experiencing acute kidney injury.

A recent systematic review that aimed to evaluate the accuracy and generalizability of existing prediction models found that existing prediction models had AUC-ROC ranging from 0.710 to 0.752 in predicting mortality risk in chronic dialysis patients.[46] Based on the calibration results, most models tended to overestimate the likelihood of mortality. It means that the models were not always generalizable to different populations, underscoring the significance of conducting external validation.[46] While the risk of CVD and mortality in patients with kidney failure undergoing surgery has been studied extensively, there is a lack of consensus on the best approach to manage these patients perioperatively. Therefore, there is a need for a more accurate and parsimonious risk prediction tool to predict the risk of major cardiac events in patients with kidney failure following surgery. Such a tool could guide preoperative planning and optimization, and post-operative management.

Alberta Kidney Disease Network (AKDN) models for predicting MACE events after non-cardiac surgery in adults with kidney failure:

Three novel models to predict major clinical events for people with kidney failure having surgery were developed and internally validated using the Alberta Kidney Disease Network (AKDN) database.[42] The study used a retrospective, population-based cohort from Alberta, Canada, covering surgeries between 2005-2019. Kidney failure was defined as an estimated glomerular filtration rate (eGFR) of less than 15 mL/min/1.73m^2 or those who are on maintenance dialysis (hemodialysis or peritoneal dialysis). The outcomes were death or major cardiac events within 30 days post-surgery, which included acute myocardial infarction (AMI) and non-fatal ventricular arrhythmias. The cohort consisted of 38,541 surgeries, with 1,204 outcomes.

The models were designed to be used by clinicians to identify high-risk patients and to inform preoperative decision-making. Logistic regression models were used for the analysis. The first model (model 1) includes demographic variables such as age, and sex, as well as clinical variables such as kidney failure type (dialysis or non-dialysis), type of surgery, and the setting where the surgery is performed (inpatient or outpatient). The second model (model 2) includes all variables from the first model with the addition of comorbidities such as cancer, cerebrovascular disease, chronic pulmonary disease, dementia, diabetes, heart failure, history of myocardial infarction, hypertension, liver disease, obesity, and peripheral vascular disease. The third model (model 3) includes all variables from the second model and adds preoperative albumin and hemoglobin, as these may be available in some but not all perioperative settings.

All three models were internally validated and showed good performance, with AUC-ROC (95%CI) ranging from 0.785 (0.771, 0.800) to 0.818 (0.806, 0.831). The models were internally validated using bootstrap resampling with 1000 samples; bootstrapping is a statistical resampling technique used to estimate the sampling distribution of a statistic by repeatedly resampling with replacement from the observed data. The models demonstrated

excellent calibration slopes, indicating that the predicted probabilities closely match the observed outcomes. Decision curve analysis showed that using any of these models could result in a net benefit compared to default strategies of intervening in all or not intervening at all in the participant group.

The benefits of these models over other tools are notable. First, the models are specific to patients with kidney failure, which is an important population to consider given their increased risk for post-operative complications. Second, the models include a wide range of variables that are known to be associated with post-operative complications, including demographic, clinical, and laboratory variables. Third, the models were developed and validated using a large observational cohort study, which provides strong evidence for their accuracy and reliability.

However, the models have not been externally validated. The study has limitations because the definitions of comorbidities and outcomes may be specific but not very sensitive, possibly leading to an underestimation of risk. Second, preoperative natriuretic peptides, useful in general populations for assessing postoperative risk, are hard to interpret in patients with kidney failure.[42] Additionally, significant outcome risk variability within predictor categories were not accounted for in the model estimates that can significantly impact risk assessments for individual patient contexts.[42]

This research proposal externally validated models developed using the Alberta Kidney Disease Network (AKDN) database[42] and developed and validated a risk prediction model using machine learning methods for major cardiovascular events and mortality in patients with kidney failure within 30 days of undergoing outpatient or inpatient non-cardiac surgery.

Research objectives

1. To externally validate existing the AKDN models for perioperative major cardiovascular events or death in adults with kidney failure within 30 days of outpatient or inpatient non-cardiac surgery in the Manitoba population.
2. To develop and validate a new risk prediction model using machine learning methods that can predict the risk of major cardiovascular events or death within 30 days of outpatient or inpatient non-cardiac surgery for patients with kidney failure.

Data Sources

Manitoba Health (MH) is the publicly funded health insurance agency providing comprehensive health insurance, including coverage for hospital and outpatient physician services, to the province's 1.3 million residents. Coverage is universal, with no eligibility distinction based on age or income, and participation rates are very high (>99%).[47] Insured services include hospital, physician, and preventive services, including vaccinations. MH maintains several centralized, administrative electronic databases that are linkable using a unique personal health identification number (PHIN). The completeness and accuracy of these databases are well established.[48,49]

We used the MH Population Registry (MHPR), which captured addresses and dates of birth, death and health insurance coverage for all insured persons, to identify the source population. We used the Hospital Abstracts Database (HAD), and Medical Services Database (MSD) to obtain information on surgeries and comorbidities. Since 1971, the HAD recorded virtually all services provided by hospitals in the province, including admissions and day surgeries.[49] The data collected comprise demographic as well as diagnosis and treatment information including primary diagnosis and service or procedure codes, coded using the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) before April, 2004, and the ICD-10-CA (Canadian adaptation of the ICD-10) and the Canadian Classification of Health Interventions (CCI) afterwards. Shared Health Diagnostic Services (SHDS) data were used to determine serum creatinine, hemoglobin and serum albumin. Serum

creatinine was used to calculate eGFR using CKD-EPI Equation[50] for Chronic Kidney Disease Stage 5 Not on Dialysis (CKD G5ND) cohort.

For the external validation of our newly developed risk prediction model using Manitoba data, we derived a retrospective, population-based cohort from the Alberta Kidney Disease Network (AKDN) database, utilizing linked administrative health, laboratory, and kidney failure datasets from Alberta, Canada.[51]

Cohort Definition

The Manitoba Centre for Health Policy (MCHP) databases were used to create a cohort of adult patients (≥18 year-olds) with renal failure (both non-dialysis and dialysis), who underwent an inpatient or outpatient non-cardiac surgical procedure between April 1, 2007, and December 31, 2019. The index date for each patient in the cohort was the day of the surgery.

1. Non-Cardiac Surgery

Surgical procedures were defined based on CCI coding, developed by CIHI[52], which involves the Tenth Revision International Statistical Classification of Diseases and Related Health Problems (ICD) codes that have been adapted for Canadian use (i.e. ICD-10-CA).

These procedures were categorized into 11 different types of surgeries based on CCI codes, as per derived and internally validated models from Harrison et al[42]. The surgery types included arteriovenous fistula (AVF) surgery, head and neck, intra-abdominal, kidney transplantation, musculoskeletal, neurosurgery, peritoneal dialysis (PD) catheter surgery, skin and soft tissue, vascular, low-risk surgeries (composite of anorectal, breast, ophthalmologic, thoracic, lower urologic and gynecologic, and retroperitoneal)[42], and more than one surgery type. As per our derivation cohort, we considered surgeries that meet the codes in Appendix 1 from both ambulatory and inpatient hospitalization data sources. The CCI codes, grouped by section, group, and intervention used to categorize surgeries, are included in Appendix 1. It is important to note that only therapeutic procedures, not diagnostic ones, were included by selecting certain CCI codes. Additionally, procedures typically performed as an outpatient (such as

colonoscopy, bronchoscopy) were excluded after screening the CCI code guide. Our analysis also accounted for cases where participants underwent multiple surgeries.

2. CKD G5 definition

To define the status of CKD G5 correctly, we utilized two specific criteria: Firstly, we considered patients whose estimated glomerular filtration rate (eGFR), a key indicator of kidney function, fell below 15mL/min/1.73m^2 during an outpatient assessment. Secondly, we identified patients who were already receiving chronic kidney replacement therapy, such as hemodialysis or peritoneal dialysis, as outpatients before their scheduled surgical procedure.

CKD G5 = CKD G5D + CKD G5ND

Chronic Kidney Disease Stage 5 on Dialysis (CKD G5D) was defined by receipt of chronic kidney replacement therapy (hemodialysis, peritoneal dialysis) (Appendix 2) as an outpatient for at least 90 days prior to the admission that included the procedure. It is important to note that this cohort definition excluded patients who have undergone kidney transplantation (they are typically classified under kidney failure categories CKDG1-5T)

CKD G5ND was distinguished by analyzing the levels of outpatient serum creatinine measured before the procedure. It required a minimum of two outpatient serum creatinine measurements taken within 7 to 365 days before the surgery as per a validated algorithm[53]. The estimated glomerular filtration rate (eGFR) was calculated using the CKD Epidemiology Collaboration (CKD-EPI) equation[50], utilizing the data obtained from the outpatient creatinine measurements. Patients whose calculated average eGFR is found to be less than 15 mL/min/1.73m^2 were included in the CKD G5ND group. By considering at least two outpatient serum creatinine measurements, we minimized the risk of misclassifying individuals with temporary fluctuations in kidney function due to acute kidney injury, allowing us to focus specifically on patients with stable and significant impairment in kidney function.

Patients that out-migrate (including “other” reasons for leaving the data registry) within the 30 days after the procedure were excluded.

Outcome

We were interested in predicting the probability of death or a major CV event within 30 days of the index procedure (outcome ascertainment period: April 1, 2007 to January 30, 2020). This composite outcome was defined as acute myocardial infarction (AMI), fatal and non-fatal cardiac arrest or ventricular arrhythmia, and all-cause mortality, with each component of this composite defined in Table A.

Table A. Composite outcome for prediction model, with ICD codes and relevant validation studies

Component of Outcome	ICD-9-CM diagnostic codes	ICD-10-CA diagnostic codes	Validation or Summary Studies
Acute Myocardial Infarction	410 (hospitalization data)	I21, I22 (hospitalization data)	(54, 55)
Fatal and Non-fatal Cardiac Arrest, Ventricular Arrhythmia	Hospitalization data: 427.1, 427.4, 427.41, 427.42, 427.5, 427.9, 798, 798.1, 798.2 (hospitalization data – manually converted to ICD-10)	I47.2, I49.01, I49.02, I46.9, I49.9, R99	(56)
All-cause Mortality	From Manitoba Health Insurance Registry, cancellation of coverage due to death within 30 days of index surgical procedure.		

We narrowed our focus on CV outcomes, specifically targeting acute myocardial infarction (AMI), cardiac arrest, and ventricular arrhythmia. The RCRI and NSQIP ACS algorithms define cardiac events as cardiac arrest and AMI.

By including ventricular arrhythmia as part of the composite outcome, we acknowledged that ventricular arrhythmias can often progress to cardiac arrest without immediate intervention, similar to “non-fatal” cardiac arrest. Validation studies reinforced our choice to incorporate ventricular arrhythmia into the composite outcome. This composite is consistent

with other postoperative risk tools and was guided by guidelines for perioperative cardiac risk assessment.[37] Moreover, incorporating all-cause mortality as part of the composite helped minimize concerns related to competing risks. A competing risk is an event whose occurrence precludes the occurrence of the primary event of interest. Since patients with CKD G5 have a significantly high risk of mortality, their eligibility for the AMI outcome could be compromised if they were to pass away within the 30-day post-operative period. Including all-cause mortality in the composite helped address this issue and allowed us to account for the potential impact of competing risks. Composite endpoints were recommended in settings where the risk of competing outcomes, such as death, is frequent.[57,58]

Predictors

We identified demographics characteristics such as sex and birth date from the MHPR. Surgeries were categorized into 11 surgery types based on CCI codes. The surgery setting was classified using the HAD as an outpatient, inpatient elective, or inpatient emergency. Chronic conditions and other comorbidities including AMI, cancer, chronic pulmonary disease, dementia, diabetes, heart failure, hypertension, liver disease, peripheral vascular disease, and stroke were identified using the HAD and MSD based on validated algorithms of ICD-9-CM and ICD-10-CA codes[55] with an unrestricted lookback period for permanent conditions and 5 years for temporary conditions (Appendix 2). Kidney failure treatment was classified as non-dialysis, hemodialysis, or peritoneal dialysis. Most recent preoperative outpatient serum albumin (in g/L) and serum hemoglobin (in g/L) within the year before surgery were also included.

Exclusions

1. Patients that out-migrate within 30 days after the procedure.
2. Missing serum albumin or hemoglobin: Approximately 2.6% of participants were excluded to preserve the integrity and accuracy of the analysis, given the clinical significance of these biomarkers.
3. Unknown surgery settings.

Analysis

Objective 1: External validation of AKDN models

Descriptive analyses were conducted to understand the characteristics of the cohort. This included the calculation of medians and interquartile ranges (IQRs) for continuous variables and counts and percentages for categorical variables. Age was centred at 18 years. We assessed the performance of all the three risk prediction models developed using the Alberta Kidney Disease Network (AKDN) database on Manitoba data as follows:

1. Model deployment: It involved utilizing a predictive model's parameters developed using AKDN database and applied them to Manitoba data. The coefficients from AKDN models were extracted and then used to make predictions on Manitoba data.
2. Model refitting: It involved re-estimating the coefficients of the models using Manitoba data while we maintained the same variable structure as used in the AKDN models.

We used the following statistical methods and metrics to evaluate the performance of the models:

1. Area Under the Receiver Operating Characteristic Curve (AUC-ROC) or Concordance Statistic (AUC-ROC): A comprehensive measure that evaluates the model's ability to distinguish between classes across various thresholds, with a higher AUC indicating better performance.
2. Area Under the Precision-Recall Curve (AUC-PR): Measures the trade-off between precision and recall for different probability thresholds. It is more informative of model discrimination with infrequent outcomes.[59]
3. Calibration Plot: Visualizes the relationship between predicted probabilities and actual outcomes, assessing the reliability of probabilistic predictions.
4. Calibration Slope: The calibration slope assesses the relationship between predicted and observed probabilities. A slope of 1 indicates perfect calibration[60,61], while a slope greater than 1 and less than 1 indicates underprediction and overprediction, respectively.

5. Brier Score: It is the mean squared difference between observed binary outcomes and predicted probabilities. A lower Brier score indicates better model performance.[19]
6. Sensitivity (True Positive Rate): The proportion of true positives correctly identified by the model.
7. Specificity (True Negative Rate): The proportion of true negatives correctly identified by the model.
8. Positive Predictive Value (PPV): The proportion of positive results that are true positives.
9. Negative Predictive Value (NPV): The proportion of negative results that are true negatives.

   We examined the sensitivity, specificity, positive and negative predictive values of the models at 5%, 10%, and 15% predictive thresholds, which were identified as critical to both patients and care providers in perioperative literature.[62,63]
10. Decision curve analysis: This method estimates the net benefit of prediction methods across a range of risk threshold probabilities for an intervention.[64,65] Net benefit is calculated as the weighted balance of true and false positives and is compared to a strategy where there is no "intervention," which is context specific. This is where threshold probabilities become important as they are patient and care provider specific, and defined as the risk threshold that warrants intervention.[37]
11. Net Reclassification Improvement (NRI): This evaluate the improvement in reclassification of risk when additional predictors are included in the model. Clinically significant categories of < 5%, 5-15%, and >15%, were utilized. These thresholds were recommended in perioperative prognostic literature as pertinent to both patients and healthcare providers.[62,63]
12. Net Absolute Reclassification Index (NARI): This quantifies the improvement in classification accuracy of a predictive model. We computed NARI using same risk categories as NRI, determined as the net count of correctly reclassified individuals per 1,000 patients.

Objective 2: Develop and validate a machine learning model

This objective involved developing a machine learning model on Manitoba data and externally validating it on Alberta data. Descriptive analyses were conducted to understand the demographics and other characteristics of the cohort. We used Pearson's correlation to measure the strength of association between continuous variables and Chi-Square Test of Independence to determine if there was a significant relationship between categorical variables. We used machine learning to develop the model. Once we developed the final model, we validated the model externally on Alberta data. The dataset used for external validation was same as used by Dr. Harrison to develop AKDN models.[42]

Different machine learning models can be used for risk prediction:

1. Decision trees: A decision tree is a flowchart-like tree structure where an internal node represents a feature (or attribute), a branch represents a decision rule, and each leaf node represents the outcome. The problems with decision trees are that they are prone to overfitting (which tends to capture the random noise in the data rather than the underlying relationship), and small changes in the data can lead to completely different trees and bias to the dominant class.[66] We did not use this algorithm for risk prediction.
2. Neural Networks: They are a set of algorithms modeled loosely after the human brain, designed to recognize patterns in data. They generally work well on large amounts of data to train effectively and not easily interpretable. We did not use this algorithm for risk prediction.
3. Random forest[67,68]: Random forest is an ensemble learning method used for classification. It operates by constructing multiple decision trees during training and outputs the mode of the classes (for classification) of the individual trees. It handles overfitting and is more robust than decision trees. They also work well on small datasets and are less complex than neural networks, thus making them more interpretable.[69] We used this algorithm for risk prediction.
4. XGBoost (eXtreme Gradient Boosting): XGBoost uses gradient boosting algorithms, which create new models that predict the residuals or errors of prior models and then add them together to make the final prediction.[70] It is widely used to solve real-world

problems. XGBoost includes a regularized model to prevent overfitting. We used this algorithm for risk prediction. We used the XGBoost library for the machine learning analysis in this study.[70]

We used random forest and XGBoost in our analysis and finalized the model based on the performance as described in the following sections.

We prepared an analytical dataset that included all the demographic and clinical characteristics of the patient. We used the “get_dummies” function to transform categorical variables with multiple categories into binary variables, making them suitable for use in machine learning models. The dataset was unbalanced, therefore, we used stratified sampling on outcome variable to split the dataset into training, validation, and testing sets. The data split ratios were 70% for training, 15% for validation, and 15% for testing. The training set was used to tune the hyperparameters and train the models, the validation dataset was used for feature selection and evaluating model performance, while the testing set was used to evaluate the model’s final performance.

We used the following metrics to evaluate the performance of the models:

1. Area Under the Receiver Operating Characteristic Curve (AUC-ROC): A comprehensive measure that evaluates the model's ability to distinguish between classes across various thresholds, with a higher AUC indicating better performance.
2. Area Under the Precision-Recall Curve (AUC-PR): Measures the trade-off between precision and recall for different probability thresholds. It is more informative of model discrimination with infrequent outcomes.[59]
3. Calibration Plot: Visualizes the relationship between predicted probabilities and actual outcomes, assessing the reliability of probabilistic predictions.
4. Brier Score: It measures the accuracy of probabilistic predictions. A lower Brier score indicates better model performance.

5. Sensitivity (True Positive Rate): The proportion of actual positives correctly identified by the model.
6. Specificity (True Negative Rate): The proportion of actual negatives correctly identified by the model.
7. Positive Predictive Value (PPV): The proportion of positive results that are true positives.
8. Negative Predictive Value (NPV): The proportion of negative results that are true negatives.

We examined the sensitivity, specificity, positive and negative predictive values for the models at different predictive thresholds, which were identified as critical to both patients and care providers in perioperative literature.[62,63] We utilized 'confidenceinterval' library to calculate sensitivity, specificity, PPV, NPV, and their confidence intervals using the Wilson method. [71] The ROC AUC score and its confidence interval were determined using the DeLong method, while the Brier score confidence interval was calculated through bootstrapping with 5000 iterations.[71] Additionally, the AUC-PR confidence interval was obtained using bootstrapping with 5000 iterations.

Once we had an analytical dataset, we used the following steps to create a machine learning model:

1. Tuning the hyperparameters: A hyperparameter is a parameter whose values help guide the learning process. The grid search method (GridSearchCV) was employed to fine-tune hyperparameters such as the number of trees, depth of trees, and learning rate. GridSearchCV systematically explores combinations of specified hyperparameters using cross-validation, ensuring optimal model tuning and thereby improving the predictive performance of the model.[72] Predicting the best hyperparameter values manually would be time-consuming and resource-intensive, which was why GridSearchCV was employed for hyperparameter tuning. In GridSearchCV, we used a 5-fold cross-validation (cv=5) approach. This meant the data was divided into five parts, with the model being trained on four parts and validated on the fifth. This cycle was repeated five times, each time with a different part used as the validation set. We used all the available features to fine-

tune the hyperparameters. We made hyperparameters selection based on the highest AUC-ROC score from GridSearchCV. We used AUC-ROC because it provided a comprehensive measure of a model's performance across all classification thresholds.

a. To find best hyperparameters for random forest model, we used the following range of values:

Parameter	Range	Best value[a]
Number of trees (n_estimators)	300 to 550 in increments of 50	450
Tree depth (max_depth)	10 to 50 in increments of 10	10
Samples required to split a node (min_samples_split)	2 to 6 in increments of 1	2
Minimum number of samples required at a leaf node (min_samples_leaf)	3 to 19 in increments of 1	15
Class weight (class_weight)	balanced, balanced_subsample, and None	None
Samples bootstrapping	True	True

[a] Based on highest AUC-ROC score from GridSearchCV

b. To find best hyperparameters for XGBoost model, we used the following range of values:

Parameter	Range	Best value[a]
Learning rate (learning_rate)	0 to 0.45 in increments of 0.05	0.35000000000000003
Number of trees (n_estimators)	50 to 450 in increments of 50	50

Minimum sum of instance weight (hessian) needed in a child (min_child_weight)	0 to 5 in increments of 1	2
Tree depth (max_depth)	1 to 9 in increments of 1	3
L1 regularization term on weights (reg_alpha)	1 to 13 in increments of 2	11
Minimum loss reduction (gamma)	0 to 0.4 in increments of 0.1	0
Subsample ratio (subsample)	0 to 0.9 in increments of 0.1	0.8
Subsample ratio of columns for each split (colsample_bytree)	0 to 0.9 in increments of 0.1	0.6000000000000001
Objective	binary:logistic	binary:logistic

[a] Based on highest AUC-ROC score from GridSearchCV

2. Feature selection: After we found the best hyperparameters, feature selection played a crucial role in selecting the best model. We implemented the following methods on the training set to evaluate the importance of features using the best hyperparameters:
   a. Variable Importance (VIMP): This method involved measuring the importance of each feature in a predictive model. Both random forest and XGBoost have built-in mechanisms for selecting important features within their respective algorithms. Random forest uses the impurity-based feature importances, known as Mean Decrease in Impurity (MDI) or Gini importance. It is computed as the (normalized) total reduction of the criterion brought by that feature. The higher, the more important the feature.[73] XGBoost used the average gain across all splits the feature is used in to get impotance of each feature.
   b. Lasso Regression (L1 Regularization): It was used for feature selection as it tends to shrink the coefficients of less important features to zero.

c. We compiled the feature rankings derived from the above methods (VIMP including random forest and XGBoost, and Lasso Regression), computed the average rank for each variable, and then re-ranked them based on these averages. This process helped in pinpointing the most influential features across different models.

For each of the above method (a, b, and c), features were ranked by their importance. Starting with the most important feature, we iteratively added one feature at a time, according to the importance ranking, and retrained the model on the training set. After each iteration, the model was evaluated on a validation set, and AUC-ROC and AUC-PR were calculated. We also included or excluded certain features to the model, retrained the model and evaluated on a validation set. In refining our model selection process, we encountered instances where a model displayed a high AUC-ROC score but a lower AUC-PR value, or vice-versa. So, we selected models that presented reasonable and balanced values for both metrics. This helped us in selecting models that were consistently reliable across varied scenarios.

3. Once we had the final model, we evaluated its performance on the testing set. We also reported performance at different thresholds, and so we were able to control the trade-off between various evaluation metrics such as PPV and sensitivity. After we were satisfied with the model, we performed cross-validation on the full dataset. Cross-validation is a sampling-based method used to assess the effectiveness of the model. It involves partitioning data into two or more subsets, systematically performing analysis on training set while validating the analysis on the testing set. The purpose of cross-validation was to highlight problems like overfitting and to give an insight on the generalization of the model. Specifically, we used 10-fold cross-validation, in which the dataset was divided into 10 distinct folds and the cross-validation process was conducted in 10 iterations. In each iteration, 9 folds were used for training the model, and the remaining fold was used for testing the model. We computed the mean and standard deviation (SD) of two metrics to evaluate the performance of our model: Area Under the Receiver Operating Characteristic Curve (AUC-ROC) and the Area Under the

Precision-Recall Curve (AUC-PR). The mean indicated the average performance of the model across different subsets of the data, thereby offering insights into the model's overall effectiveness. The standard deviation measured the variation of the metric values from the mean. A lower standard deviation indicates that the metric values across folds are close to the mean, suggesting that the model's performance is consistent across different data subsets.

In our study, we utilized two models. The first model, designated as the baseline model, included all available features. This comprehensive model served as a reference point. The second model was developed with a focus on optimizing performance. The aim was to construct a model that not only performed better in terms of performance metrics but also offered greater simplicity and interpretability. We finalized the second model based on its performance in validation set. We then evaluated the chosen model on testing dataset that the models have not seen during training or validation. The final models were saved to disk using "joblib". "joblib" saved the entire state of a model, including all the model features, hyperparameters, and model structure.

External validation:

The saved models were shared with our Alberta counterparts to externally validate. The dataset used for external validation was the same dataset used by Dr. Harrison to develop AKDN models.[42] Prior to validation, the Alberta dataset underwent several preprocessing steps to align it with the format of the Manitoba data like encoding of categorical variables and aligning variable names. Model performance was evaluated using the same metrics as described for the internal validation.

Sensitivity analysis

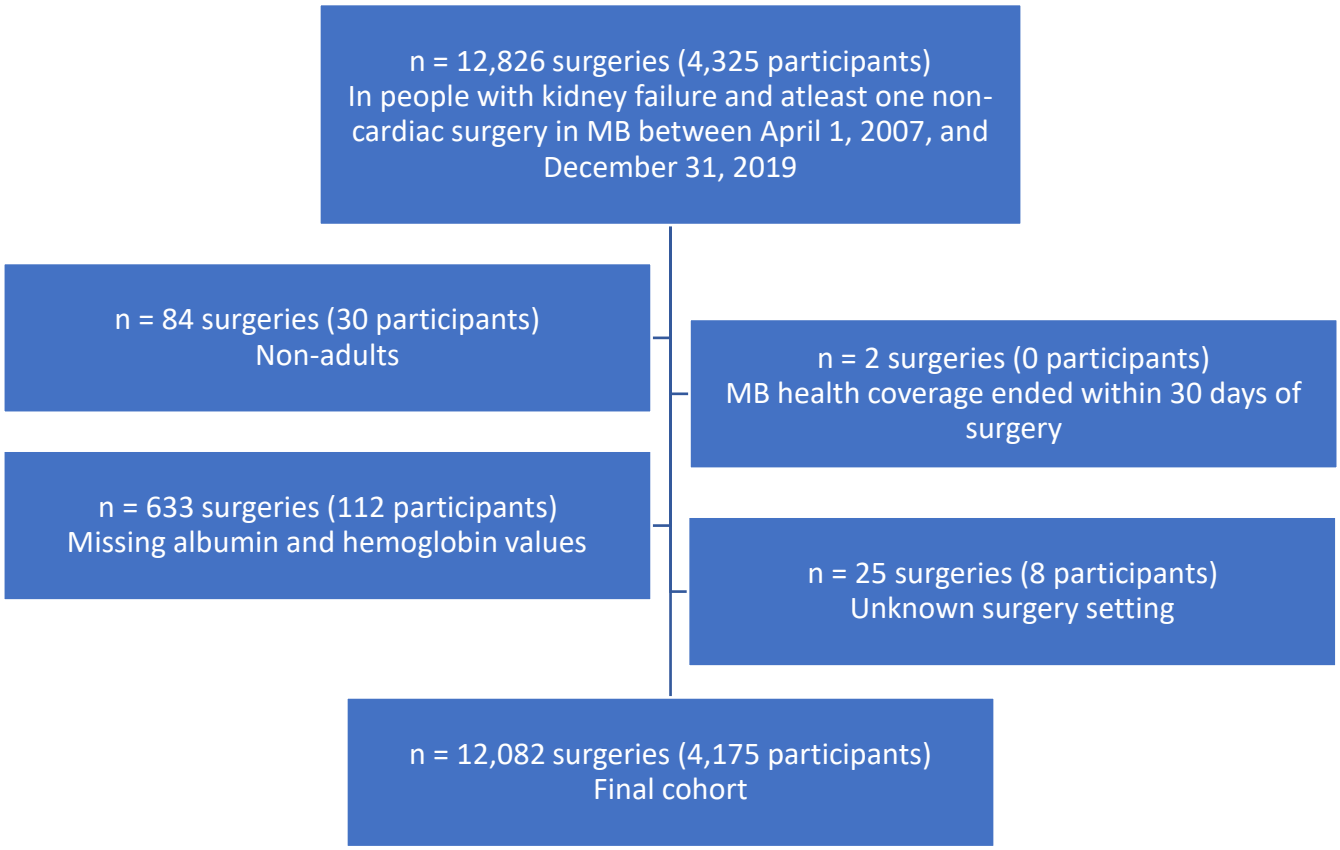
We restricted the MB cohort to include only the first surgery for each patient (first surgery cohort). The data was split into a training set and a test set in a 70:30 ratio. We applied the

same features and hyperparameters as used in the original ML models. The models were trained on the training set and their performance was evaluated on the test set.

An additional sensitivity analysis was performed to evaluate the robustness of our original ML models predicting the risk of major cardiovascular events and death in patients with kidney failure undergoing non-cardiac surgery. For this analysis, we used first surgery cohort as defined above. We applied the pre-trained original ML models directly to the entire cohort and assessed their performance.

Statistical analyses were conducted using SAS (9.4), and Python (v. 3.9.13). The analysis was performed using Scikit-learn, a Python-based library.[75]

Figure 2: Cohort flow diagram. The number of patients that were included and excluded at each step of cohort formation



Descriptives

Table 1 Frequency (%) of surgeries by outcome and selected demographics and clinical characteristics

Characteristic	Mortality or MACE		
	No	Yes	Total
Total Surgeries	**11,513 (100.0)**	**569 (100.0)**	**12,082 (100.0)**
Sex, No. (%)			
Male	6,529 (56.7)	334 (58.7)	6,863 (56.8)
Female	4,984 (43.3)	235 (41.3)	5,219 (43.2)
Age, Median (Q1-Q3)	60.6 (50.3-69.5)	65.0 (56.2-72.9)	60.9 (50.5-69.7)
Surgery type, No. (%)			
Intra-abdominal	848 (7.4)	47 (8.3)	895 (7.4)
Musculoskeletal	1,216 (10.6)	124 (21.8)	1,340 (11.1)
Head and neck	527 (4.6)	17 (3.0)	544 (4.5)
Vascular	1,412 (12.3)	151 (26.5)	1,563 (12.9)
Skin and soft tissue	S	S	255 (2.1)
Neurosurgery	S	<6 (<1.0)	84 (0.7)
Peritoneal dialysis catheter	1,479 (12.8)	32 (5.6)	1,511 (12.5)
Arteriovenous Fistula	2,688 (23.3)	39 (6.9)	2,727 (22.6)
Kidney Transplant	388 (3.4)	12 (2.1)	400 (3.3)
Low-risk[a]	1,810 (15.7)	34 (6.0)	1,844 (15.3)
More than one	827 (7.2)	92 (16.2)	919 (7.6)
Surgery setting, No. (%)			
Outpatient	7,231 (62.8)	89 (15.6)	7,320 (60.6)
Elective inpatient	2,153 (18.7)	84 (14.8)	2,237 (18.5)
Emergency inpatient	2,129 (18.5)	396 (69.6)	2,525 (20.9)
Kidney failure type, No. (%)			

Non-dialysis	3,195 (27.8)	137 (24.1)	3,332 (27.6)
Hemodialysis	6,591 (57.2)	312 (54.8)	6,903 (57.1)
Peritoneal dialysis	1,727 (15.0)	120 (21.1)	1,847 (15.3)
Cerebrovascular disease, No. (%)	3,453 (30.0)	246 (43.2)	3,699 (30.6)
Chronic pulmonary disease, No. (%)	8,390 (72.9)	445 (78.2)	8,835 (73.1)
Dementia, No. (%)	443 (3.8)	48 (8.4)	491 (4.1)
Diabetes, No. (%)	8,437 (73.3)	470 (82.6)	8,907 (73.7)
Heart failure, No. (%)	5,322 (46.2)	382 (67.1)	5,704 (47.2)
History of myocardial infarction, No. (%)	2,768 (24.0)	358 (62.9)	3,126 (25.9)
Hypertension disease, No. (%)	10,482 (91.0)	540 (94.9)	11,022 (91.2)
Liver disease (mild and moderate/severe), No. (%)	1,549 (13.5)	99 (17.4)	1,648 (13.6)
Peripheral vascular disease, No. (%)	3,141 (27.3)	273 (48.0)	3,414 (28.3)
Cancer, No. (%)	1,665 (14.5)	99 (17.4)	1,764 (14.6)
Albumin, Median (Q1-Q3)	33.0 (29.0-36.0)	29.0 (23.0-33.3)	32.9 (29.0-36.0)
Hemoglobin, Median (Q1-Q3)	105.0 (95.0-115.0)	103.0 (92.0-113.0)	105.0 (94.0-115.0)
Mortality, No. (%)	0 (0.0)	302 (53.1)	302 (2.5)

[a] includes anorectal, breast, ophthalmologic, thoracic, lower urologic and gynecologic, and retroperitoneal surgeries
S = Suppressed, because number of events involved was between one and five.

We identified 12,082 surgeries performed in 4,175 participants (Figure 2) and 569 (5%) of those led to death or a major cardiac event (outcome) within 30 days of the surgery (Table 1). Among these 569 outcome events, the deaths constituted 53% of the events. About 57% of the total surgeries were performed on males. The median age among the group with outcome was 65.0 years (IQR: 56.2-72.9) and 60.6 years (IQR: 50.3-69.5) among group with no outcome. Arteriovenous fistula (23%), low-risk (15%) and vascular (13%) were the most common surgery types. Notably, musculoskeletal surgeries accounted for about 22% in the group having the outcome compared to 11% in the group with no outcome. Vascular surgeries were also more

prevalent in the group with the outcome (27%) than in the no outcome group (12%). About 61% of surgeries were outpatient surgeries. About 63% of surgeries in the group with no outcome were outpatient, while 70% of surgeries in the group with outcome were emergency inpatient. Approximately 57% of the surgeries were conducted on individuals undergoing hemodialysis. The most prevalent comorbid conditions were hypertension (91%), followed closely by diabetes (74%), and chronic pulmonary disease (73%). There was a notable difference in the history of myocardial infarction between the two groups. While 24% of the surgeries in the no outcome group had a history of myocardial infarction, this rate was significantly higher (63%) in the group with outcome. The preoperative median hemoglobin level was recorded at 105 g/L, with an interquartile range (IQR) of 94 to 115 g/L, and the median albumin level was 33 g/L, with an IQR spanning from 29 to 36 g/L.

Objective 1: External validation of AKDN models (Model deployment)

Table 2 Overall model performance for each risk prediction model

	Model 1	Model 2	Model 3
Outcome/ Surgeries (N)	569/12,082	569/12,082	569/12,082
AUC-ROC (95% CI)	0.821 (0.803-0.838)	0.868 (0.853-0.883)	0.874 (0.859-0.888)
AUC-PR	0.179	0.411	0.393
Calibration slope	1.32	1.4	1.24
Brier score	0.041	0.038	0.037

Model 1 included age, sex, dialysis modality, surgery type and setting.
Model 2 included model 1 and comorbidities
Model 3 included model 2 and preoperative hemoglobin and albumin

We assigned coefficients from the AKDN models to MB data and calculated AUC-ROC, AUC-PR, calibration slope, and Brier score (Table 2). AUC-ROC ranged from 0.821 (95% CI 0.803-0.838) for model 1 to 0.874 (95% CI 0.859-0.883) for model 3, indicating good discriminative ability. Model 1 had an AUC-PR of 0.179, which was substantially lower compared to the other models, indicating less precision and recall balance. Model 2 and model 3 had AUC-PR of 0.411 and 0.393 respectively, reflecting a significantly better performance in capturing the positive

class. Models 1, 2, and 3 had calibration slopes of 1.32, 1.4, and 1.24 respectively, indicating a tendency to under-predict outcomes which means model's predicted probabilities are systematically lower than the actual observed outcomes. The low Brier scores in all the three models indicated high level of accuracy in the model's probabilistic predictions.

Figure 3: Calibration plot for model 1 (deployment)

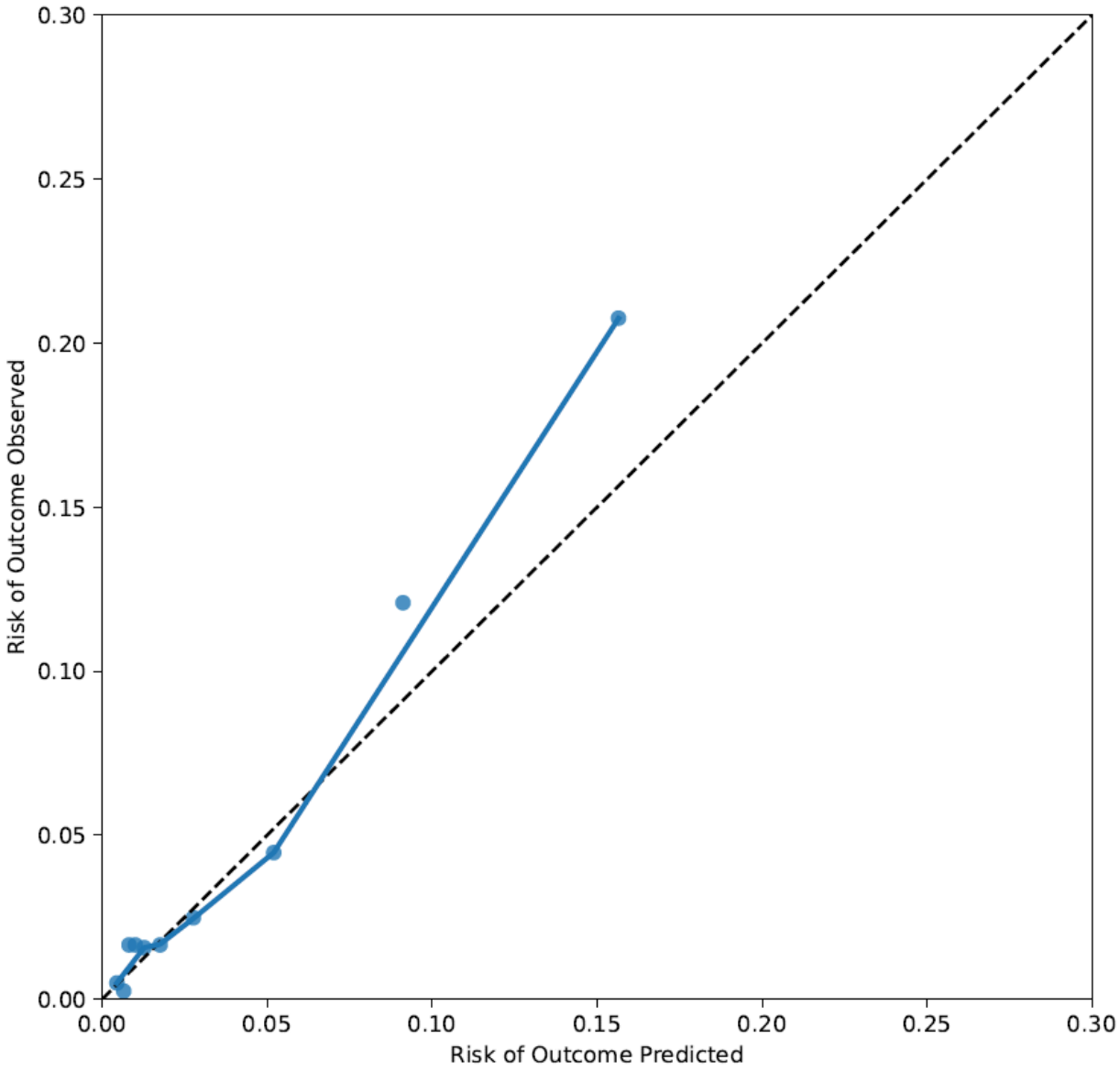


Figure 4: Calibration plot for model 2 (deployment)

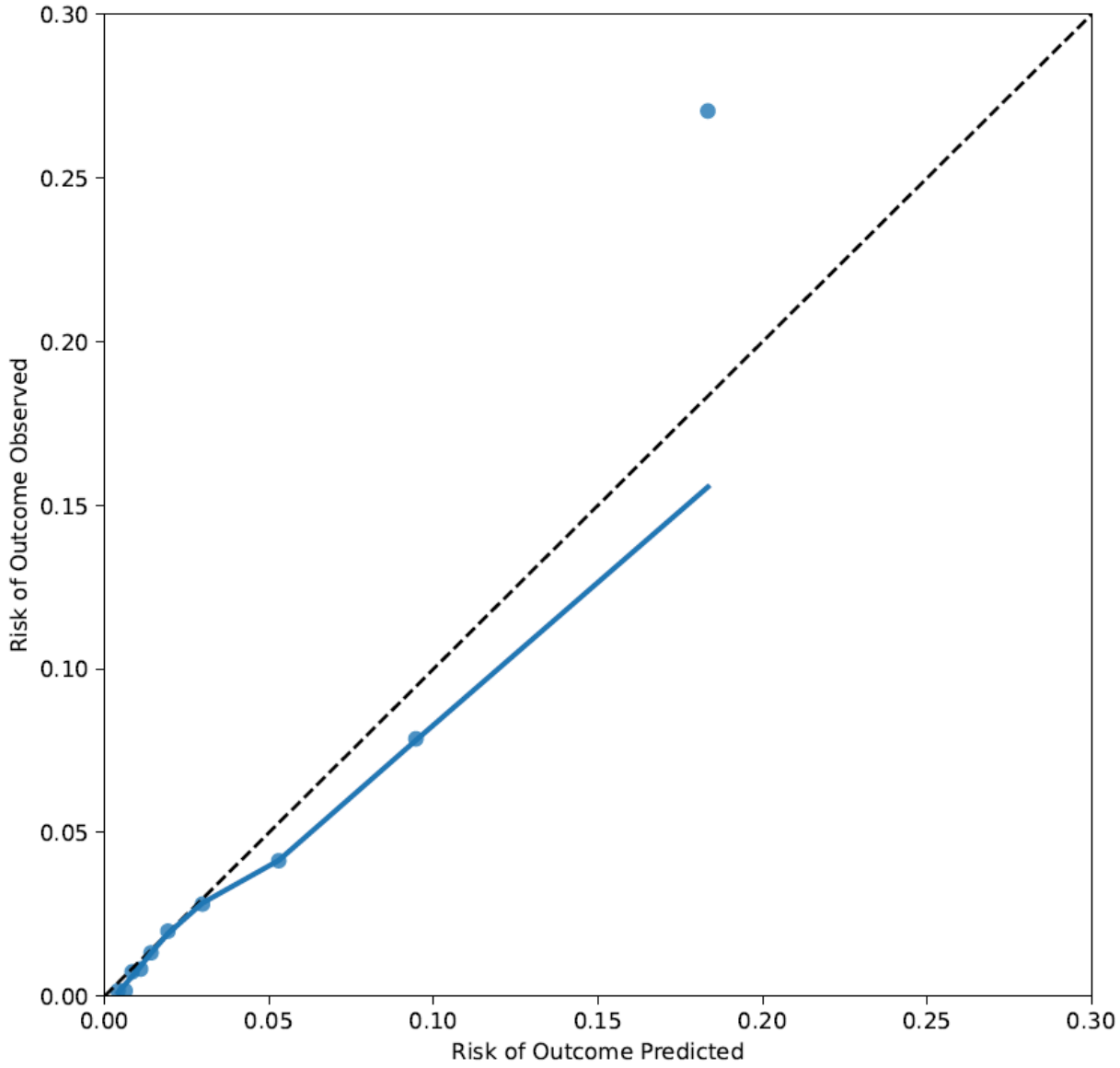


Figure 5: Calibration plot for model 3 (deployment)

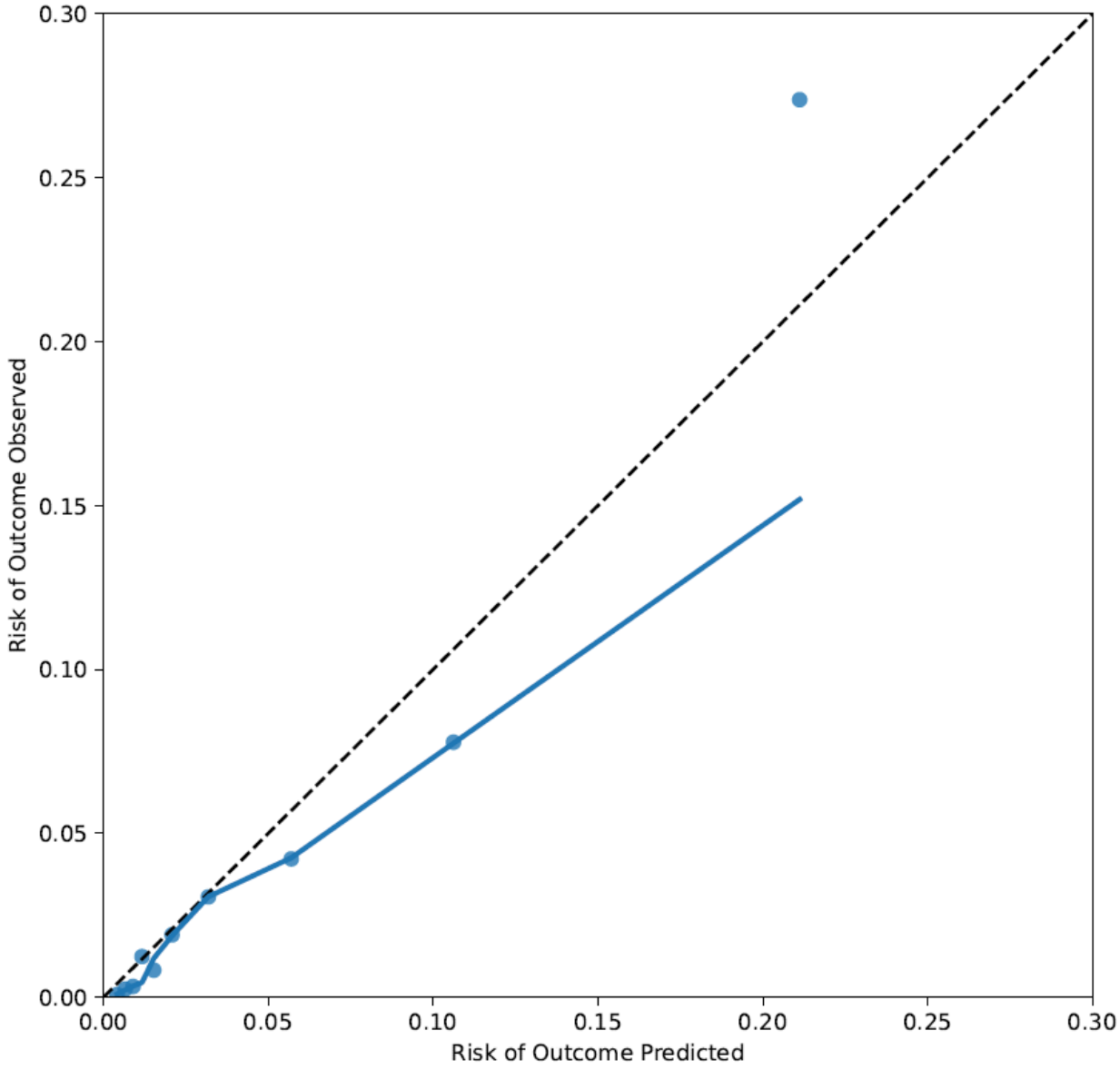
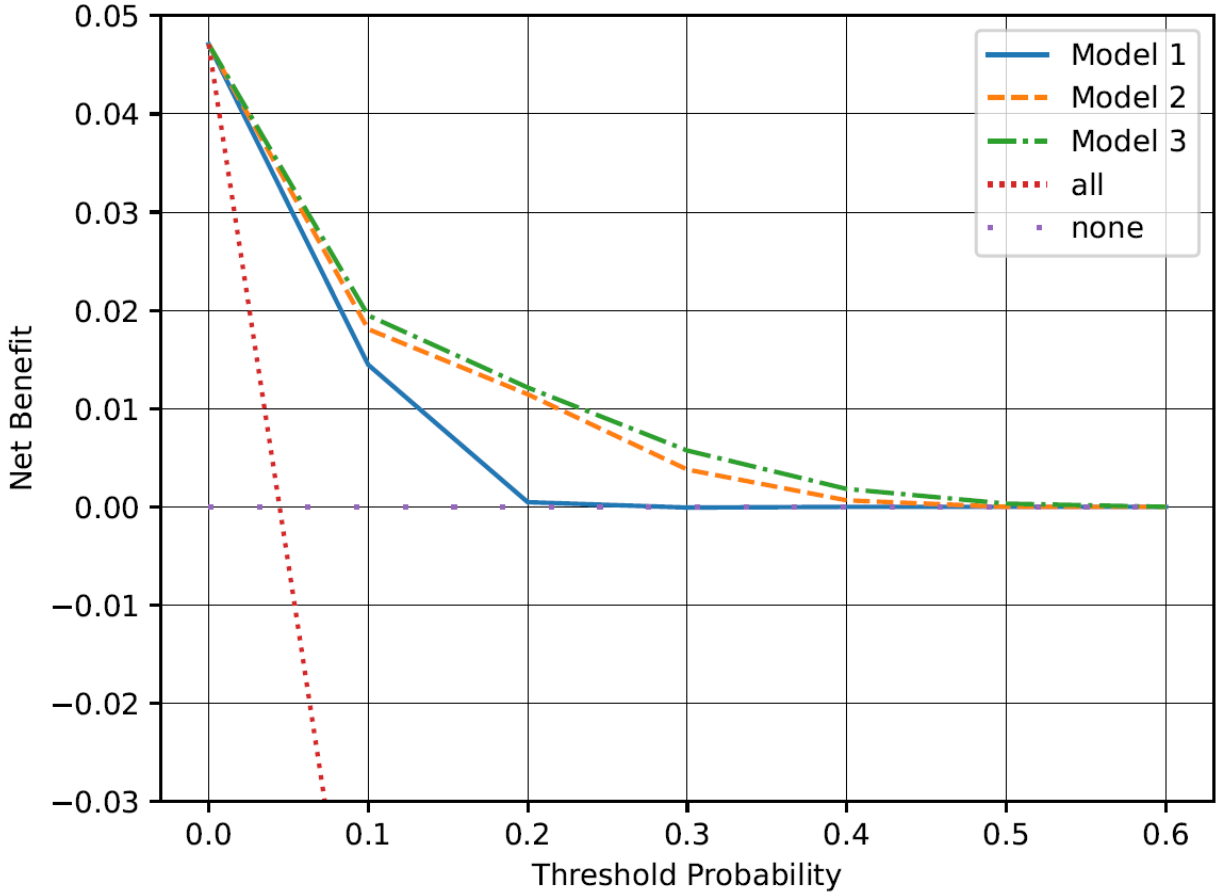


Figure 6: Decision curve analysis



The calibration plots for all the models demonstrated strong calibration throughout, except for overestimation at the highest predicted risks, as shown in Figures 3 through 5. The clinical usefulness of the three models was assessed through Decision Curve Analysis (DCA). Each model exhibited comparable positive net benefit relative to approaches where all or none of the participants received a perioperative cardiac monitoring intervention (Figure 6). The models were advantageous, yielding net benefit at predicted risk thresholds ranging from 0 to 40%.

Table 3 Sensitivity, specificity, NPV and PPV (%) at different thresholds of predicted probability of major postoperative outcomes

	Sensitivity	Specificity	PPV	NPV
At 5%				
Model 1	76.3	76.9	14.0	98.5
Model 2	78.9	76.8	14.4	98.7
Model 3	81.2	76.0	14.3	98.8
At 10%				
Model 1	55.5	89.0	19.9	97.6
Model 2	63.8	88.7	21.8	98.0
Model 3	70.5	87.1	21.2	98.4
At 15%				
Model 1	22.5	95.8	20.8	96.2
Model 2	48.0	95.4	33.8	97.4
Model 3	55.4	93.5	29.6	97.7

Model 1 included age, sex, dialysis modality, surgery type and setting.
Model 2 included model 1 and comorbidities
Model 3 included model 2 and preoperative hemoglobin and albumin

Table 3 shows sensitivity, specificity, NPV and PPV (%) at different thresholds of predicted probability of major postoperative outcomes. As the threshold increases from 5% to 15%, there is a general trend of decreasing sensitivity and increasing specificity across all models. Each model exhibits a trade-off between sensitivity and specificity at different thresholds, with model 3 consistently showing higher sensitivity and model 1 maintaining higher specificity at higher thresholds. The PPV tends to increase with higher thresholds, indicating a higher percentage of true positives among positive predictions. Conversely, the NPV slightly decreased, reflecting the proportion of true negatives among negative predictions.

Table 4 Event and non-event Reclassification Tables between models, stratified by clinically important probability categories

Model 2 versus model 1

		Risk model 2 predicted probability threshold						
Outcome status	**Risk model 1** predicted probability threshold	<5%	5-15%	> 15%	Total	Correct percentage reclassification	Incorrect percentage reclassification	NRI
Present	< 5%	109	26	0	135	32.9%	4.7%	28.1%
	5-15%	11	134	161	306			
	> 15%		16	112	128			
	Total	120	176	273	569			
Absent	< 5%	8573	280		8853	3.9%	4.4%	-0.5%
	5-15%	272	1674	227	2173			
	> 15%		179	308	487			
	Total	8845	2133	535	11513			
Category-based Net Absolute Reclassification Index (NARI)	NARI = 8.6 per 1000 patients							

Model 3 versus model 2

		Risk model 3 predicted probability threshold						
Outcome status	**Risk model 2** predicted probability threshold	<5%	5-15%	> 15%	Total	Correct percentage reclassification	Incorrect percentage reclassification	NRI
Present	< 5%	105	15	0	120	12.8%	3.2%	9.7%
	5-15%	2	116	58	176			
	> 15%		16	257	273			
	Total	107	147	315	569			
Absent	< 5%	8614	231		8845	1.9%	4.6%	-2.7%
	5-15%	135	1700	298	2133			
	> 15%		83	452	535			
	Total	8749	2014	750	11513			
Category-based Net Absolute Reclassification Index (NARI)	NARI = -21.2 per 1000 patients							

Table 4 demonstrated event and non-event reclassification tables between models and calculated NRI and NARI. In comparing model 2 against model 1, for events (presence of

outcome), a NRI of 28.1% was achieved. For non-events (absence of outcome), there was a slight negative NRI of -0.5%. The Net Absolute Reclassification Index (NARI) showed an improvement of 8.6 per 1,000 patients. In the comparison between model 3 and model 2, for the presence of the outcome, model 3 achieved a NRI of 9.7%. For absence of outcome, the NRI was negative (-2.7%). The Net Absolute Reclassification Index (NARI) was -21.2 per 1000 patients, indicating a decrease in the predictive accuracy of a model 3 compared to model 2.

Objective 1: External validation of AKDN models (Model refit)

Table 5 Odds ratio (95% CI) and overall model performance for each risk prediction model

	Model 1	**Model 2**	**Model 3**
Sex			
Male	Ref.	Ref.	Ref.
Female	0.90 (0.75-1.08)	0.95 (0.79-1.14)	0.90 (0.75-1.09)
Age	1.03 (1.02-1.03)	1.02 (1.01-1.03)	1.02 (1.01-1.03)
Surgery description			
Intra-abdominal	Ref.	Ref.	Ref.
Head and Neck	1.03 (0.56-1.89)	1.16 (0.63-2.15)	1.22 (0.66-2.26)
Vascular	3.22 (2.24-4.61)	2.57 (1.77-3.72)	2.67 (1.84-3.89)
Skin and Soft Tissue	1.07 (0.61-1.90)	1.01 (0.56-1.81)	0.88 (0.49-1.59)
Neurosurgery	1.05 (0.24-4.59)	0.97 (0.22-4.30)	0.98 (0.22-4.34)
Peritoneal Dialysis Catheter	0.72 (0.44-1.18)	0.74 (0.45-1.22)	0.69 (0.42-1.15)
Arteriovenous fistula	1.02 (0.64-1.63)	1.03 (0.64-1.66)	1.01 (0.63-1.63)
Kidney Transplant	0.99 (0.50-1.95)	1.25 (0.63-2.51)	1.50 (0.74-3.04)
Low-risk	1.14 (0.71-1.84)	1.09 (0.67-1.78)	1.09 (0.67-1.78)
Musculoskeletal	1.06 (0.74-1.52)	0.91 (0.63-1.32)	0.86 (0.59-1.24)
More than one	1.79 (1.22-2.62)	1.65 (1.11-2.45)	1.56 (1.05-2.32)
Surgery setting			
Outpatient	Ref.	Ref.	Ref.

Elective inpatient	3.23 (2.30-4.54)	3.17 (2.25-4.48)	3.17 (2.25-4.49)
Emergency inpatient	14.33 (10.91-18.81)	12.15 (9.18-16.08)	11.06 (8.33-14.69)
Kidney failure type			
Non-dialysis	Ref.	Ref.	Ref.
Hemodialysis	0.88 (0.71-1.10)	0.82 (0.65-1.04)	0.83 (0.66-1.05)
Peritoneal dialysis	1.25 (0.94-1.65)	1.32 (0.99-1.77)	1.02 (0.76-1.38)
Cerebrovascular disease		1.15 (0.95-1.40)	1.17 (0.96-1.41)
Chronic pulmonary disease		1.11 (0.89-1.39)	1.14 (0.91-1.42)
Dementia		1.36 (0.96-1.94)	1.31 (0.92-1.86)
Diabetes		1.14 (0.89-1.47)	1.08 (0.84-1.39)
Heart failure		1.14 (0.92-1.40)	1.10 (0.90-1.36)
History of myocardial infarction		3.31 (2.71-4.05)	3.30 (2.70-4.04)
Hypertension disease		0.83 (0.55-1.26)	0.90 (0.59-1.38)
Liver disease (mild and moderate/severe)		1.16 (0.91-1.48)	1.12 (0.88-1.43)
Peripheral vascular disease		1.07 (0.88-1.30)	0.98 (0.80-1.19)
Cancer		1.10 (0.86-1.40)	1.11 (0.87-1.43)
Albumin			0.95 (0.93-0.96)
Hemoglobin			1.01 (1.00-1.01)
Model Fit and Performance			
Outcome/ Surgeries (N)	569/12,082	569/12,082	569/12,082
AIC, BIC	3726, 3852	3552, 3752	3513, 3727
AUC-ROC	0.830 (0.812-0.847)	0.853 (0.837-0.870)	0.861 (0.845-0.876)
Area under precision recall curve	0.226	0.297	0.290
Calibration slope	1.01	0.98	1.01

Brier score	0.040	0.038	0.038

Model 1 included age, sex, dialysis modality, surgery type and setting.
Model 2 included model 1 and comorbidities
Model 3 included model 2 and preoperative hemoglobin and albumin

We developed a prediction model using logistic regression incorporating the same predictors as those used in the AKDN models, and allowed coefficients and intercept to vary. Table 5 shows odds ratios, along with their respective 95% confidence interval (CI) for all the three models. Models 2 and 3 show lower AIC and BIC values suggesting an improved model fit compared to model 1. AUC-ROC ranged from 0.83 (95% CI 0.812-0.847) for model 1 to 0.861 (95% CI 0.845-0.876) for model 3, indicating good discrimination capability. AUC-PR increased from 0.226 for model 1 to 0.297 and 0.29 for models 2 and 3 respectively. All the models demonstrated excellent calibration at 1. Furthermore, the Brier scores were consistently low across all three models, reflecting a higher precision in the probability estimates. The calibration plots for all models displayed excellent calibration across all predicted risks (Figures 7-9).

Figure 7: Calibration plot for model 1 (refit)

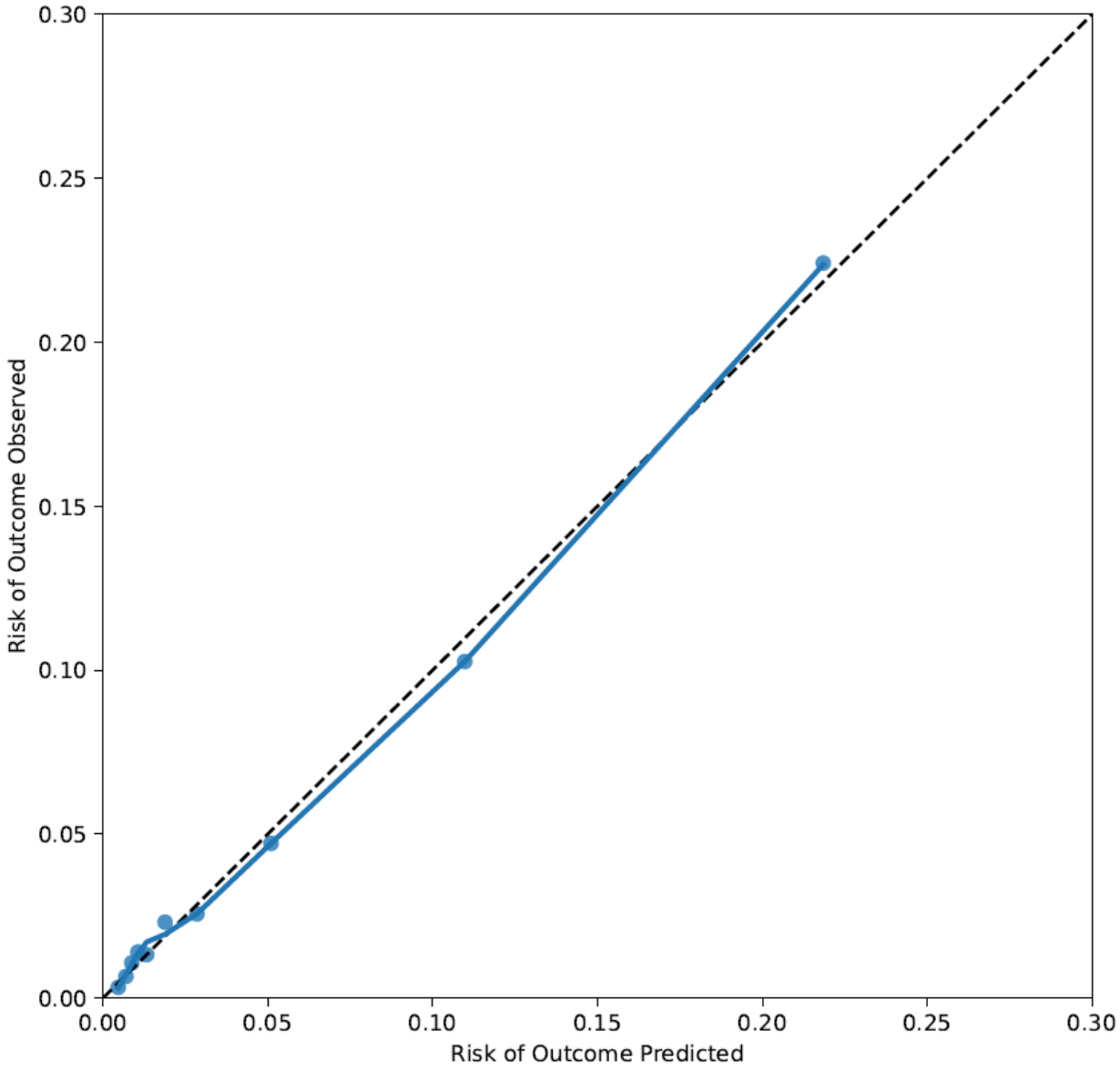


Figure 8: Calibration plot for model 2 (refit)

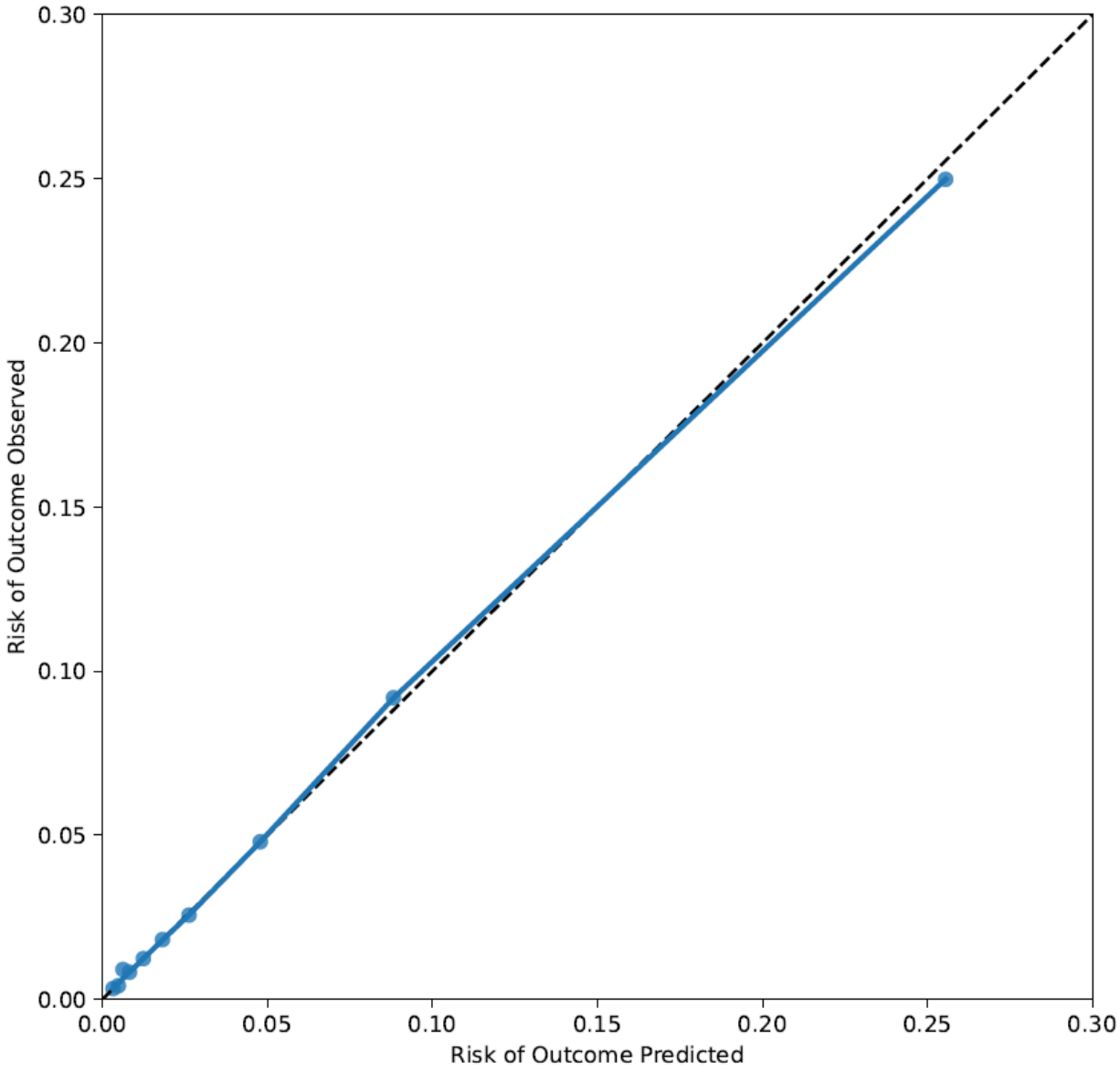


Figure 9: Calibration plot for model 3 (refit)

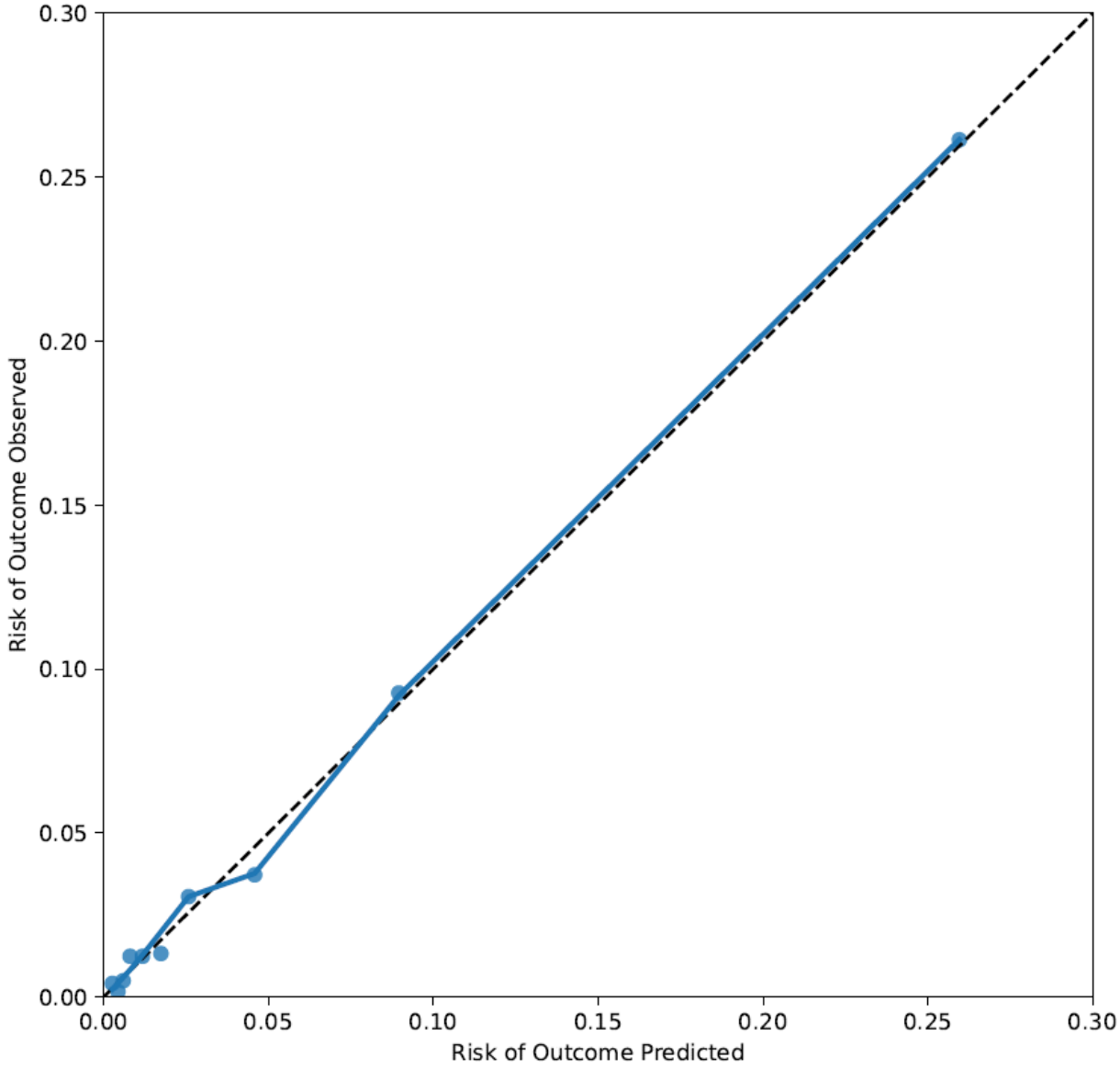
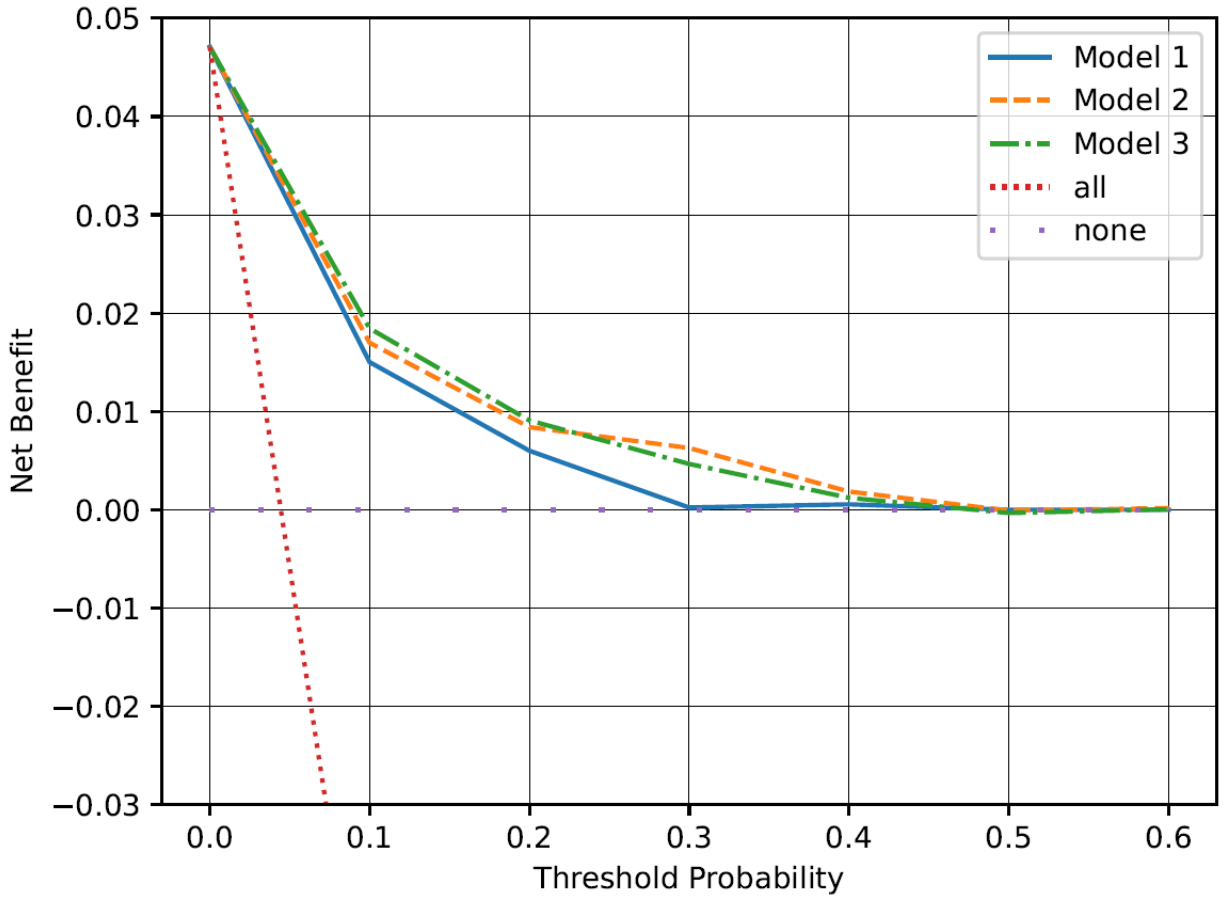


Figure 10: Decision curve analysis



Decision Curve Analysis (DCA) was employed to evaluate the clinical usefulness of the three models. All models demonstrated a similar positive net benefit when compared to strategies where all or none of the participants received a perioperative cardiac monitoring intervention (Figure 10). The models were beneficial, offering a net benefit in guiding perioperative cardiac monitoring decisions at predicted risk thresholds ranging from 0 to 40%.

Table 6 Sensitivity, specificity, NPV and PPV (%) at different thresholds of predicted probability of major postoperative outcomes

	Sensitivity	Specificity	PPV	NPV
5% Threshold				
Model 1	76.4	78.4	14.9	98.5
Model 2	77.5	78.4	15.0	98.6
Model 3	77.2	79.1	15.5	98.6
10% Threshold				
Model 1	63.6	85.9	18.2	97.9
Model 2	59.8	89.5	21.9	97.8
Model 3	63.3	89.3	22.6	98.0
15% Threshold				
Model 1	44.6	92.8	23.5	97.1
Model 2	50.4	93.0	26.3	97.4
Model 3	52.0	93.1	27.1	97.5

Model 1 included age, sex, dialysis modality, surgery type and setting.
Model 2 included model 1 and comorbidities
Model 3 included model 2 and preoperative hemoglobin and albumin

Table 6 shows sensitivity, specificity, NPV and PPV (%) at different thresholds of predicted probability of major postoperative outcomes. Across all thresholds, the models demonstrated a trade-off between sensitivity and specificity. As the threshold increased, sensitivity generally decreased from about 77% at 5% threshold to 50% at 15% threshold for all the models while specificity increased from 78% at 5% threshold to 93% at 15% threshold for all the models. The PPV improved with higher thresholds (about 15% at 5% threshold to 26% at 15% threshold), indicating a greater proportion of true positive predictions as the threshold increased, while the NPV slightly decreased (about 98.5% at 5% threshold to 97% at 15% threshold).

Table 7 Event and non-event reclassification tables between models, stratified by clinically important probability categories

Model 2 versus model 1

		Risk model 2 predicted probability threshold						
Outcome status	**Risk model 1** Predicted probability threshold	<5%	5-15%	> 15%	Total	Correct percentage reclassification	Incorrect percentage reclassification	NRI
Present	< 5%	106	28	0	134	19.9%	13.0%	6.9%
	5-15%	22	74	85	181			
	> 15%		52	202	254			
	Total	128	154	287	569			
Absent	< 5%	8587	435		9022	7.3%	7.1%	0.2%
	5-15%	435	848	379	1662			
	> 15%		405	424	829			
	Total	9022	1688	803	11513			
Category-based Net Absolute Reclassification Index (NARI)	NARI = 5.4 per 1000 patients							

Model 3 versus model 2

		Risk model 3 predicted probability threshold						
Outcome status	**Risk model 2** predicted probability threshold	<5%	5-15%	> 15%	Total	Correct percentage reclassification	Incorrect percentage reclassification	NRI
Present	< 5%	119	9	0	128	5.4%	4.2%	1.2%
	5-15%	11	121	22	154			
	> 15%		13	274	287			
	Total	130	143	296	569			
Absent	< 5%	8847	175		9022	3.2%	2.3%	0.8%
	5-15%	264	1332	92	1688			
	> 15%		100	703	803			
	Total	9111	1607	795	11513			
Category-based Net Absolute Reclassification Index (NARI)	NARI = 8.6 per 1000 patients							

Table 7 demonstrated event and non-event reclassification tables between models and calculated NRI and NARI. In the evaluation of model 2 versus model 1, a NRI of 6.9% was observed for events, with a marginal NRI of 0.2% for non-events. The Net Absolute Reclassification Index (NARI) showed an improvement of 5.4 per 1,000 patients. When comparing model 3 to model 2, the NRI for events achieved by model 3 was 1.2%, and for non-events, there was a slight net improvement of 0.8%. The Net Absolute Reclassification Index (NARI) was 8.6 per 1,000 patients.

Objective 2: Developed and validated a machine learning model

Table 8 shows the distribution of clinical characteristics across the training, validation, and internal testing cohorts. The table suggested a high degree of similarity, indicating that the cohorts are comparable and well-balanced. This similarity supported the validity of the study's findings by ensuring that the predictive models developed from the training data were applicable to similarly composed validation and testing cohorts.

Table 8: Baseline characteristics of development, validation, and testing cohorts

Characteristics	Training	Validation	Testing
Outcome, No. (%)	398 (4.7)	85 (4.7)	86 (4.7)
Age, Median (Q1-Q3)	60.8 (50.4-69.6)	61.4 (50.8-69.8)	60.8 (51.0-69.8)
Female, No. (%)	3642 (43.1)	788 (43.5)	789 (43.5)
Surgery type, No. (%)			
Intra-abdominal	608 (7.2)	136 (7.5)	151 (8.3)
Musculoskeletal	926 (10.9)	205 (11.3)	209 (11.5)
Head and neck	376 (4.4)	84 (4.6)	84 (4.6)
Vascular	1075 (12.7)	268 (14.8)	220 (12.1)
Skin and soft tissue	172 (2.0)	44 (2.4)	39 (2.2)
Neurosurgery	60 (0.7)	13 (0.7)	11 (0.6)

Peritoneal dialysis catheter	1080 (12.8)	211 (11.6)	220 (12.1)
Arteriovenous fistula	1940 (22.9)	388 (21.4)	399 (22.0)
Kidney transplant	278 (3.3)	68 (3.8)	54 (3.0)
Low-risk[a]	1277 (15.1)	276 (15.2)	291 (16.1)
More than one	665 (7.9)	119 (6.6)	135 (7.4)
Surgery setting, No. (%)			
Outpatient	5120 (60.5)	1095 (60.4)	1105 (60.9)
Elective inpatient	1577 (18.6)	327 (18.0)	333 (18.4)
Emergency inpatient	1760 (20.8)	390 (21.5)	375 (20.7)
Kidney failure type, No. (%)			
Non-dialysis	2366 (28.0)	468 (25.8)	498 (27.5)
Hemodialysis	4821 (57.0)	1044 (57.6)	1038 (57.3)
Peritoneal dialysis	1270 (15.0)	300 (16.6)	277 (15.3)
Comorbidities, No. (%)			
Cerebrovascular disease	2525 (29.9)	577 (31.8)	597 (32.9)
Chronic pulmonary disease	6175 (73.0)	1357 (74.9)	1303 (71.9)
Dementia	320 (3.8)	89 (4.9)	82 (4.5)
Diabetes	6202 (73.3)	1350 (74.5)	1355 (74.7)
Heart failure	3956 (46.8)	869 (48.0)	879 (48.5)
History of myocardial infarction	2160 (25.5)	520 (28.7)	446 (24.6)
Hypertension disease	7717 (91.2)	1660 (91.6)	1645 (90.7)
Liver disease (mild and moderate/severe)	1127 (13.3)	264 (14.6)	257 (14.2)
Peripheral vascular disease	2367 (28.0)	512 (28.3)	535 (29.5)
Cancer	1225 (14.5)	279 (15.4)	260 (14.3)

Albumin, Median (Q1-Q3)	33.0 (29.0-36.0)	33.0 (29.0-36.0)	32.2 (28.0-36.0)
Hemoglobin, Median (Q1-Q3)	105.0 (94.0-116.0)	104.0 (94.0-115.0)	105.0 (95.0-114.0)

[a] includes anorectal, breast, ophthalmologic, thoracic, lower urologic and gynecologic, and retroperitoneal surgeries

The results of the random forest models in predicting the risk of major cardiovascular events and death for individuals with kidney failure undergoing non-cardiac surgery were presented in Table 9. This table includes information about the performance of two models across different threshold probabilities. Model 1 was the baseline model and included all the features. Model 2 was optimized model and included selected features namely age, sex, surgery type (vascular, more than one, peritoneal dialysis, arteriovenous fistula, musculoskeletal, low-risk surgeries), surgery setting (emergency inpatient and outpatient), kidney failure type (hemodialysis), comorbidities (history of myocardial infarction, dementia, chronic pulmonary disease, liver disease, peripheral vascular disease, cancer), albumin and hemoglobin values.

Table 9: Model performance (95% CI) of the random forest model at different thresholds of predicted probability of major postoperative outcomes in the testing cohort

Threshold	**5%**		**10%**		**15%**	
Model	**1**	**2**	**1**	**2**	**1**	**2**
Sensitivity	81% (72%-88%)	83% (73%-89%)	62% (51%-71%)	59% (49%-69%)	42% (32%-52%)	44% (34%-55%)
Specificity	76% (74%-78%)	76% (74%-78%)	88% (87%-90%)	89% (87%-90%)	94% (93%-95%)	94% (93%-95%)
PPV	14% (12%-18%)	15% (12%-18%)	21% (16%-26%)	21% (16%-26%)	25% (19%-33%)	26% (20%-34%)
NPV	99% (c	99% (98%-99%)	98% (97%-99%)	98% (97%-98%)	97% (96%-98%)	97% (96%-98%)

AUC-ROC	0.863 (0.825-0.902)	0.863 (0.825-0.902)	0.863 (0.825-0.902)	0.863 (0.825-0.902)	0.863 (0.825-0.902)	0.863 (0.825-0.902)
AUC-PR	0.315 (0.212-0.417)	0.332 (0.232-0.440)	0.315 (0.212-0.417)	0.332 (0.232-0.440)	0.315 (0.212-0.417)	0.332 (0.232-0.440)
Brier	0.039 (0.032-0.047)	0.038 (0.032-0.046)	0.039 (0.032-0.047)	0.038 (0.032-0.046)	0.039 (0.032-0.047)	0.038 (0.032-0.046)

Model 1: Baseline model includes all features namely age, sex, surgery type, surgery setting, kidney failure type, comorbidities, albumin and hemoglobin values

Model 2: Includes selected features namely age, sex, surgery type (vascular, more than one, peritoneal dialysis, arteriovenous fistula, musculoskeletal, low-risk surgeries), surgery setting (emergency inpatient and outpatient), kidney failure type (hemodialysis), comorbidities (history of myocardial infarction, dementia, chronic pulmonary disease, liver disease, peripheral vascular disease, cancer), albumin and hemoglobin values

For a threshold probability of 5%, model 1 demonstrated a sensitivity of 81% (95% CI 72%-88%) and a specificity of 76%(95% CI 74%-78%), with a PPV of 14%(95% CI 12%-18%) and a NPV of 99%(95% CI 12%-18%). The AUC-ROC was 0.863 (95% CI 0.825-0.902) and the area under the precision-recall curve was 0.315 (95% CI 0.212-0.417). AUC-ROC indicated good discriminative ability and AUC-PR reflected the good balance between precision and recall. Brier score for both the models was low, 0.039 (95% CI 0.032-0.047) for model 1 and 0.038 (95% CI 0.032-0.046) for model 2, indicating good model performance. Model 2, under the same threshold, shows slightly higher sensitivity and PPV but maintained the same specificity and NPV. The AUC-ROC was similar to model 1 but AUC-PR was slightly higher than those of model 1.

As the threshold probability increased to 10%, both models displayed a decrease in sensitivity to 62% (95% CI 51%-71%) for model 1 and 59% (95% CI 49%-69%) for model 2 while specificity increased to 88% (95% CI 87%-90%) for model 1 and 89% (95% CI 87%-90%) for model 2. The PPV and NPV also adjusted to 21% and 98%, respectively. At a threshold probability of 15%, there was a further decrease in sensitivity, 42% (95% CI 32%-52%) for model 1 and 44% (95% CI 34%-55%) for model 2, and an increase in specificity to 94% (95% CI 93%-

95%) for both models. The PPV increased to about 26%, and the NPV slightly decreased to 97% (95% CI 96%-98%). The calibration plot for both the random forest models demonstrated strong calibration throughout, except for overestimation at the highest predicted risks, as shown in figures 11 and 12.

Figure 11: Calibration plot for random forest (model 1) in the testing cohort

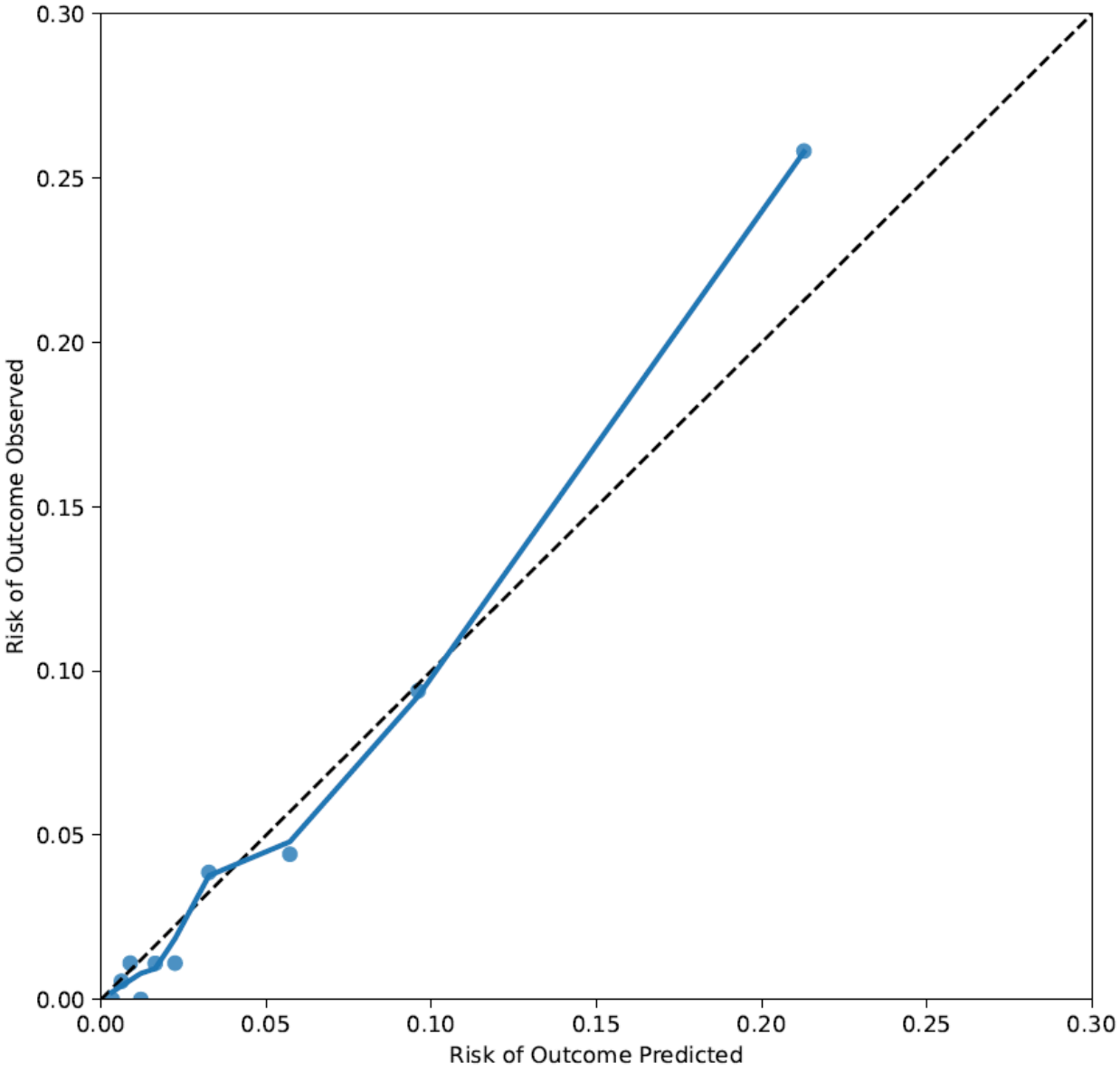
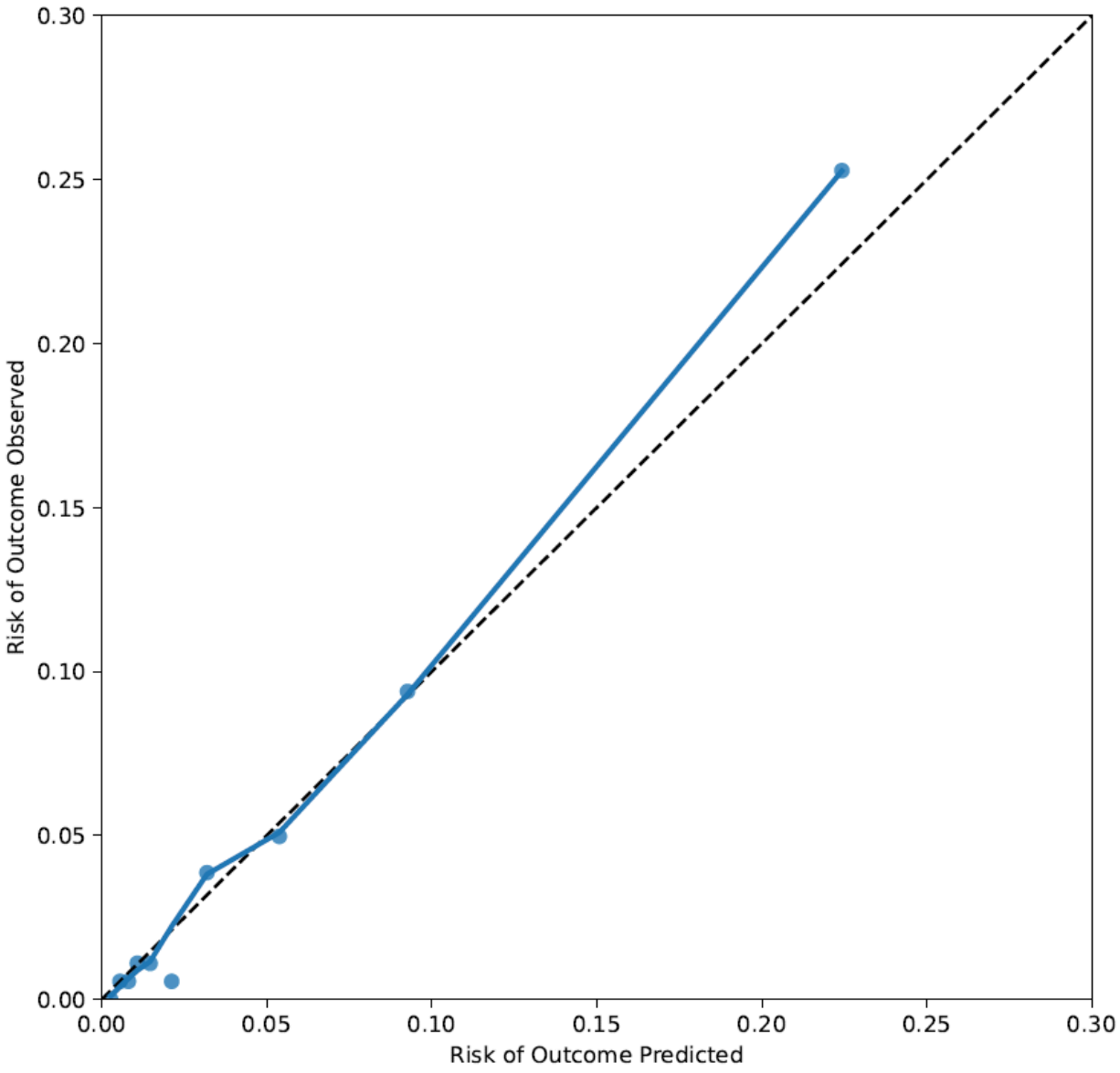


Figure 12: Calibration plot for random forest (model 2) in the testing cohort



We further evaluated the XGBoost model results and presented them in Table 10. Model 1 was the baseline model and included all the features. Model 2 was optimized model and included selected features namely surgery type (vascular, more than one, peritoneal dialysis), surgery setting (emergency inpatient and outpatient), comorbidities (history of myocardial infarction), albumin and hemoglobin values. At a threshold probability of 5%, the model 1 demonstrated a sensitivity of 74% (95% CI 64%-82%) and specificity of 81% (95% CI 79%-83%). The PPV and NPV were 16% (95% CI 13%-20%) and 98%(95% CI 98%-99%), respectively. The

AUC-ROC for this model was 0.863 (95% CI 0.826-0.900), indicating a strong discriminative ability between those who would and would not experience the event. The AUC-PR was 0.306 (95% CI 0.215-0.410), reflecting the model's effectiveness in identifying true positives among the predicted positives.

Table 10: Model performance (95% CI) of the XGBoost model at different thresholds of predicted probability of major postoperative outcomes in the testing cohort

Threshold	**5%**		**10%**		**15%**	
Model	**1**	**2**	**1**	**2**	**1**	**2**
Sensitivity	74% (64%-82%)	74% (64%-82%)	59% (49%-69%)	60% (50%-70%)	50% (40%-60%)	49% (39%-59%)
Specificity	81% (79%-83%)	80% (78%-82%)	90% (89%-92%)	89% (88%-91%)	93% (92%-95%)	93% (92%-94%)
PPV	16% (13%-20%)	16% (13%-20%)	23% (18%-29%)	22% (17%-28%)	28% (21%-35%)	26% (20%-33%
NPV	98% (98%-99%)	98% (98%-99%)	98% (97%-98%)	98% (97%-98%)	97% (97%-98%)	97% (96%-98%)
AUC-ROC	0.863 (0.826-0.900)	0.861 (0.824-0.899)	0.863 (0.826-0.900)	0.861 (0.824-0.899)	0.863 (0.826-0.900)	0.861 (0.824-0.899)
AUC-PR	0.306 (0.215-0.410)	0.304 (0.210-0.411)	0.306 (0.215-0.410)	0.304 (0.210-0.411)	0.306 (0.215-0.410)	0.304 (0.210-0.411)
Brier	0.038 (0.032-0.046)	0.038 (0.032-0.046)	0.038 (0.032-0.046)	0.038 (0.032-0.046)	0.038 (0.032-0.046)	0.038 (0.032-0.046)

Model 1: Baseline model includes all features namely age, sex, surgery type, surgery setting, kidney failure type, comorbidities, albumin and hemoglobin values

Model 2: Includes selected features namely surgery type (vascular, more than one, peritoneal dialysis), surgery setting (emergency inpatient and outpatient), comorbidities (history of myocardial infarction), albumin and hemoglobin values

For the same threshold probability, model 2 showed a sensitivity of 74% (95% CI 64%-82%) and a slightly lower specificity of 80% (95% CI 78%-82%). The PPV and NPV were 16% (95% CI 13%-20%) and 98% (95% CI 98%-99%), respectively, with a AUC-ROC of 0.861 (95% CI 0.824-0.899) and an AUC-PR of 0.304 (95% CI 0.210-0.411). These metrics indicated a comparable performance to the first model. Brier score for both the models was low, 0.038 (95% CI 0.032-0.046) for model 1 and model 2.

As the threshold probability increased to 10% and then to 15%, both models exhibited a general trend of increased specificity and NPV, alongside decreased sensitivity and PPV. This trend illustrated the models enhanced ability to accurately exclude individuals without the risk at higher thresholds, albeit at the expense of missing some true positive cases. At a threshold of 10%, model 1 reported a sensitivity of 59% (95% CI 49%-69%), specificity of 90% (95% CI 89%-92%), PPV of 23% (95% CI 18%-29%), and NPV of 98% (95% CI 97%-98%). Model 2, under the same threshold, reported a sensitivity of 60% (95% CI 50%-70%), specificity of 89% (95% CI 88%-91%), PPV of 22% (95% CI 17%-28%), and NPV of 98% (95% CI 97%-98%). At a threshold of 15%, model 1 reported a sensitivity of 50% (95% CI 40%-60%), specificity of 93% (95% CI 92%-95%), PPV of 28% (95% CI 21%-35%), and NPV of 97% (95% CI 97%-98%). Model 2, under the same threshold, showed a sensitivity of 49% (95% CI 39%-59%), specificity of 93% (95% CI 92%-94%), PPV of 26% (95% CI 92%-94%), and NPV of 97% (95% CI 96%-98%). The calibration plot for both the XGBoost models demonstrated strong calibration throughout as shown in figures 13 and 14.

Figure 13: Calibration plot for XGBoost model (model 1) in the testing cohort

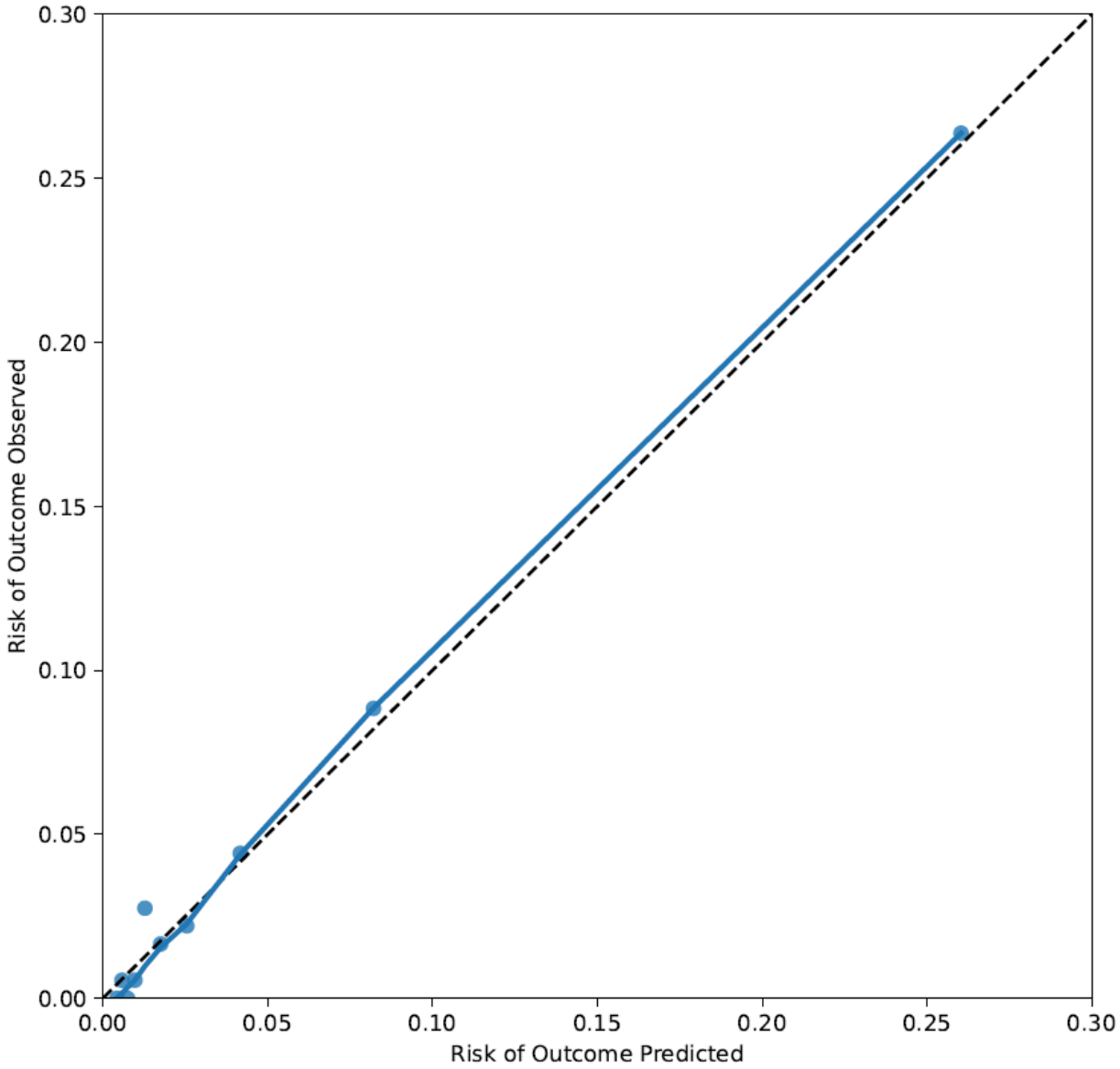


Figure 14: Calibration plot for XGBoost model (model 2) in the testing cohort

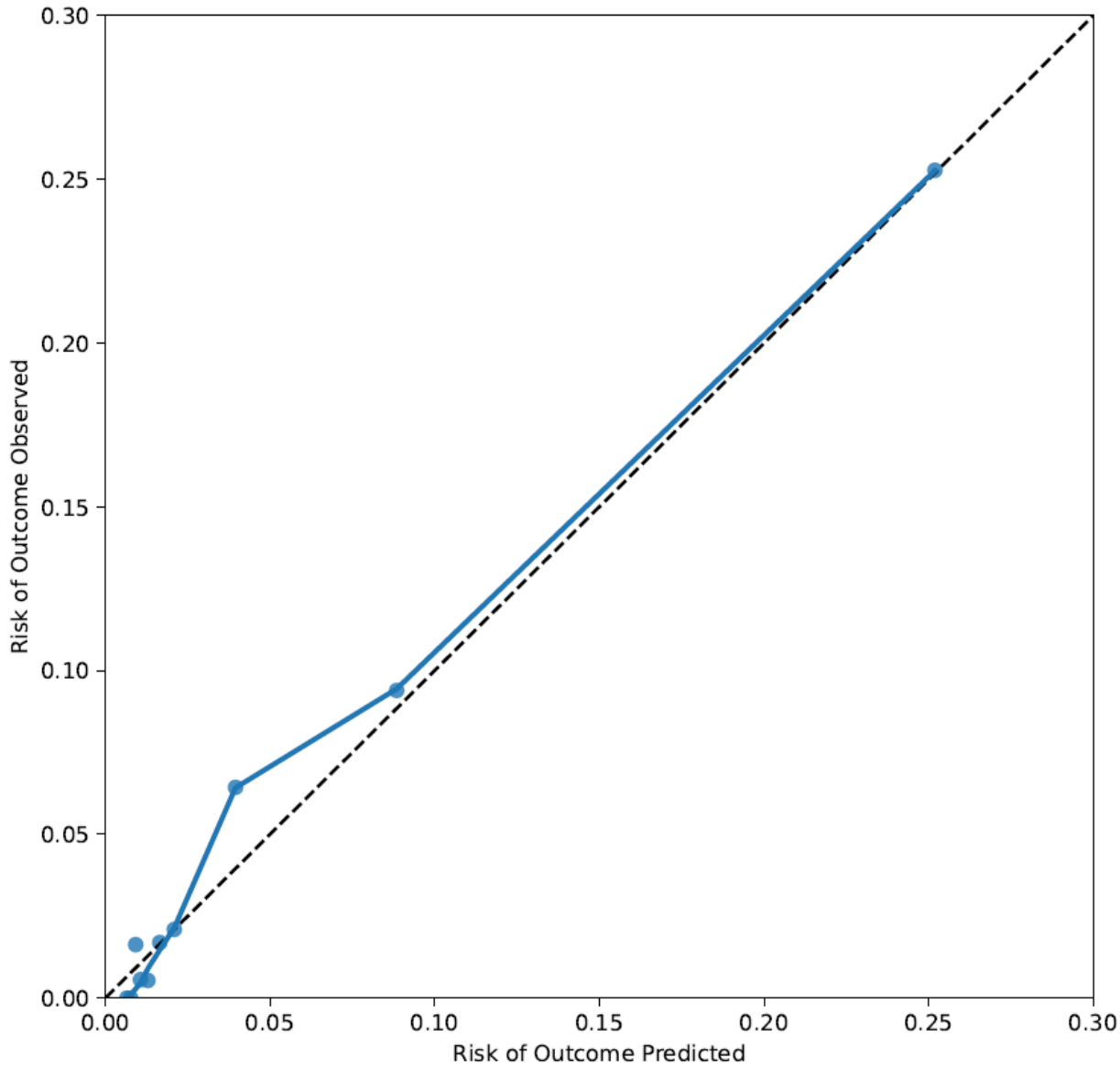


Table 11: Cross-validation results of random forest models

Model no.	Mean AUC-ROC	SD AUC-ROC	Mean AUC-PR	SD AUC-PR
1	0.91	0.009	0.424	0.043
2	0.911	0.008	0.424	0.043

Model 1: Baseline model includes all features namely age, sex, surgery type, surgery setting, kidney failure type, comorbidities, albumin and hemoglobin values

Model 2: Includes selected features namely age, sex, surgery type (vascular, more than one, peritoneal dialysis, arteriovenous fistula, musculoskeletal, low-risk surgeries), surgery setting (emergency inpatient and outpatient), kidney failure type

(hemodialysis), comorbidities (history of myocardial infarction, dementia, chronic pulmonary disease, liver disease, peripheral vascular disease, cancer), albumin and hemoglobin values

Table 12: Cross-validation results for XGBoost models

Model no.	Mean AUC-ROC	SD AUC-ROC	Mean AUC-PR	SD AUC-PR
1	0.875	0.011	0.354	0.042
2	0.862	0.015	0.328	0.044

Model 1: Baseline model includes all features namely age, sex, surgery type, surgery setting, kidney failure type, comorbidities, albumin and hemoglobin values
Model 2: Includes selected features namely surgery type (vascular, more than one, peritoneal dialysis), surgery setting (emergency inpatient and outpatient), comorbidities (history of myocardial infarction), albumin and hemoglobin values

Table 11 and 12 presented the cross-validation results for random forest and XGBoost models, showcasing the mean and SD for both the AUC-ROC and the AUC-PR for each model. In random forest cross-validation results (table 11), model 1 achieved a mean AUC-ROC of 0.91 with a SD of 0.009, alongside a mean AUC-PR of 0.424 with an SD of 0.043. Model 2 reported slightly higher performance metrics, with a mean AUC-ROC of 0.911 and an SD of 0.008, and a mean AUC-PR of 0.424 with an SD of 0.043.

In cross-validation of XGBoost models (table 12), model 1 exhibited a mean AUC-ROC of 0.875 with a standard deviation of 0.011, and a mean AUC-PR of 0.354 with a standard deviation of 0.042. Model 2 shows a slightly lower mean AUC-ROC of 0.862 with the standard deviation of 0.015, and mean AUC-PR of 0.328 with a standard deviation of 0.044.

The relatively low standard deviations in both AUC-ROC and AUC-PR indicated a high level of consistency in their performance across the different folds of the cross-validation process.

Table 13 presented the performance metrics of random forest on Alberta data evaluated at different thresholds (5%, 10%, and 15%). At a 5% threshold, model 1 shows a sensitivity of 63% and a specificity of 87%, with a positive predictive value (PPV) of 13% and a negative predictive value (NPV) of 99%. Model 2 has similar metrics with slightly lower sensitivity and

higher specificity and PPV. As thresholds increase to 10% and 15%, the sensitivity of both models decreases significantly, while specificity and PPV improve, especially at 15% where the PPV reaches 54% and 60% for model 1 and model 2, respectively. The AUC-ROC was 0.842 (95% CI 0.831-0.854) and 0.839 (95% CI 0.827-0.851) for models 1 and 2, respectively. These scores were slightly lower than the AUC-ROC of 0.863 obtained from the testing cohort (table 9). However, AUC-PR scores were slightly higher in external validation, 0.334 (95% CI 0.308-0.361) for model 1 and 0.340 (95% CI 0.314-0.367) for model 2, compared to 0.315 for model 1 and 0.332 for model 2 in the testing cohort (table 9). The calibration plot for both the random forest models demonstrated strong calibration throughout in Alberta cohort, except for underestimation at the highest predicted risks, as shown in figures 15 and 16.

Table 13: Model performance in external validation of the random forest model at different thresholds of predicted probability of major postoperative outcomes in the Alberta cohort

Threshold	**5%**		**10%**		**15%**	
Model	**1**	**2**	**1**	**2**	**1**	**2**
Sensitivity	63% (60%-66%)	61% (58%-64%)	41% (38%-43%)	39% (37%-42%)	23% (21%-25%)	23% (21%-26%)
Specificity	87% (86%-87%)	88% (88%-89%)	96% (96%-96%)	96% (96%-97%)	99% (99%-99%)	99% (99%-100%)
PPV	13% (12%-14%)	14% (12%-15%)	24% (22%-26%)	26% (24%-28%)	54% (50%-59%)	60% (55%-64%)
NPV	99% (99%-99%)	99% (98%-99%)	98% (98%-98%)	98% (98%-98%)	98% (97%-98%)	98% (97%-98%)
AUC-ROC	0.842 (0.831-0.854)	0.839 (0.827-0.851)	0.842 (0.831-0.854)	0.839 (0.827-0.851)	0.842 (0.831-0.854)	0.839 (0.827-0.851)
AUC-PR	0.334 (0.308-0.361)	0.340 (0.314-0.367)	0.334 (0.308-0.361)	0.340 (0.314-0.367)	0.334 (0.308-0.361)	0.340 (0.314-0.367)

Brier	0.027 (0.026-0.028)	0.027 (0.025-0.028)	0.027 (0.026-0.028)	0.027 (0.025-0.028)	0.027 (0.026-0.028)	0.027 (0.025-0.028)

Model 1: Baseline model includes age, sex, surgery type, surgery setting, kidney failure type, comorbidities, albumin and hemoglobin values

Model 2: Includes selected features namely age, sex, surgery type (vascular, more than one, peritoneal dialysis, arteriovenous fistula, musculoskeletal, low-risk surgeries), surgery setting (emergency inpatient and outpatient), kidney failure type (hemodialysis), comorbidities (history of myocardial infarction, dementia, chronic pulmonary disease, liver disease, peripheral vascular disease, cancer), albumin and hemoglobin values

Figure 15: Calibration plot for random forest (model 1) in the Alberta cohort

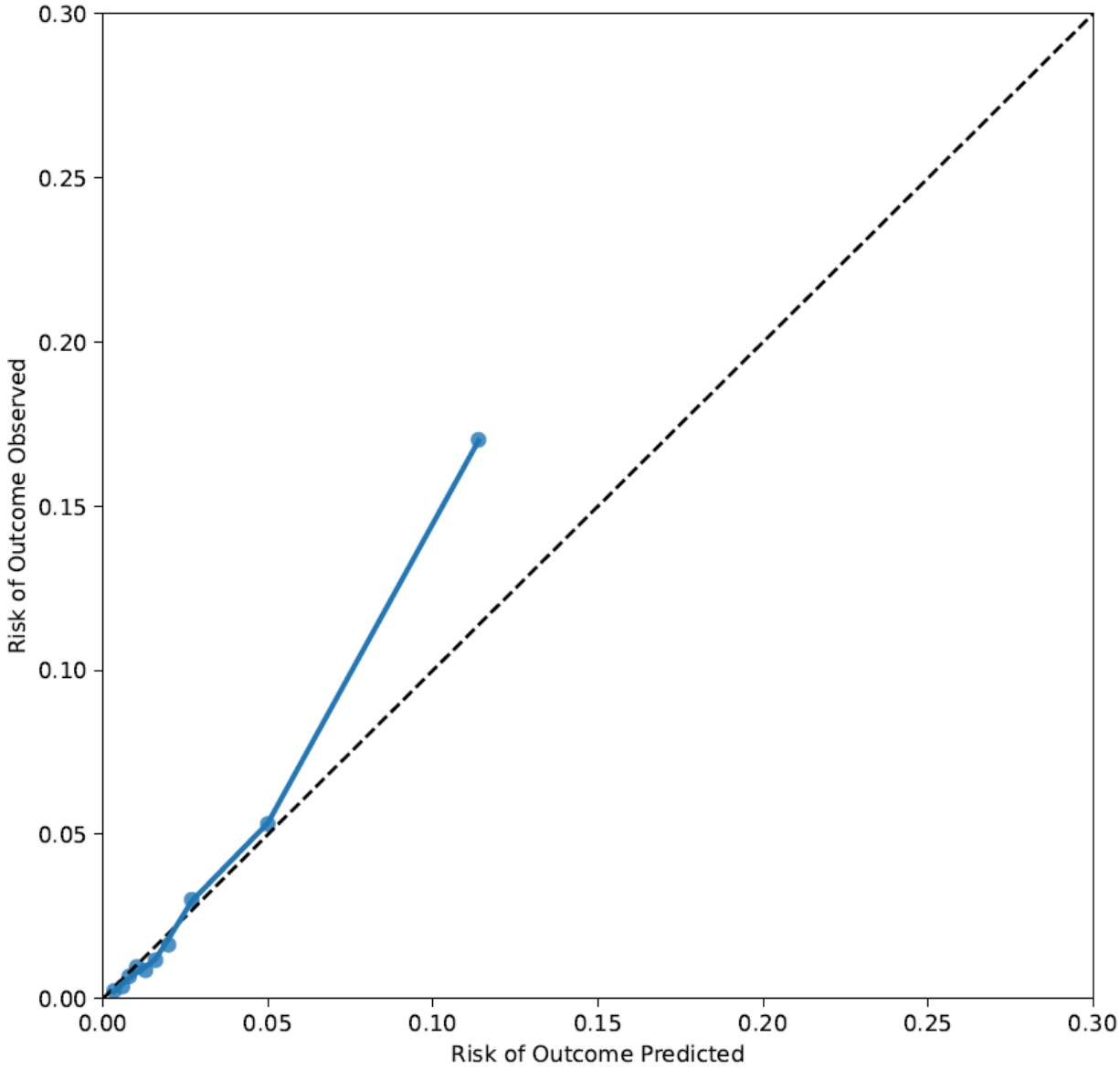


Figure 16: Calibration plot for random forest (model 2) in the Alberta cohort

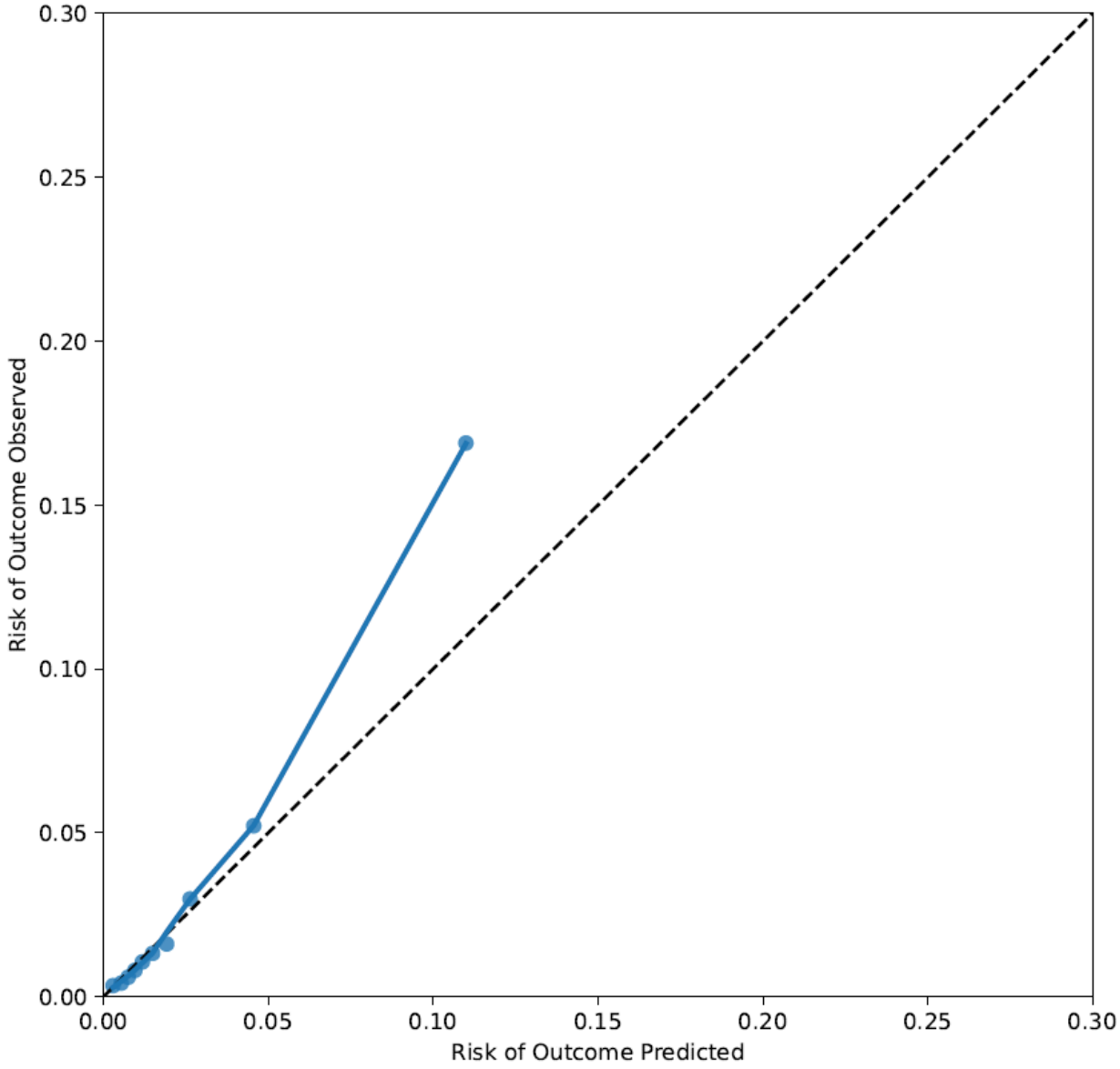


Table 14 displayed the performance metrics of XGBoost on Alberta data evaluated at different thresholds (5%, 10%, and 15%). At a 5% threshold, both models have a sensitivity of 56%. Model 1 shows a slightly higher specificity (91%) compared to model 2 (90%). Both models have very high NPVs (98%) and the PPV was relatively low (16% for both models). At a 10% threshold, the sensitivity for both models decreased, with model 1 at 34% and model 2 at 32%. Specificity increased to 98% for both model 1 and model 2. The PPV almost doubled, 31% for

model 1 and 34% for model 2. Both models maintained a high NPV of 98%. At a 15% threshold, there was a further reduction in sensitivity, with model 1 at 22% and model 2 at 23%, while specificity improved for both models (99% for model 1 and 100% for model 2). PPV improved drastically to 57% for model 1 and 67% for model 2, while NPV was constant at 98%. The AUC-ROC values were 0.849 (95% CI 0.837-0.861) and 0.841 (95% CI 0.828-0.854) in models 1 and 2, respectively. They were slightly lower than AUC-ROC in the testing cohort (0.863 in model 1 and 0.861 in model 2, table 10). However, the AUC-PR was 0.335 (95% CI 0.309-0.362) and 0.350 (95% CI 0.323-0.376) for models 1 and 2, respectively, which was slightly higher than in testing cohorts (0.306 in model 1 and 0.304 in model 2, table 10). Brier score for both the models was low at 0.027 (95% CI 0.025-0.028). The calibration plot for both the XGBoost models demonstrated strong calibration throughout in the Alberta cohort, except for underestimation at the highest predicted risks, as shown in figures 17 and 18.

Table 14: Model performance in external validation of the XGBoost at different thresholds of predicted probability of major postoperative outcomes in the Alberta cohort

Threshold	**5%**		**10%**		**15%**	
Model	**1**	**2**	**1**	**2**	**1**	**2**
Sensitivity	56% (53%-59%)	56% (53%-59%)	34% (31%-37%)	32% (29%-34%)	22% (20%-25%)	23% (21%-25%)
Specificity	91% (90%-91%)	90% (90%-91%)	98% (97%-98%)	98% (98%-98%)	99% (99%-100%)	100% (100%-100%)
PPV	16% (15%-17%)	16% (15%-17%)	31% (28%-33%)	34% (31%-37%)	57% (52%-61%)	67% (63%-72%)
NPV	98% (98%-99%)	98% (98%-99%)	98% (98%-98%)	98% (98%-98%)	98% (97%-98%)	98% (97%-98%)
AUC-ROC	0.849 (0.837-0.861)	0.841 (0.828-0.854)	0.849 (0.837-0.861)	0.841 (0.828-0.854)	0.849 (0.837-0.861)	0.841 (0.828-0.854)

AUC-PR	0.335 (0.309-0.362)	0.350 (0.323-0.376)	0.335 (0.309-0.362)	0.350 (0.323-0.376)	0.335 (0.309-0.362)	0.350 (0.323-0.376)
Brier	0.027 (0.025-0.028)	0.027 (0.025-0.028)	0.027 (0.025-0.028)	0.027 (0.025-0.028)	0.027 (0.025-0.028)	0.027 (0.025-0.028)

Model 1: Baseline model includes all features namely age, sex, surgery type, surgery setting, kidney failure type, comorbidities, albumin and hemoglobin values

Model 2: Includes selected features namely surgery type (vascular, more than one, peritoneal dialysis), surgery setting (emergency inpatient and outpatient), comorbidities (history of myocardial infarction), albumin and hemoglobin values

Figure 17: Calibration plot for XGBoost model (model 1) in the Alberta cohort

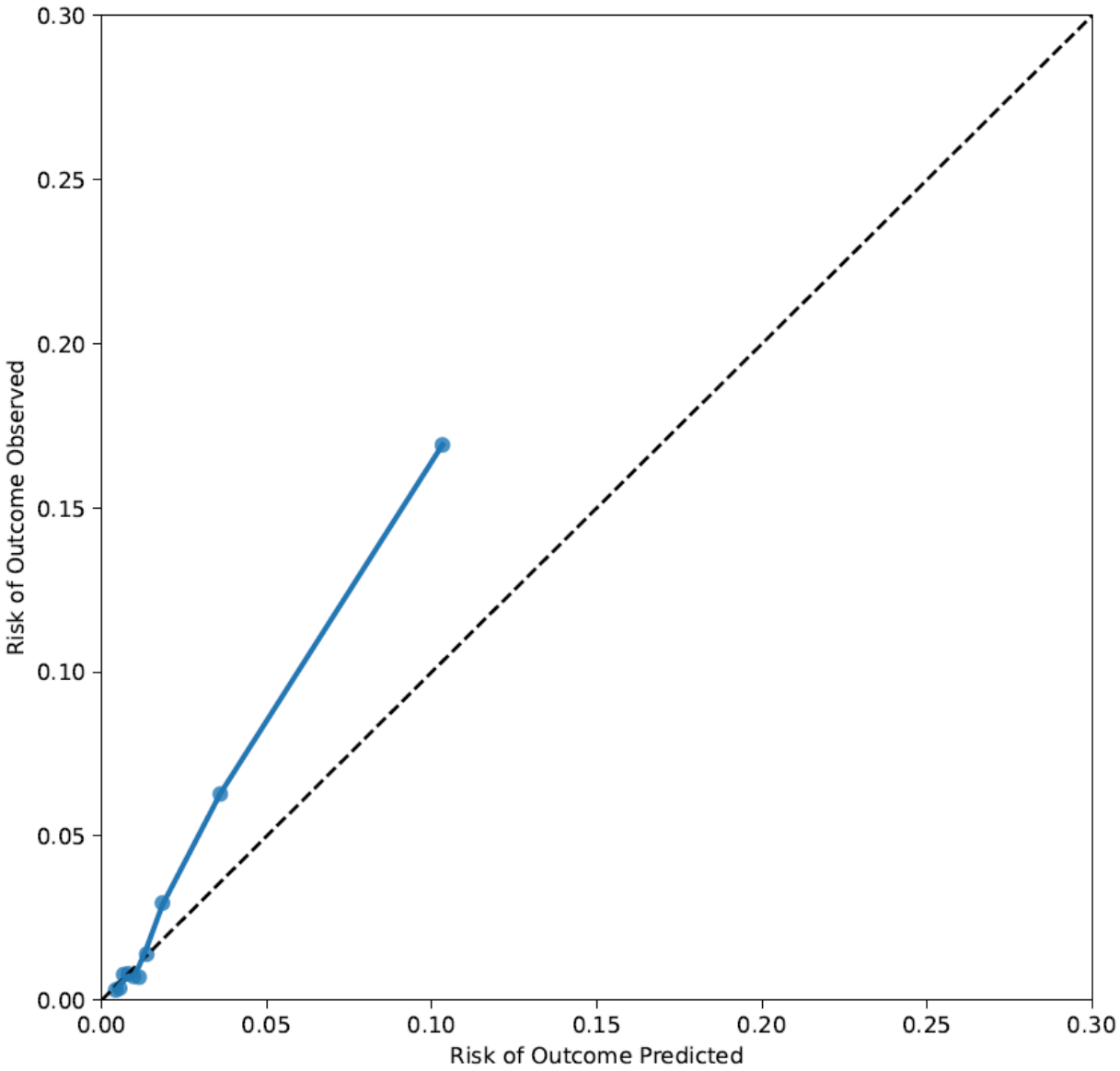
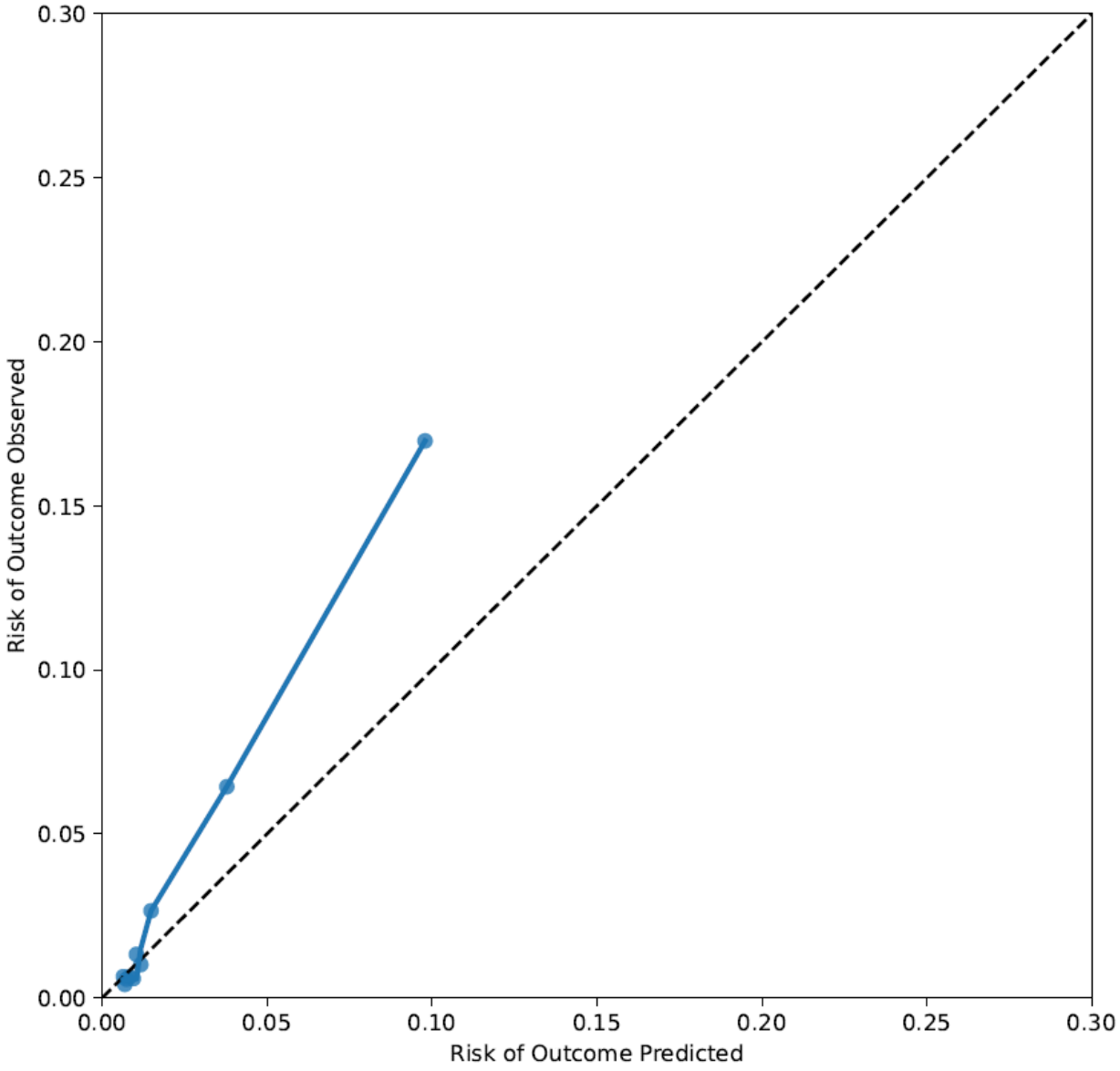


Figure 18: Calibration plot for XGBoost model (model 2) in the Alberta cohort



Both random forest (RF) and XGBoost models exhibited commendable and comparable performance. In the testing cohort, while XGBoost utilized a smaller set of features, random forest achieved a slightly higher AUC-ROC and AUC-PR scores. However, in the Alberta cohort, XGBoost model displayed a slightly higher AUC-ROC and AUC-PR scores than random forest model. We provided a 8 variable XGBoost model and a 20 variable random forest model, both of which demonstrated excellent performance.

Sensitivity analysis:

Table 15: Random forest model performance (95% CI) in first surgery cohort

	Retrain and apply[1]		Apply model[2]	
Surgeries	4,175			
Model	1	2	1	2
ROC-AUC	0.851 (0.792-0.910)	0.842 (0.779-0.904)	0.915 (0.896-0.935)	0.917 (0.898-0.937)
AUC-PR	0.339 (0.206-0.471)	0.310 (0.189-0.445)	0.430 (0.357-0.506)	0.439 (0.365-0.514)
Brier	0.035 (0.028-0.045)	0.035 (0.028-0.045)	0.033 (0.029-0.037)	0.032 (0.028-0.037)

[1] Split the cohort into training (70%) and test set (30%). Retrained the models on training set and assessed performance on test set.

[2] Applied the pretrained models on the cohort and assessed their performance.

Model 1: Baseline model includes all features namely age, sex, surgery type, surgery setting, kidney failure type, comorbidities, albumin and hemoglobin values

Model 2: Includes selected features namely age, sex, surgery type (vascular, more than one, peritoneal dialysis, arteriovenous fistula, musculoskeletal, low-risk surgeries), surgery setting (emergency inpatient and outpatient), kidney failure type (hemodialysis), comorbidities (history of myocardial infarction, dementia, chronic pulmonary disease, liver disease, peripheral vascular disease, cancer), albumin and hemoglobin values

The performance of the random forest models on the cohort restricted to only the first surgery per participant, which consisted of 4,175 surgeries, was summarized in table 15. Model 1 achieved a ROC-AUC of 0.851, an AUC-PR of 0.339, and a Brier score of 0.035 when retrained on the training set and evaluated on the test set. Model 2 showed a ROC-AUC of 0.842, an AUC-PR of 0.310, and a Brier score of 0.035. However, when the original pre-trained models were applied to the full cohort, model 1 ROC-AUC was 0.915, an AUC-PR of 0.430, and the Brier score of 0.033. Model 2 exhibited a slight increase in performance with a ROC-AUC of 0.917, AUC-PR of 0.439, and a Brier score of 0.032.

Table 16: XGBoost model performance (95% CI) in first surgery cohort

	Retrain and apply[1]		Apply model[2]	
Surgeries	4,175			
Model	1	2	1	2
ROC-AUC	0.855 (0.795-0.915)	0.852 (0.793-0.911)	0.876 (0.850-0.903)	0.856 (0.826-0.887)

AUC-PR	0.342 (0.215-0.484)	0.307 (0.176-0.428)	0.353 (0.279-0.430)	0.329 (0.029-0.038)
Brier	0.034 (0.027-0.044)	0.035 (0.027-0.044)	0.033 (0.256-0.406)	0.034 (0.030-0.039)

[1] Split the cohort into training (70%) and test set (30%). Retrained the models on training set and assessed performance on test set.

[2] Applied the pretrained models on the cohort and assessed their performance.

Model 1: Baseline model includes all features namely age, sex, surgery type, surgery setting, kidney failure type, comorbidities, albumin and hemoglobin values

Model 2: Includes selected features namely surgery type (vascular, more than one, peritoneal dialysis), surgery setting (emergency inpatient and outpatient), comorbidities (history of myocardial infarction), albumin and hemoglobin values

For the XGBoost model (table 16), model 1 achieved a ROC-AUC of 0.855, an AUC-PR of 0.342, and a Brier score of 0.034 when retrained and tested. Model 2 had slightly lower performance metrics with a ROC-AUC of 0.852, AUC-PR of 0.307, and a Brier score of 0.035. When the pre-trained models were applied to the full cohort, model 1 exhibited ROC-AUC was 0.876, an AUC-PR of 0.353, and the Brier score was 0.033. Model 2 showed ROC-AUC of 0.856, AUC-PR of 0.329, and a Brier score of 0.034.

When the pre-trained models were applied to the first surgery cohort, the calibration plots for both random forest and XGBoost models (figures 19 and 20) demonstrated strong calibration except for underestimation at the highest predicted risks. However, when models were retrained and applied, they exhibited poor calibration. Both discrimination and calibration results indicated that applying the pre-trained models on the first surgery cohort performed better then retrained models.

Figure 19: Calibration plot for random forest models in the first surgery cohort

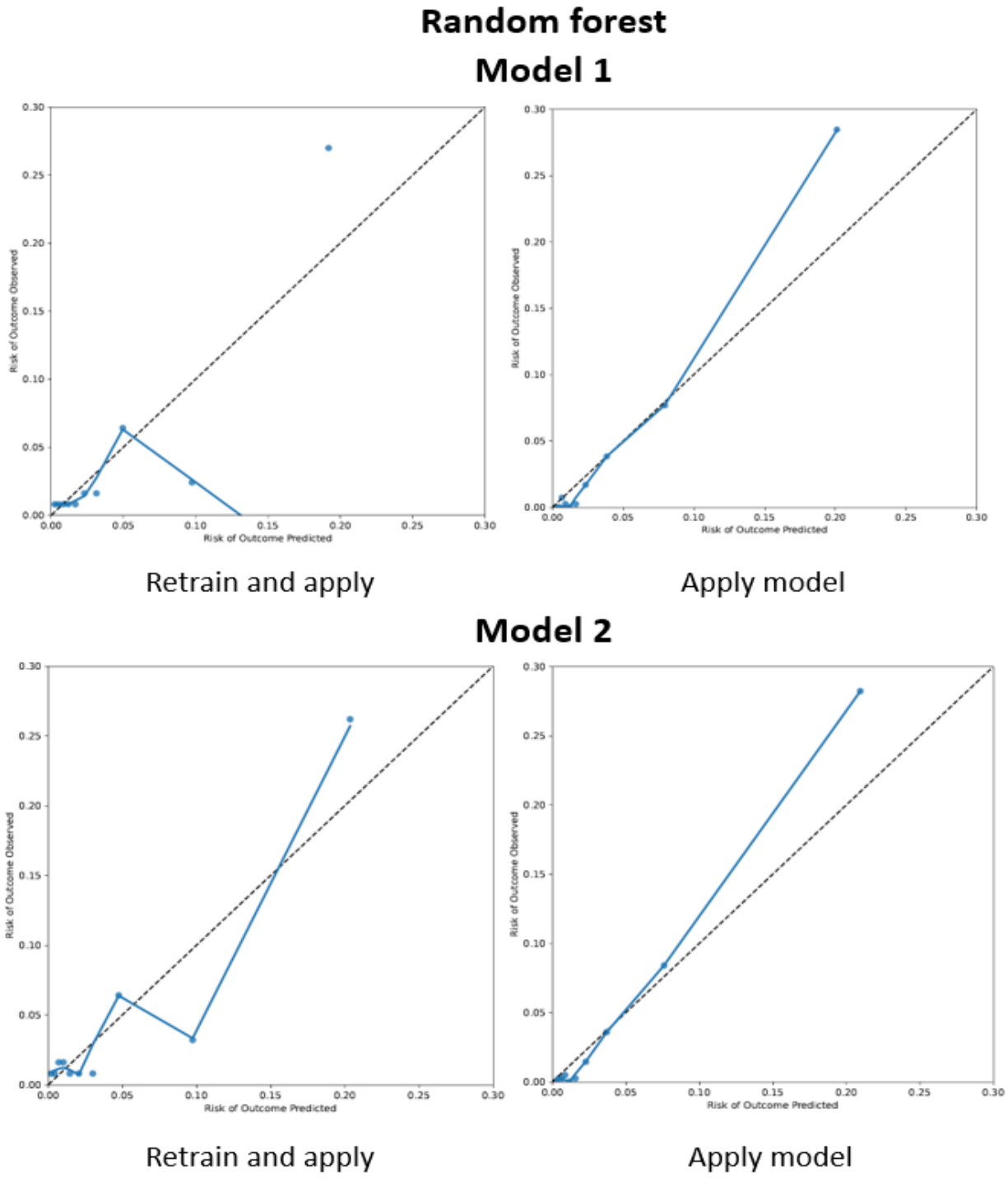
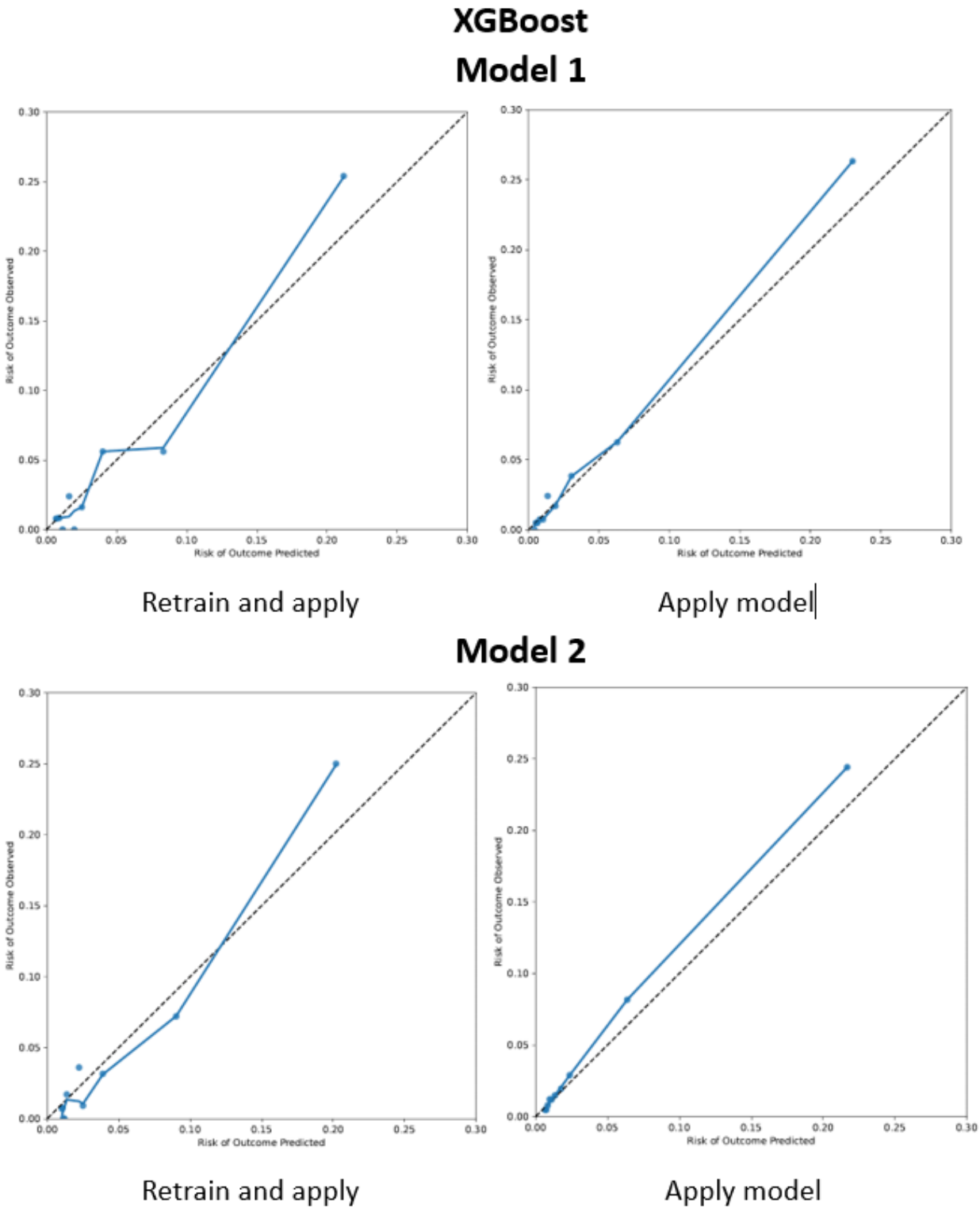


Figure 20: Calibration plot for XGBoost models in the first surgery cohort



Discussion

In this external validation study, we assessed the performance of the AKDN models in an independent population. The models indicated good discriminative ability between those who would and would not experience the event. The calibration plots showed strong calibration throughout, except for overestimation at the highest predicted risks. We developed a new parsimonious machine-learning models for perioperative outcomes in patients with kidney failure undergoing non-cardiac surgery. Our findings show that machine-based approaches to predict peri-operative events performed well. These methods also demonstrated clinical utility and usability through their use of routinely collected data. Our findings support the use of either modeling approach and turn the focus to knowledge translation of these findings to clinical decision support through integration in electronic health records and patient-facing decision aids.

Most patients facing major surgery are appropriately concerned about the risk of adverse perioperative outcomes. In the setting of elective surgery, the risk of outcomes, if high, may inform the decision to forego the procedure and consider a conservative approach. Existing risk prediction tools may not adequately consider the specific variables associated with kidney failure, which can lead to underestimation or overestimation of risk. RCRI is a commonly used tool that provides perioperative cardiac risk assessment and management guidelines for patients undergoing non-cardiac surgery. However, the tool showed poor external validity among the cohort of people with kidney failure.[43] The inclusion of kidney failure-specific predictors as well as a focus on building models exclusively for this population, as done in these models helps fill this gap, providing clinicians with a more refined tool. To our knowledge, this research describes the first two externally validated prediction models that predict the risk of major cardiovascular events and death for people with kidney failure having non-cardiac surgery.

Patients with kidney failure were more likely to develop post-operative complications and mortality.[9-13] Incorporating unique predictors that affect those with kidney failure

addressed the limitations of the current perioperative risk prediction tools. Both AKDN and ML models offer a more tailored and precise approach to risk prediction, allowing for better preparedness and personalized care strategies. The potential for these models to transform care practices underscores the need for continued research and validation in diverse clinical settings, aiming to standardize and improve the quality of care for patients with kidney failure undergoing surgery.

We also developed and externally validated machine learning algorithms to predict the risk of major cardiovascular events and death for people with kidney failure having non-cardiac surgery. Previously, much clinical data was either overlooked or not collected due to its complexity and the lack of collection and storage methods. Now, advancements in data collection are opening doors to better analyze this data. Machine learning algorithms are being utilized more frequently to assist in predicting clinical outcomes, to enhance healthcare epidemiology by leveraging the collected data more effectively.[25]

In various studies, machine learning models have consistently demonstrated superior performance over logistic regression. For instance, the random forest model outperformed the logistic regression model in predicting diabetes with higher precision, recall, and area under the curve values in a study conducted in Saudi Arabia.[75] Another meta-analysis compared logistic regression with machine learning models for prognostic prediction studies in pregnancy care and found that machine learning algorithms performed better than logistic regression models for several outcomes.[76] Similarly, machine learning models outshone traditional logistic regression models in predicting cardiovascular risks among Chinese patients with hypertension.[77] At the same time, there are also studies where machine learning models performed at or below par with logistic regression models and logistic regression was chosen for better model interpretability and simplicity.[35,78-82]

Our results indicated that random forest and XGBoost models could effectively identify high-risk patients for major cardiovascular events and death. We tested several thresholds of

the predicted probabilities, and each threshold provided a different balance in accurately detecting those truly at risk and correctly ruling out those who are not. This offered crucial insights for customizing risk assessment strategies when caring for kidney failure patients who are undergoing non-cardiac surgery. We provided a 8 variable XGBoost model and a 20 variable random forest model, both of which demonstrated excellent performance.

Study strengths

A noteworthy strength of our study was the external validation of the AKDN models, which were developed using Alberta data, using Manitoba data. By successfully applying these models to a separate cohort, we demonstrated their potential for widespread use and contributed valuable evidence to support their effectiveness outside their initial development cohort. This significantly bolsters the generalizability of the AKDN models, offering a solid foundation for their adoption in healthcare environment and facilitating improved patient outcomes through informed decision-making.

We leveraged a retrospective design and utilized secondary data from the MCHP databases to develop a risk prediction model, which saved significant time and financial resources compared to developing a prospective registry for CKD G5 patients undergoing major non-cardiac surgery.[83] Another strength of our study was external validation on Alberta data, which is critical for machine learning models. This was particularly vital because external validation verified model's generalizability and provided evidence about its robustness and suitability for clinical adoption.[20]

Study limitations

A significant limitation of this study was the lack of external validation on non-Canadian populations. Our models were developed using data from Manitoba and then externally validated using data from Alberta; both are Canadian provinces and may not universally represent diverse patient populations. Without validation through external datasets outside Canada, the generalizability of our models in other countries remains uncertain. We did not

account for clustering in our models, however, when we restricted to one surgery per patient, both random forest and XGBoost models performed well.

We were limited by the available data and were unable to include biomarkers (such as Troponin T or BNP/NT-proBNP) in our model. Some of these markers are known to have strong associations with perioperative outcomes[84], but their limited availability in both provinces did not allow us to examine their potential utility in perioperative risk prediction

Next steps

We will seek to externally validate the machine learning model in different populations to verify the model's generalizability. We will consider incorporation into relevant electronic databases to facilitate use in clinical care. We will also consider the development of an online medical calculator such as Calculate by QxMD or the perioperative risk calculator database, to assist healthcare professionals in making patient centered decision.[85,86]

Conclusion

In conclusion, we present two highly accurate and externally validated models to predict MACE or mortality in patients with kidney failure undergoing non-cardiac surgery using population-based routinely collected data. In addition, we also demonstrate the potential benefits of utilizing machine-learning methods in clinical settings through the innovative application of machine learning techniques in developing a new parsimonious model. This study lays the groundwork for future model development for perioperative events using machine learning methods in the general population.

Ethics

Data access approvals have been received from the Provincial Health Research Privacy Committee (PHRPC, P2024-23) and Shared Health. Ethics approvals have been received from the University of Manitoba Research Ethics Board (HS26359 (H2024:083)). This work was of

minimal risk to patients due to the protective privacy measures put in place by the Manitoba Centre for Health Policy and Shared Health's administrative computer encryption. Participants will not require informed consent as health information will be accessed through scrambled PHIN.

The Conjoint Health Research Ethics Board at the University of Calgary and the Health Research Ethics Board at the University of Alberta granted ethics approval and waived the requirement for informed consent, as the study involved population-based data. All ethical guidelines and regulations were adhered to during this study.

Literature Cited

1. Kidney Disease: Improving Global Outcomes (KDIGO). (2012). KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease (Vol. 3, No. 1). KDIGO. https://kdigo.org/wp-content/uploads/2017/02/KDIGO_2012_CKD_GL.pdf
2. Kitzler, T. M., & Chun, J. (2023, January). Understanding the current landscape of kidney disease in Canada to advance precision medicine guided personalized care. Canadian journal of kidney health and disease, 10, 205435812311541. https://doi.org/10.1177/20543581231154185
3. Lin, Q., Mao, L., Shao, L., Zhu, L., Han, Q., Zhu, H., Jin, J., & You, L. (2020, December 15). Global, regional, and national burden of chronic myeloid leukemia, 1990–2017: A systematic analysis for the global burden of disease study 2017. Frontiers in oncology, 10. https://doi.org/10.3389/fonc.2020.580759
4. Chartier, M. J., Tangri, N., Komenda, P., Walld, R., Koseva, I., Burchill, C., McGowan, K. L., & Dart, A. (2018, October 10). Prevalence, socio-demographic characteristics, and comorbid health conditions in pre-dialysis chronic kidney disease: results from the Manitoba chronic kidney disease cohort. BMC nephrology, 19(1). https://doi.org/10.1186/s12882-018-1058-3
5. Meara, J. G., & Greenberg, S. L. (2015, May). The Lancet Commission on Global Surgery Global surgery 2030: Evidence and solutions for achieving health, welfare and economic development. Surgery, 157(5), 834–835. https://doi.org/10.1016/j.surg.2015.02.009
6. Canadian Institute for Health Information. (2023). Surgeries impacted by COVID-19, March 2020 to September 2022 — Data Tables. CIHI. https://www.cihi.ca/sites/default/files/document/surgeries-impacted-by-covid-19-march-2020-to-sept-2022-data-tables-en.xlsx
7. Surgeries impacted by COVID-19: An update on volumes and wait times | CIHI. (n.d.). https://www.cihi.ca/en/surgeries-impacted-by-covid-19-an-update-on-volumes-and-wait-times

8. Nepogodiev, D., Martin, J., Biccard, B., Makupe, A., Bhangu, A., Ademuyiwa, A. O., Adisa, A., Aguilera, M. L., Chakrabortee, S., Fitzgerald, J. E., Ghosh, D., Glasbey, J., Harrison, E. M., Ingabire, J. A., Salem, H., Lapitan, M. C., Lawani, I., Lissauer, D., Magill, L., . . . Morton, D. (2019, February 1). Global burden of post-operative death. The lancet; Elsevier BV. https://doi.org/10.1016/s0140-6736(18)33139-8
9. Palamuthusingam, D., Nadarajah, A., Pascoe, E. M., Craig, J., Johnson, D. W., Hawley, C. M., & Fahim, M. (2020, June 26). Post-operative mortality in patients on chronic dialysis following elective surgery: A systematic review and meta-analysis. PLOS ONE, 15(6), e0234402. https://doi.org/10.1371/journal.pone.0234402
10. Harrison, T. G., Ruzycki, S. M., James, M. T., Ronksley, P. E., Zarnke, K. B., Tonelli, M., Manns, B. J., McCaughey, D., Schneider, P., Dixon, E., Hartley, R. L., Owen, V. S., Ma, Z., & Hemmelgarn, B. R. (2021, March). Estimated GFR and incidence of major surgery: A population-based cohort study. American journal of kidney diseases, 77(3), 365-375.e1. https://doi.org/10.1053/j.ajkd.2020.08.009
11. Palamuthusingam, D., Nadarajah, A., Johnson, D. W., Pascoe, E. M., Hawley, C. M., & Fahim, M. (2021, March 18). Morbidity after elective surgery in patients on chronic dialysis: a systematic review and meta-analysis. BMC nephrology, 22(1). https://doi.org/10.1186/s12882-021-02279-0
12. Harrison, T. G., Hemmelgarn, B. R., James, M. T., Manns, B. J., Tonelli, M., Brindle, M. E., McCaughey, D., Ruzycki, S. M., Zarnke, K. B., Wick, J., & Ronksley, P. E. (2023, January 10). Association of kidney function with major post-operative events after non-cardiac ambulatory surgeries. Annals of surgery, 277(2), e280–e286. https://doi.org/10.1097/sla.0000000000005040
13. Gajdos, C., Hawn, M. T., Kile, D., Robinson, T. N., & Henderson, W. G. (2013, February 1). Risk of major nonemergent inpatient general surgical procedures in patients on long-term dialysis. JAMA surgery, 148(2), 137. https://doi.org/10.1001/2013.jamasurg.347
14. Yasuda, K., Tahara, K., Kume, M., Tsutsui, S., Higashi, H., & Kitano, S. (2007, December 1). Risk factors for morbidity and mortality following abdominal surgery in patients on maintenance hemodialysis. Hepato-gastroenterology, 54(80), 2282-2284.

15. Drolet, S., Maclean, A. R., Myers, R. P., Shaheen, A. A. M., Dixon, E., & Buie, W. D. (2010, November). Morbidity and mortality following colorectal surgery in patients with end-stage renal failure: A population-based study. Diseases of the colon & rectum, 53(11), 1508–1516. https://doi.org/10.1007/dcr.0b013e3181e8fc8e
16. Rigatto, C., Sood, M. M., & Tangri, N. (2012, November). Risk prediction in chronic kidney disease. Current opinion in nephrology and hypertension, 21(6), 612–618. https://doi.org/10.1097/mnh.0b013e328359072f
17. Ramspek, C. L., Jager, K. J., Dekker, F. W., Zoccali, C., & van Diepen, M. (2020, November 24). External validation of prognostic models: what, why, how, when and where? Clinical kidney journal, 14(1), 49–58. https://doi.org/10.1093/ckj/sfaa188
18. Debray, T. P., Vergouwe, Y., Koffijberg, H., Nieboer, D., Steyerberg, E. W., & Moons, K. G. (2015, March). A new framework to enhance the interpretation of external validation studies of clinical prediction models. Journal of clinical epidemiology, 68(3), 279–289. https://doi.org/10.1016/j.jclinepi.2014.06.018
19. Brier GW. Verifcation of forecasts expressed in terms of probability. Mon Weather Rev. 1950;78(1):1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2
20. Steyerberg, E. W., & Harrell, F. E. (2016, January). Prediction models need appropriate internal, internal–external, and external validation. Journal of clinical epidemiology, 69, 245–247. https://doi.org/10.1016/j.jclinepi.2015.04.005
21. Bleeker, S., Moll, H., Steyerberg, E., Donders, A., Derksen-Lubsen, G., Grobbee, D., & Moons, K. (2003, September). External validation is necessary in prediction research: Journal of clinical epidemiology, 56(9), 826–832. https://doi.org/10.1016/s0895-4356(03)00207-5
22. Zhang, X., Zhou, K., You, L., Zhang, J., Chen, Y., Dai, H., Wan, S., Guan, Z., Hu, M., Kang, J., Liu, Y., & Shang, H. (2023). Risk prediction models for mortality and readmission in patients with acute heart failure: A protocol for systematic review, critical appraisal, and meta-analysis. PloS One, 18(7), e0283307. https://doi.org/10.1371/journal.pone.0283307

23. Goodacre, S. (2023). Using clinical risk models to predict outcomes: what are we predicting and why? Emergency Medicine Journal, 40(10), 728–730. https://doi.org/10.1136/emermed-2022-213057
24. Habehh, H., & Gohel, S. (2021, December 16). Machine learning in healthcare. Current genomics, 22(4), 291–300. https://doi.org/10.2174/1389202922666210705124359
25. Wiens, J., & Shenoy, E. S. (2017, August 21). Machine learning for healthcare: On the verge of a major shift in healthcare epidemiology. Clinical infectious diseases, 66(1), 149–153. https://doi.org/10.1093/cid/cix731
26. Lee, C. S., & Lee, A. Y. (2020, June). Clinical applications of continual learning machine learning. The lancet digital health, 2(6), e279–e281. https://doi.org/10.1016/s2589-7500(20)30102-3
27. Shamout, F., Zhu, T., & Clifton, D. A. (2021). Machine learning for clinical outcome prediction. IEEE reviews in biomedical engineering, 14, 116–126. https://doi.org/10.1109/rbme.2020.3007816
28. Jamal, S., Goyal, S., Grover, A., Shanker, A. (2018). Machine learning: What, why, and how?. In: Shanker, A. (eds) Bioinformatics: Sequences, Structures, Phylogeny. Springer, Singapore. https://doi.org/10.1007/978-981-13-1562-6_16
29. Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017, April 4). Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLOS ONE, 12(4), e0174944. https://doi.org/10.1371/journal.pone.0174944
30. Lee, Y. W., Choi, J. W., & Shin, E. H. (2021, February). Machine learning model for predicting malaria using clinical information. Computers in biology and medicine, 129, 104151. https://doi.org/10.1016/j.compbiomed.2020.104151
31. Sone, D., & Beheshti, I. (2021, June 22). Clinical application of machine learning models for brain imaging in epilepsy: A review. Frontiers in neuroscience, 15. https://doi.org/10.3389/fnins.2021.684825
32. Feng, J. Z., Wang, Y., Peng, J., Sun, M. W., Zeng, J., & Jiang, H. (2019, December). Comparison between logistic regression and machine learning algorithms on survival

prediction of traumatic brain injuries. Journal of critical care, 54, 110–116. https://doi.org/10.1016/j.jcrc.2019.08.010

33. Charilaou, P., & Battat, R. (2022, February 7). Machine learning models and over-fitting considerations. World journal of gastroenterology, 28(5), 605–607. https://doi.org/10.3748/wjg.v28.i5.605
34. Chen, P. H. C., Liu, Y., & Peng, L. (2019, April 18). How to develop machine learning models for healthcare. Nature materials, 18(5), 410–414. https://doi.org/10.1038/s41563-019-0345-0
35. Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019, June). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. Journal of clinical epidemiology, 110, 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004
36. Anderson, B., Zou, X., Evison, F., Gallier, S., Inston, N., & Sharif, A. (2022, November 2). Major adverse cardiovascular events and all-cause mortality after emergency general surgery among kidney failure patients. BJS open, 6(6). https://doi.org/10.1093/bjsopen/zrac130
37. Duceppe, E., Parlow, J., MacDonald, P., Lyons, K., McMullen, M., Srinathan, S., Graham, M., Tandon, V., Styles, K., Bessissow, A., Sessler, D. I., Bryson, G., & Devereaux, P. (2017, January). Canadian cardiovascular society guidelines on perioperative cardiac risk assessment and management for patients who undergo non-cardiac surgery. Canadian journal of cardiology, 33(1), 17–32. https://doi.org/10.1016/j.cjca.2016.09.008
38. Prytherch, D. R., Whiteley, M. S., Higgins, B., Weaver, P. C., Prout, W. G., & Powell, S. J. (1998, September). POSSUM and Portsmouth POSSUM for predicting mortality. British journal of surgery, 85(9), 1217–1220. https://doi.org/10.1046/j.1365-2168.1998.00840.x
39. Sutton, R., Bann, S., Brooks, M., & Sarin, S. (2002, June). The surgical risk scale as an improved tool for risk-adjusted analysis in comparative surgical audit. British journal of surgery, 89(6), 763–768. https://doi.org/10.1046/j.1365-2168.2002.02080.x
40. Neri, L., Lonati, C., Titapiccolo, J. I., Nadal, J., Meiselbach, H., Schmid, M., Baerthlein, B., Tschulena, U., Schneider, M. P., Schultheiss, U. T., Barbieri, C., Moore, C., Steppan, S.,

Eckardt, K. U., Stuard, S., & Bellocchio, F. (2022, July 12). The cardiovascular literature-based risk algorithm (CALIBRA): Predicting cardiovascular events in patients with non-dialysis dependent chronic kidney disease. Frontiers in nephrology, 2. https://doi.org/10.3389/fneph.2022.922251

41. Lee, T. H., Marcantonio, E. R., Mangione, C. M., Thomas, E. J., Polanczyk, C. A., Cook, E. F., Sugarbaker, D. J., Donaldson, M. C., Poss, R., Ho, K. K. L., Ludwig, L. E., Pedan, A., & Goldman, L. (1999, September 7). Derivation and prospective validation of a simple index for prediction of cardiac risk of major non-cardiac surgery. Circulation, 100(10), 1043–1049. https://doi.org/10.1161/01.cir.100.10.1043
42. Harrison, T. G., Hemmelgarn, B. R., James, M. T., Sawhney, S., Manns, B. J., Tonelli, M., Ruzycki, S. M., Zarnke, K. B., Wilson, T. A., McCaughey, D., & Ronksley, P. E. (2023, March 10). Prediction of major post-operative events after non-cardiac surgery for people with kidney failure: derivation and internal validation of risk models. BMC nephrology, 24. https://doi.org/10.1186/s12882-023-03093-6
43. Harrison, T. G., Hemmelgarn, B. R., James, M. T., Sawhney, S., Lam, N. N., Ruzycki, S. M., Wilson, T. A., & Ronksley, P. E. (2022, October). Using the revised cardiac risk index to predict major post-operative events for people with kidney failure: An external validation and update. CJC open, 4(10), 905–912. https://doi.org/10.1016/j.cjco.2022.07.008
44. Gupta, P. K., Gupta, H., Sundaram, A., Kaushik, M., Fang, X., Miller, W. J., Esterbrooks, D. J., Hunter, C. B., Pipinos, I. I., Johanning, J. M., Lynch, T. G., Forse, R. A., Mohiuddin, S. M., & Mooss, A. N. (2011, July 26). Development and validation of a risk calculator for prediction of cardiac risk after surgery. Circulation, 124(4), 381–387. https://doi.org/10.1161/circulationaha.110.015701
45. Bilimoria, K. Y., Liu, Y., Paruch, J. L., Zhou, L., Kmiecik, T. E., Ko, C. Y., & Cohen, M. E. (2013, November). Development and evaluation of the universal ACS NSQIP surgical risk calculator: A decision aid and informed consent tool for patients and surgeons. Journal of the American College of Surgeons, 217(5), 833-842e3. https://doi.org/10.1016/j.jamcollsurg.2013.07.385

46. Ramspek, C., Voskamp, P., van Ittersum, F., Krediet, R., Dekker, F., & van Diepen, M. (2017, September). Prediction models for the mortality risk in chronic dialysis patients: a systematic review and independent external validation study. Clinical Epidemiology, Volume 9, 451–464. https://doi.org/10.2147/clep.s139748
47. Martens P, Nickel N, Lix L, Turner D, Prior H, Walld R, Soodeen R, Rajotte L, & Ekuma O. (2015). The cost of smoking: A Manitoba study. Manitoba Centre for Health Policy, Winnipeg, MB.
48. Roos, L. L., Mustard, C. A., Nicol, J. P., Comm, B., McLerran, D. F., Malenka, D. J., Young, T. K., & Cohen, M. M. (1993, March). Registries and administrative data. Medical care, 31(3), 201–212. https://doi.org/10.1097/00005650-199303000-00002
49. Robinson, J. R., Young, T. K., Roos, L. L., & Gelskey, D. E. (1997, September). Estimating the burden of disease. Medical care, 35(9), 932–947. https://doi.org/10.1097/00005650-199709000-00006
50. Levey, A. S., Stevens, L. A., Schmid, C. H., Zhang, Y. L., Castro, A. F., Feldman, H. I., Kusek, J. W., Eggers, P., Van Lente, F., Greene, T., & Coresh, J. (2009, May 5). A new equation to estimate glomerular filtration rate. Annals of internal medicine, 150(9), 604. https://doi.org/10.7326/0003-4819-150-9-200905050-00006
51. Hemmelgarn, B. R., Clement, F., Manns, B. J., Klarenbach, S., James, M. T., Ravani, P., Pannu, N., Ahmed, S. B., MacRae, J., Scott-Douglas, N., Jindal, K., Quinn, R., Culleton, B. F., Wiebe, N., Krause, R., Thorlacius, L., & Tonelli, M. (2009). Overview of the Alberta Kidney Disease Network. BMC Nephrology, 10(1). https://doi.org/10.1186/1471-2369-10-30
52. Canadian Institute for Health Information. Canadian Classification of Health Interventions (CCI): Alphabetical index and tabular list. Ottawa, ON: CIHI; 2022.
53. Siew, E. D., Ikizler, T. A., Matheny, M. E., Shi, Y., Schildcrout, J. S., Danciu, I., Dwyer, J. P., Srichai, M., Hung, A. M., Smith, J. P., & Peterson, J. F. (2012, May). Estimating baseline kidney function in hospitalized patients with impaired kidney function. Clinical journal of the American Society of Nephrology, 7(5), 712–719. https://doi.org/10.2215/cjn.10821011

54. Austin, P. C., Daly, P. A., & Tu, J. V. (2002, August). A multicenter study of the coding accuracy of hospital discharge administrative data for patients admitted to cardiac care units in Ontario. American heart journal, 144(2), 290–296. https://doi.org/10.1067/mhj.2002.123839
55. Tonelli, M., Wiebe, N., Fortin, M., Guthrie, B., Hemmelgarn, B. R., James, M. T., Klarenbach, S. W., Lewanczuk, R., Manns, B. J., Ronksley, P., Sargious, P., Straus, S., & Quan, H. (2015, April 17). Methods for identifying 30 chronic conditions: application to administrative data. BMC medical informatics and decision making, 15(1). https://doi.org/10.1186/s12911-015-0155-5
56. Ye, Y., Larrat, E. P., & Caffrey, A. R. (2017, December 11). Algorithms used to identify ventricular arrhythmias and sudden cardiac death in retrospective studies: a systematic literature review. Therapeutic advances in cardiovascular disease, 12(2), 39–51. https://doi.org/10.1177/1753944717745493
57. Manja, V., AlBashir, S., & Guyatt, G. (2017, February). Criteria for use of composite end points for competing risks—a systematic survey of the literature with recommendations. Journal of clinical epidemiology, 82, 4–11. https://doi.org/10.1016/j.jclinepi.2016.12.001
58. Noordzij, M., Leffondre, K., van Stralen, K. J., Zoccali, C., Dekker, F. W., & Jager, K. J. (2013, August 24). When do we need competing risks methods for survival analysis in nephrology? Nephrology dialysis transplantation, 28(11), 2670–2677. https://doi.org/10.1093/ndt/gft355
59. Saito, T., & Rehmsmeier, M. (2015, March 4). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432
60. Steyerberg, E. W. (2019). Clinical prediction models: A practical approach to development, validation, and updating. Springer Cham. https://doi.org/10.1007/978-3-030-16399-0
61. Steyerberg, E. W., & Lingsma, H. F. (2008, April 10). Validating prediction models. BMJ, 336(7648), 789. https://doi.org/10.1136/bmj.39542.610000.3a

62. Rodseth, R. N., Biccard, B. M., Le Manach, Y., Sessler, D. I., Lurati Buse, G. A., Thabane, L., Schutt, R. C., Bolliger, D., Cagini, L., Cardinale, D., Chong, C. P., Chu, R., Cnotliwy, M., Di Somma, S., Fahrner, R., Lim, W. K., Mahla, E., Manikandan, R., Puma, F., . . . Devereaux, P. (2014, January). The prognostic value of preoperative and post-operative B-type natriuretic peptides in patients undergoing non-cardiac surgery. Journal of the American College of Cardiology, 63(2), 170–180. https://doi.org/10.1016/j.jacc.2013.08.1630
63. Roshanov, P. S., Walsh, M., Devereaux, P. J., MacNeil, S. D., Lam, N. N., Hildebrand, A. M., Acedillo, R. R., Mrkobrada, M., Chow, C. K., Lee, V. W., Thabane, L., & Garg, A. X. (2017, January). External validation of the Revised Cardiac Risk Index and update of its renal variable to predict 30-day risk of major cardiac complications after non-cardiac surgery: rationale and plan for analyses of the VISION study. BMJ Open, 7(1), e013510. https://doi.org/10.1136/bmjopen-2016-013510
64. Vickers, A. J., & Elkin, E. B. (2006, November). Decision curve analysis: A novel method for evaluating prediction models. Medical decision making, 26(6), 565–574. https://doi.org/10.1177/0272989x06295361
65. Vickers, A. J., van Calster, B., & Steyerberg, E. W. (2019, October 4). A simple, step-by-step guide to interpreting decision curve analysis. Diagnostic and prognostic research, 3, 18. https://doi.org/10.1186/s41512-019-0064-7
66. Ranganathan, S., Nakai, K., & Schonbach, C. (Eds.). (2018). Decision trees and random forests. In Encyclopedia of bioinformatics and computational biology: ABC of bioinformatics (pp. 374-383). Elsevier. https://books.google.ca/books?id=rs51DwAAQBAJ
67. Rigatti, S. J. (2017). Random forest. Journal of insurance medicine, 47(1), 31–39. https://doi.org/10.17849/insm-47-01-31-39.1
68. Liu, Y., Wang, Y., & Zhang, J. (2012). New machine learning algorithm: random forest. In Lecture notes in computer science (Vol. 7473, pp. 246–252). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-34062-8_32
69. Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., & Chica-Rivas, M. (2015, December). Machine learning predictive models for mineral prospectivity: An evaluation

of neural networks, random forest, regression trees and support vector machines. Ore geology reviews, 71, 804–818. https://doi.org/10.1016/j.oregeorev.2015.01.001

70. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785
71. Gildenblat, J. (2023). A python library for confidence intervals. GitHub. Retrieved from https://github.com/jacobgil/confidenceinterval
72. Ahmad, G. N., Fatima, H., Ullah, S., Saidi, A. S., & Imdadullah, N. (2022). Efficient medical diagnosis of human heart diseases using machine learning techniques with and without GridSearchCV. IEEE Access, 10, 80151–80173. https://doi.org/10.1109/access.2022.3165792
73. Gordon, A. D., Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and Regression Trees. Biometrics, 40(3), 874. https://doi.org/10.2307/2530946
74. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html
75. Daghistani, T., & Alshammari, R. (2020). Comparison of statistical logistic regression and RandomForest machine learning techniques in predicting diabetes. Journal of advances in information technology, 11. 78. https://doi.org/10.12720/jait.11.2.78-83
76. Sufriyana, H., Husnayain, A., Chen, Y. L., Kuo, C. Y., Singh, O., Yeh, T. Y., Wu, Y. W., & Su, E. C. Y. (2020, November 17). Comparison of multivariable logistic regression and other machine learning algorithms for prognostic prediction studies in pregnancy care: Systematic review and meta-analysis. JMIR Medical Informatics, 8(11), e16503. https://doi.org/10.2196/16503
77. Xi, Y., Wang, H., & Sun, N. (2022, November 14). Machine learning outperforms traditional logistic regression and offers new possibilities for cardiovascular risk prediction: A study involving 143,043 Chinese patients with hypertension. Frontiers in Cardiovascular Medicine, 9. https://doi.org/10.3389/fcvm.2022.1025705

78. Song, X., Liu, X., Liu, F., & Wang, C. (2021, July). Comparison of machine learning and logistic regression models in predicting acute kidney injury: A systematic review and meta-analysis. International Journal of Medical Informatics, 151, 104484. https://doi.org/10.1016/j.ijmedinf.2021.104484
79. S. Tangprasert, R. Sonthana, Y. Nilsiam, S. Nuchitprasitchai and N. Bhumpenpein, "Data-Driven Heart Disease Risk Prediction with Machine Learning on Healthcare Datasets," 2023 Research, Invention, and Innovation Congress: Innovative Electricals and Electronics (RI2C), Bangkok, Thailand, 2023, pp. 220-223. https://ieeexplore.ieee.org/document/10355977
80. Stylianou, N., Akbarov, A., Kontopantelis, E., Buchan, I., & Dunn, K. W. (2015, August). Mortality risk prediction in burn injury: Comparison of logistic regression with machine learning approaches. Burns, 41(5), 925–934. https://doi.org/10.1016/j.burns.2015.03.016
81. Song, Y., Yang, X., Luo, Y., Ouyang, C., Yu, Y., Ma, Y., Li, H., Lou, J., Liu, Y., Chen, Y., Cao, J., & Mi, W. (2022, October 11). Comparison of logistic regression and machine learning methods for predicting postoperative delirium in elderly patients: A retrospective study. CNS Neuroscience & Therapeutics, 29(1), 158–167. https://doi.org/10.1111/cns.13991
82. Faisal, M., Scally, A., Howes, R., Beatson, K., Richardson, D., & Mohammed, M. A. (2018, November 29). A comparison of logistic regression models with alternative machine learning methods to predict the risk of in-hospital mortality in emergency medical admissions via external validation. Health Informatics Journal, 26(1), 34–44. https://doi.org/10.1177/1460458218813600
83. Boslaugh, S. (2015). Using secondary data. In Public Health Research Methods (pp. 253-277). SAGE Publications, Inc. https://doi.org/10.4135/9781483398839
84. Harrison, T. G., Shukalek, C. B., Hemmelgarn, B. R., Zarnke, K. B., Ronksley, P. E., Iragorri, N., Graham, M. M., & James, M. T. (2020). Association of NT-proBNP and BNP With Future Clinical Outcomes in Patients With ESKD: A Systematic Review and Meta-analysis. American Journal of Kidney Diseases, 76(2), 233–247. https://doi.org/10.1053/j.ajkd.2019.12.017

85. QxMD | Moving Research into Practice | QxMD. (2019, July 23). QxMD. https://qxmd.com/
86. Roshanov P. Perioperative Risk Prediction: Evidence-based prediction guides for complications in surgical patients 2019 Available from: https://perioperativerisk.com/

Appendix 1. Surgical Categories by Canadian Classification of Health Intervention (CCI) codes

Surgery Type	Section, Group, Intervention Components Codes based on Canadian Classification of Interventions (CCI)
Arteriovenous fistula	1JM50 1JM51 1JM57 1JM58 1JM76 1JM80 1JM82 1JM87 1KY76LA 1KY76LASJ 1KY76LAXXA 1KY76LAXXL 1KY76LAXXN 1KY80 1KY76
Head and Neck	1DA56LA 1DA58 1DA59HAT9 1DA59JAGX 1DA80 1DA82 1DA83 1DA84 1DA87 1DA89 1DA91 1DE56LA 1DE59JAGX 1DE80 1DE84 1DE86 1DE87 1DE91 1DF53 1DF58 1DF71JA 1DF72LA 1DF80 1DF85 1DF87 1DF89 1DG 1DJ53 1DJ80 1DK59LAGX 1DK80 1DK85 1DK87 1DK91 1DE53 1DL59LAGX 1DL80 1DL87 1DL89 1DL91 1DM53 1DN76QRQB 1DP53 1DR57 1DR59LAKD 1DR59QRAZ 1DR59QRKD 1DR59QRX7 1DR72 1DR80 1DR89 1DR91 1DZ70 1DZ94 1EA56 1EA58 1EA72 1EA73 1EA74 1EA80 1EA87 1EA92 1EB73LA 1EB74 1EB80 1EB87 1EC 1ED56 1ED73 1ED74 1ED79 1ED80 1ED83 1ED87 1ED91 1EE56 1EE58 1EE71 1EE73 1EE78 1EE79 1EE80 1EE83 1EE87 1EE91 1EF73 1EF74 1EF80 1EG 1EH 1EJ87 1EJ89 1EL53 1EL57LA 1EL72DA 1EL72LA 1EL74 1EL80 1EL83 1EM53 1EM73LA 1EM74 1EM80 1EM86 1EM87 1EN53 1EN73LA 1EN74 1EN80 1EN87 1EN91 1EP58 1EP72 1EP80 1EP87 1EQ56LA 1EQ59 1EQ70 1EQ87 1EQ94 1ES58 1ES80 1ES87 1ET56LA 1ET57LA 1ET59LAAD 1ET59LAAG 1ET59LAGX 1ET72LA 1ET73 1ET80 1ET82 1ET86 1ET87 1ET89 1EU87 1EU89 1EV87 1EW80 1EW86 1EW87 1EW91 1EX59 1EX80 1EX87 1EY87 1EY91 1FA53 1FA56LA 1FA84 1FA87 1FA91 1FB53 1FB80 1FB86 1FB87 1FB91 1FC56LA 1FC80LAXXB 1FC80LAXXE 1FC87 1FG56LA 1FG58 1FG59JAGX 1FG72 1FG80 1FG87 1FH56LA 1FH59JAGX 1FH78 1FH80 1FH87 1FJ56LA 1FJ59JAGX 1FJ72 1FJ74 1FJ80 1FJ87 1FJ91 1FK94 1FL51 1FL57 1FL80 1FL87 1FL89 1FM50 1FM51 1FM57 1FM80 1FM83 1FM87 1FM89 1FM91 1FN51 1FN57 1FN59 1FN80 1FN83 1FN87 1FN89 1FQ56LA 1FQ59HAAW 1FQ78 1FQ80 1FQ87 1FQ89 1FR56LA 1FR59JAGX 1FR87 1FR89 1FU71 1FU87 1FU89 1FU91 1FV83 1FV87 1FV89 1FX56LA 1FX80 1FX86 1FX87 1FX91 1GA74 1GA80 1GA83 1GA87 1GA89 1GB 1GC50 1GC59 1GD53 1GD74 1GD83 1GD87 1GD89 1GE50 1GE56LA 1GE80 1GE87 1GE89 1GE91 1GH71 1GH84 1GJ50LA 1GJ50LANR 1GJ53LAPM 1GJ56LA 1GJ77 1GJ80 1GJ82 1GJ85 1GJ86 1GJ87 1GK52LA 1GK59LAGX 1GK74LA 1GK80 1GK83 1GK87 1GK89 1MB87LA 1MC87 1MC89 1MC91 1ML59 1ML87
Intra-abdominal	1MG87 1MG89 1MJ87 1MJ89 1MJ91 1MP50 1MP51 1MP59 1MP76 1MP80 1MP87 1NF80DAXXE 1NF80DAXXN 1NF80LA 1NF80LAXXE 1NF80LAXXN 1NF82 1NF84 1NF86 1NF87 1NF89 1NF90 1NF91 1NF92 1NK53DATS 1NK53LAQB 1NK53LATS 1NK56DA 1NK56LA 1NK58 1NK74 1NK76 1NK77 1NK80 1NK82 1NK84 1NK85 1NK87DA 1NK87DN 1NK87DP 1NK87DX 1NK87DY 1NK87LA 1NK87RE 1NK87RF 1NK87TF 1NK87TG 1NM56DA 1NM56LA 1NM58 1NM74 1NM76 1NM77 1NM80 1NM82 1NM87 1NM89 1NM91 1NP58 1NP72 1NP73LA 1NP85 1NP86 1NV89 1OA53 1OA58 1OA59DAGX 1OA59LAGX 1OA74 1OA85 1OA87 1OB59DAGX 1OB59LAGX 1OB74 1OB83 1OB85 1OB87 1OB89 1OD57 1OD76 1OD80 1OD86 1OD89 1OE57DAAG 1OE57DAAM 1OE57DAAS

	1OE57DAAZ 1OE57DABD 1OE57DAGX 1OE57HAAG 1OE57HAAM 1OE57HAAS 1OE57HAAZ 1OE57HABD 1OE57HAGX 1OE57LAAG 1OE57LAAM 1OE57LAAS 1OE57LAAZ 1OE57LABD 1OE57LAGX 1OE59KQ 1OE76 1OE80 1OE84 1OE86 1OE87 1OE89 1OJ53 1OJ56 1OJ76 1OJ83 1OJ85 1OJ87 1OJ89 1OK58 1OK85 1OK87 1OK89 1OK91 1OT56 1OT58 1OT70 1OT72 1OT80 1OT87 1OT91 1OW12 1OW80DA 1OW80LA 1OW87 1OW89 1OZ94LA
Kidney Transplant	1PC85
Musculoskeletal (MSK)	1VZ94LA 1VZ70LA 1VS80 1VS72 1VS58 1VR80 1VR72 1VR58 1VR57 1VQ93 1VQ91 1VQ87 1VQ83 1VQ82 1VQ80 1VQ79 1VQ74 1VQ73LA 1VQ58 1VP89 1VP87 1VP80 1VP74 1VP73LA 1VP72 1VP53 1VN 1VM 1VL 1VK89 1VK87 1VK80 1VG93 1VG87 1VG83 1VG80 1VG75 1VG74 1VG73LA 1VG72LA 1VG72DA 1VG58DA 1VG57 1VG53 1VE80 1VE72 1VE58 1VD87 1VD80 1VD72 1VD58 1VD57 1VC93 1VC91 1VC87 1VC83 1VC82 1VC80 1VC79 1VC74 1VC73LA 1VC58 1VA93 1VA87 1VA83 1VA80 1VA75 1VA74 1VA73LA 1VA72LA 1VA72DA 1VA58DA 1VA57 1VA53 1UV80 1UV72 1UV58 1UU84 1UU80 1UU72 1UU53 1UT84 1UT80 1UT72LA 1UT53 1US80 1US72 1US58 1UK93 1UK87 1UK80 1UK75 1UK74 1UK73LA 1UK72LA 1UK53 1UJ93 1UJ87 1UJ82 1UJ80 1UJ75 1UJ74 1UJ73LA 1UJ71 1UJ58 1UG93 1UG87 1UG80 1UG75 1UG74 1UG73LA 1UG72LA 1UG72JAHB 1UG72JAAZ 1UG57 1UG53 1UF93 1UF87 1UF84 1UF82 1UF80 1UF79 1UF74 1UF73LA 1UC89 1UC87 1UC82 1UC80 1UC79 1C75 1UC74 1UC73LA 1UC72 1UC57 1UC53 1UB93 1UB87 1UB83 1UB80 1UB75 1UB74 1UB73LA 1UB72LA 1UB72DA 1UB58 1UB57 1UB53 1TZ94LA 1TZ70LA 1TV93 1TV91 1TV87 1TV84 1TV83 1TV82 1TV80 1TV79 1TV74 1TV73LA 1TV58 1TS80 1TS72 1TS58 1TQ80 1TQ72 1TQ58 1TQ57 1TM93 1TM87 1TM83 1TM80 1TM75 1TM74 1TM73LA 1TM72LA 1TM72DA 1TM58 1TM57 1TM53 1TK93 1TK91 1TK87 1TK83 1TK82 1TK80 1TK79 1TK74 1TK73LA 1TK58 1TH 1TF80 1TF72 1TF58 1TF57 1TC80 1TC72 1TC57 1TB87 1TB80 1TB74 1TB72LA 1TB72JAHB 1TB72JAAZ 1TB72DA 1TA93 1TA87 1TA83 1TA80 1TA75 1TA74 1TA73LA 1TA72LA 1TA72DA 1TA58 1TA57 1TA53 1SY87 1SY84 1SY80 1SY72 1SY58 1SY57 1SY53 1SW87 1SW74 1SQ93 1SQ91 1SQ87 1SQ83 1SQ80 1SQ74 1SQ58 1SQ53 1SN93 1SN87 1SN75 1SN74 1SN72 1SN58 1SM87 1SM80 1SM74 1SM73LA 1SL91 1SL89 1SL87 1SL80 1SL74 1SL73LA 1SL58 1SK87 1SK80 1SK74 1SK73LA 1SG87 1SG80 1SG72WK 1SG72WJ 1SG58 1SF89 1SF87 1SF80 1SF75 1SF74 1SF73PF 1SE89 1SE87PF 1SE59LAGX 1SE53 1SC89 1SC87 1SC80 1SC75 1SC74 1SC72JAHB 1SC72JAAZ 1SA89 1SA80 1SA75 1SA74
Neurosurgery	1AA53SZPL 1AA80 1AA87 1AB86 1AB87 1AC50 1AC52 1AC53 1AC59 1AC87 1AE53 1AE59 1AE85 1AE87 1AF59 1AF87 1AG87 1AJ53 1AJ87 1AN53 1AN56 1AN59 1AN70 1AN73 1AN87 1AP52 1AP53 1AP59 1AP72 1AP87 1AW59 1AW72 1AW87 1AX52MESJ 1AX52MQSJ 1AX53 1AX56 1AX73 1AX80 1AX86 1AX87 1AZ94 1BA53SZDV 1BA59 1BA72 1BA80 1BA87 1BB53 1BB58 1BB59 1BB72 1BB80 1BB87 1BD58 1BD59 1BD72 1BD80 1BD87 1BF 1BG 1BJ53 1BJ59 1BK59 1BM 1BN 1BP 1BQ 1BS58 1BS59 1BS72 1BS80 1BS87 1BT 1BX53 1BX59

	1BX72 1BX80 1BX87 1BZ 1JW50 1JW51 1JW57 1JW59 1JW76 1JW89 1JW86 1JW87
Peritoneal Dialysis Catheter	1OT53 1OT54 1YY53LATS 1SY55
Skin and Soft Tissue	1MD87 1MD89 1MK87 1MK89 1MR50 1MR51 1MR52HA 1MR59 1MR76 1MR80 1MR87 1MR91 1MS50 1MS51 1MS52 1MS59 1MS76 1MS80 1MS87 1MS91 1SH56LA 1SH59 1SH87 1SZ56LA 1SZ59 1SZ87 1TX56LA 1TX59LA 1TX59LAGX 1TX87 1UY56LA 1UY57 1UY59 1UY72 1UY80 1UY87 1VX56LA 1VX59LAGX 1VX87 1WV56LA 1WV57 1WV58 1WV59LAGX 1WV72 1WV80 1WV87 1YA14JAXXK 1YA14JAXXL 1YA14JAXXN 1YA14JAXXP 1YA53 1YA58 1YA59JAGX 1YA59JALV 1YA80 1YA83 1YA87 1YB14JAXXK 1YB14JAXXL 1YB14JAXXN 1YB14JAXXP 1YB53 1YB58 1YB59JAGX 1YB74 1YB80LAXXA 1YB80LAXXB 1YB80LAXXE 1YB80LAXXF 1YB87 1YC14JAXXK 1YC14JAXXL 1YC14JAXXN 1YC14JAXXP 1YC58LAXXA 1YC59JAGX 1YC80LAXXA 1YC80LAXXB 1YC80LAXXE 1YC87 1YD14JAXXK 1YD14JAXXL 1YD14JAXXN 1YD14JAXXP 1YD59JAGX 1YD80LAXXA 1YD80LAXXB 1YD80LAXXE 1YD80LAXXF 1YD87 1YE14JAXXK 1YE14JAXXL 1YE14JAXXN 1YE14JAXXP 1YE59JAGX 1YE74 1YE78 1YE79LAXXA 1YE79LAXXL 1YE79LAXXN 1YE80LAXXA 1YE80LAXXB 1YE80LAXXE 1YE87 1YF14JAXXK 1YF14JAXXL 1YF14JAXXN 1YF14JAXXP 1YF53 1YF57JACF 1YF58 1YF59JAGX 1YF74 1YF80LAXXA 1YF80LAXXB 1YF80LAXXE 1YF80LAXXF 1YF87 1YG14JAXXK 1YG14JAXXL 1YG14JAXXN 1YG14JAXXP 1YG53 1YG58 1YG59JAGX 1YG74 1YG80LAXXA 1YG80LAXXB 1YG80LAXXE 1YG80LAXXF 1YG87 1YR14JAXXK 1YR14JAXXL 1YR14JAXXN 1YR14JAXXP 1YR59JAGX 1YR78LA 1YR78LAAZ 1YR80LAXXA 1YR80LAXXB 1YR80LAXXE 1YR80LAXXF 1YR87 1YS14JAXXK 1YS14JAXXL 1YS14JAXXN 1YS14JAXXP 1YS53 1YS58 1YS59JAGX 1YS74 1YS78LA 1YS78LAAZ 1YS80 1YS87 1YT14JAXXK 1YT14JAXXL 1YT14JAXXN 1YT14JAXXP 1YT53 1YT58 1YT59JAGX 1YT78HA 1YT78LA 1YT78LAAZ 1YT80 1YT87 1YU14JAXXK 1YU14JAXXL 1YU14JAXXN 1YU14JAXXP 1YU53 1YU58 1YU59JAGX 1YU80LAXXA 1YU80LAXXB 1YU80LAXXE 1YU80LAXXF 1YU80LAXXG 1YU87 1YV14JAXXK 1YV14JAXXL 1YV14JAXXN 1YV14JAXXP 1YV53 1YV58 1YV59JAGX 1YV74 1YV78LA 1YV78LAAZ 1YV80LAXXA 1YV80LAXXB 1YV80LAXXE 1YV80LAXXF 1YV87 1YW14JAXXK 1YW14JAXXL 1YW14JAXXN 1YW14JAXXP 1YW53 1YW58 1YW59JAGX 1YW80LAXXA 1YW80LAXXB 1YW80LAXXE 1YW80LAXXF 1YW87 1YX80 1YX87 1YX89 1YY14JAXXK 1YY14JAXXL 1YY14JAXXN 1YY14JAXXP 1YY80 1YY87 1YY87 1YY89 1YZ14JAXXK 1YZ14JAXXL 1YZ14JAXXN 1YZ14JAXXP 1YZ53 1YZ58 1YZ59JAGX 1YZ59JALV 1YZ78LA 1YZ78LAAZ 1YZ80LAXXA 1YZ80LAXXB 1YZ80LAXXE 1YZ80LAXXF 1YZ87 1YZ94 1YZ94LA
Vascular (excluding fistula procedures)	1ID57 1ID76 1ID80 1ID82 1ID86 1ID87 1JD53 1JD59 1JD89 1JE50 1JE51 1JE57 1JE58 1JE59 1JE76 1JE80 1JE87 1JJ50 1JJ51 1JJ57 1JJ58 1JJ76 1JJ80 1JJ83 1JJ87

	1JK50 1JK51 1JK57 1JK58 1JK76 1JK80 1JK87 1JL50 1JL51 1JL57 1JL58 1JL80 1JL87 1JQ50 1JQ51 1JQ57 1JQ80 1JQ87 1JT50 1JT51 1JT57 1JT58 1JT80 1JT87 1JU 1JY50 1JY51 1JY57 1JY76 1JY80 1JY87 1KA50 1KA53 1KA57 1KA58 1KA76 1KA80 1KA82 1KA87 1KE50 1KE51 1KE57 1KE58 1KE76 1KE80 1KE87 1KG50 1KG51 1KG57 1KG58 1KG76 1KG80 1KG82 1KG87 1KQ 1KR34 1KR50 1KR51 1KR53 1KR57 1KR58 1KR59 1KR76 1KR78 1KR80 1KR83 1KR87 1KT50 1KT51 1KT58 1KT76 1KT80 1KT82 1KT87 1KV53 1KV80 1KX80 1KZ 1KY* *IKY with the exception of arteriovenous fistula creation (in separate category)
Low-risk (composite of Ophthalmologic, breast, thoracic, retroperitoneal, lower urologic and gynecologic, and anorectal)	*Ophthalmologic:* 1CC58 1CC59JAGX 1CC80LAXXA 1CC80LAXXK 1CC80LAXXQ 1CC84 1CC85 1CC87 1CD52 1CD53 1CD80LAXXA 1CD80LAXXK 1CD87 1CE56LA 1CE56LALZ 1CF59LAGX 1CF87 1CF91 1CG59 1CG71 1CG76 1CG87 1CH59LAGX 1CH72 1CH80 1CH87 1CJ56 1CJ59LAGX 1CJ80 1CJ87 1CL53 1CL56 1CL59 1CL87 1CL89 1CM59LA 1CM89 1CN59LAGX 1CN59LAGY 1CN72 1CP53 1CP56 1CP80 1CP87 1CP89 1CP91 1CQ59LAGX 1CQ72 1CQ78 1CQ80 1CQ83 1CQ87 1CR80 1CR87 1CS56LA 1CS56LALZ 1CS59JAGX 1CS72 1CS80 1CS84 1CS87 1CT51 1CT59 1CT80 1CT87 1CT89 1CU50 1CU51 1CU56 1CU57 1CU59LAGX 1CU72 1CU76LA 1CU76LANR 1CU76ML 1CU76MLNR 1CU80 1CU87 1CU89 1CV 1CX56LA 1CX59JAGX 1CX72 1CX74 1CX78 1CX80 1CX84 1CX87 1CX88 1CZ56LA 1CZ70 1CZ94HA 1CZ94LA *Breast:* 1YK50 1YK58 1YK80 1YK83 1YK84 1YK87 1YK89 1YK90 1YL87 1YL89 1YM58 1YM74 1YM78 1YM79 1YM80 1YM87 1YM88 1YM89 1YM90 1YM91 1YM92 *Thoracic* 1GM50LA 1GM56LA 1GM80DAXXE 1GM80LA 1GM80LAXXE 1GM80LAXXG 1GM86 1GM87 1GN 1GR 1GT56 1GT58 1GT59 1GT78 1GT80 1GT85 1GT87 1GT89 1GT91 1GV56 1GV59DAGX 1GV59DAZ9 1GV59LAGX 1GV76 1GV80 1GV87 1GV89 1GW56 1GW59 1GW87 1GX78 1GX80 1GX86 1GX87 1GY56 1GY70 1GY72 1GY86 1GY94DA 1GY94LA 1ME87 1ME89 1MF87 1MM 1MN50 1MN51 1MN59 1MN74 1MN76 1MN77 1MN80 1MN87 1NA56DB 1NA56DBXXF 1NA56DBXXG 1NA56EZ 1NA56EZXXG 1NA56FA 1NA56FAXXF 1NA56FAXXG 1NA56LB 1NA56LBXXF 1NA56LBXXG 1NA56LP 1NA56LPXXF 1NA56LPXXG 1NA56QB 1NA56QBXXF 1NA56QBXXG 1NA56QFXXF 1NA56QFXXG 1NA72 1NA74 1NA76 1NA77 1NA80 1NA82 1NA84 1NA86 1NA87 1NA88 1NA89 1NA90 1NA91 1NA92 *Retroperitoneal:* 1PB87 1PB89 1PC51LALV 1PC56 1PC59DAGX 1PC59LAGX 1PC80 1PC82 1PC83 1PC86 1PE50DABD 1PE50DABF 1PE50DABJ 1PE56DA 1PE56LA 1PE59LAAG 1PE59LAGX 1PE76 1PE77 1PE80 1PE82 1PE87 1PE89 1PG50DABD 1PG50DABF 1PG50DABJ 1PG50LABJ 1PG52DA 1PG56DA 1PG56LA 1PG57DAGX 1PG57LAAM 1PG57LAGX 1PG59DAAG 1PG59DAAS 1PG59DAAT 1PG59DAAZ 1PG59DAGX 1PG59KQAP 1PG59KQAQ 1PG59KQAR 1PG59LAAG 1PG59LAGX

	1PG72 1PG74 1PG76 1PG77 1PG80DA 1PG80LA 1PG80LAXXE 1PG80LD 1PG82 1PG86 1PG87 1PG98 1PL50LAGX 1PL53 1PL59LAAD 1PL59LAAG 1PL59LAAS 1PL59LAAZ 1PL59LAGX 1PL72LA 1PL72LAAG 1PL74 1PL80 1PL87 1PM56DA 1PM56LA 1PM57DAGX 1PM57LAGX 1PM58LA 1PM59DAAG 1PM59DAAS 1PM59DAAT 1PM59DAAZ 1PM59DAGX 1PM59DAX7 1PM59KQAP 1PM59KQAQ 1PM59KQAR 1PM72 1PM77 1PM79 1PM80AF 1PM80FJ 1PM80LA 1PM82 1PM84 1PM86 1PM87LA 1PM89LA 1PM90 1PM91 1PM92 *Lower Urologic and gynecologic:* 1MH87 1MH89 1PQ53LAPZ 1PQ56DA 1PQ56LA 1PQ56QY 1PQ57LAAM 1PQ57LAGX 1PQ59FLAD 1PQ59FLAG 1PQ59FLAS 1PQ59FLAZ 1PQ59FLGX 1PQ59LAAZ 1PQ59LAGX 1PQ72 1PQ77 1PQ78 1PQ80 1PQ82 1PQ86 1PQ87 1PQ89 1PV57LAGX 1PV59LAGX 1PV80 1PZ94DA 1PZ94HA 1PZ94LA 1QD72 1QD89 1QE14JAXXK 1QE14JAXXL 1QE14JAXXN 1QE14JAXXP 1QE53 1QE56 1QE58 1QE59LAAD 1QE59LAAG 1QE59LAGX 1QE59LAX7 1QE72 1QE76 1QE80 1QE82 1QE84 1QE87 1QE89 1QG14JAXXK 1QG14JAXXL 1QG14JAXXN 1QG14JAXXP 1QG53 1QG56 1QG59LAAD 1QG59LAAG 1QG59LAGX 1QG59LAX7 1QG78 1QG80 1QG87 1QG89 1QH80 1QH87 1QJ53 1QJ58 1QJ80 1QJ87 1QJ89LA 1QM56 1QM58 1QM74 1QM80 1QM87 1QM89 1QM91 1QN 1QP51 1QP52 1QP72 1QP73 1QP87 1QQ87 1QQ89 1QT87PB 1QT87PK 1QT87PNGX 1QT87QZ 1QT87QZAG 1QT91 1QZ89 1QZ94DA 1QZ94HA 1QZ94LA 1RB56LA 1RB57DA 1RB57LA 1RB58 1RB59 1RB74 1RB80 1RB83 1RB85 1RB87 1RB89 1RD72 1RD89 1RF50DABJ 1RF50DAGX 1RF50DAKR 1RF50DANR 1RF50LABJ 1RF50LAGX 1RF50LAKR 1RF50LANR 1RF51 1RF56 1RF59DAAG 1RF59DAGX 1RF59LAAG 1RF59LAGX 1RF72 1RF74 1RF80 1RF87 1RF89 1RM56DA 1RM56LA 1RM59DAGX 1RM59LAGX 1RM72CAGX 1RM72DAGX 1RM72LAGX 1RM74 1RM80DA 1RM80LA 1RM80LAXXA 1RM80LAXXE 1RM80LAXXF 1RM80LAXXN 1RM80LAXXQ 1RM87DAAG 1RM87DAAK 1RM87DAGX 1RM87LAAK 1RM87LAGX 1RM89 1RM91 1RN74LA 1RN80DA 1RN80LA 1RN80LAFA 1RN80LAXXA 1RN80LAXXE 1RN89LA 1RN89LANRA 1RN89LANRE 1RN89LAXXA 1RN89LAXXE 1RN89LAXXQ 1RS59DAGX 1RS74 1RS80CAXXA 1RS80CAXXB 1RS80CAXXE 1RS80CAXXG 1RS80CAXXN 1RS80CAXXQ 1RS80DA 1RS80LA 1RS80LAXXA 1RS80LAXXB 1RS80LAXXE 1RS80LAXXG 1RS80LAXXN 1RS80LAXXQ 1RS84 1RS86 1RS87 1RS89 1RW56LA 1RW59 1RW72 1RW80 1RW84 1RW87 1RW88 1RW91 1RW92 1RY14JAXXK 1RY14JAXXN 1RY14JAXXP 1RY56LA 1RY59JAGX 1RY80 1RY87 1RZ94DA 1RZ94LA *Anorectal:* 1NQ56DA 1NQ56LA 1NQ59DAAD 1NQ59DAAG 1NQ59DAGX 1NQ59HAX7 1NQ59LAAD 1NQ59LAAG 1NQ59LAGX 1NQ72DA 1NQ72LA 1NQ72PB 1NQ74DW 1NQ74ED 1NQ74EJ 1NQ74PC 1NQ74PD 1NQ74PE 1NQ74SS 1NQ74TV 1NQ74VT 1NQ80 1NQ84 1NQ86 1NQ87CA 1NQ87DA 1NQ87DF 1NQ87DX 1NQ87LA 1NQ87PB 1NQ87PF 1NQ87RD 1NQ87TF 1NQ89 1NQ90 1NT53LADV 1NT53LAPM 1NT56LA 1NT72 1NT80 1NT84 1NT86 1NT87

Appendix 2. Data elements

Data Element	Data Element Source	Data element Description
Demographic Variables		
Age	Manitoba Health Insurance Registry	Continuous in years, calculated as surgery date minus date of birth**.**
Sex	Manitoba Health Insurance Registry	Dichotomous Male and Female
Surgical Variables		
Category of surgery	Hospitalization Discharge Abstracts Database (CCI codes)	Categorized into 11 surgical groups per CCI codes as outlined in Appendix 1.
Surgery setting	Hospitalization Discharge Abstracts Database	Categorized into three categories: Ambulatory surgery based on surgery location/no planned admission; Major elective based on planned hospital admission with inpatient surgery; Major Urgent/Emergent based on unplanned hospital admission with inpatient surgery.
Kidney Failure Variables		
Kidney Failure Type	Medical Claims Shared Health Diagnostic Services	Hemodialysis: at least 2 of the tariff codes on different dates (9798, 9799, 3803, 3800, 3801, 3804, 3790, 9801, 9802, 3792, 9814, 9820) Peritoneal dialysis: at least 2 of the tariff codes on different dates (9805, 9807, 3793, 3805, 3807, 9806, 9819, 9610, 3794, 9821, 3806) CKD G5-ND: laboratory data
Comorbidities (defined with unrestricted lookback, or 5 years for cancer)		
Cancer (any)	Hospitalization Discharge Abstracts Database (ICD-10-CA) and Medical claims (Enhanced ICD-9-CM)	ICD-10-CA: C00.x–C26.x, C30.x–C34.x, C37.x– C41.x, C43.x, C45.x–C58.x, C60.x–C76.x, C81.x–C85.x, C88.x, C90.x–C97.x; ICD-9-CM: 140.x–172.x, 174.x–195.8, 200.x–208.x, 238.6 ICD-10-CA: C77.x-C80.x;

		ICD-9-CM: 196.x-199.x
Cerebrovascular disease	Hospitalization Discharge Abstracts Database (ICD-10-CA) and Medical claims (Enhanced ICD-9-CM)	ICD-10-CA: G45.x, G46.x, H34.0, 160.x-169.x; ICD-9-CM: 362.34, 430.x-438.x
Chronic Pulmonary Disease	Hospitalization Discharge Abstracts Database (ICD-10-CA) and Medical claims (Enhanced ICD-9-CM)	ICD-10-CA: I27.8, I27.9, J40.x–J47.x, J60.x–J67.x, J68.4, J70.1, J70.3; ICD-9-CM: 416.8, 416.9, 490.x–505.x, 506.4, 508.1, 508.8
Dementia	Hospitalization Discharge Abstracts Database (ICD-10-CA) and Medical claims (Enhanced ICD-9-CM)	ICD-10-CA: F00.x–F03.x, F05.1, G30.x, G31.1 290.x; ICD-9-CM: 294.1, 331.2
Diabetes	Hospitalization Discharge Abstracts Database (ICD-10-CA) and Medical claims (Enhanced ICD-9-CM)	ICD-10-CA: E10-14; ICD-9-CM: 250
Heart Failure	Hospitalization Discharge Abstracts Database (ICD-10-CA) and Medical claims (Enhanced ICD-9-CM)	ICD-10-CA: I09.9, I11.0, I13.0, I13.2, I25.5, I42.0, I42.5-I42.9, I43.x, I50.x, P29.0; ICD-9-CM: 398.91, 402.01, 402.11, 402.91, 404.01, 404.03, 404.11, 404.13, 404.91, 404.93, 425.4–425.9, 428.x
History of Myocardial Infarction	Hospitalization Discharge Abstracts Database (ICD-10-CA) and Medical claims (Enhanced ICD-9-CM)	ICD-10-CA: I21.x, I22.x, I25.2; ICD-9-CM:410.x, 412.x
Hypertension	Hospitalization Discharge Abstracts Database (ICD-10-CA) and Medical claims (Enhanced ICD-9-CM)	ICD-10-CA: I10-I13, I15 ICD-9-CM: 401-405
Liver disease (mild and moderate/severe)	Hospitalization Discharge Abstracts Database (ICD-10-CA) and Medical claims (Enhanced ICD-9-CM)	ICD-10-CA: B18.x, K70.0–K70.3, K70.9, K71.3–K71.5, K71.7, K73.x, K74.x, K76.0, K76.2–K76.4, K76.8, K76.9, Z94.4; ICD-9-CM: 070.22, 070.23, 070.32, 070.33, 070.44, 070.54, 070.6, 070.9, 570.x, 571.x, 573.3, 573.4, 573.8, 573.9, V42.7 ICD-10-CA: I85.0, I85.9, I86.4, I98.2, K70.4, K71.1, K72.1, K72.9, K76.5, K76.6,

		K76.7; ICD-9-CM: 456.0–456.2, 572.2–572.8
Peripheral Vascular disease	Hospitalization Discharge Abstracts Database (ICD-10-CA) and Medical claims (Enhanced ICD-9-CM)	ICD-10-CA: I70.x, I71.x, I73.1, I73.8, I73.9, I77.1, I79.0, I79.2, K55.1, K55.8, K55.9, Z95.8, Z95.9; ICD-9-CM: 093.0, 437.3, 440.x, 441.x, 443.1–443.9, 47.1, 557.1, 557.9, V43.4
Laboratory investigations		
Hemoglobin (g/L)	Shared Health Diagnostic Services	Most recent preoperative outpatient hemoglobin that was drawn prior to the procedure, i.e. date of hemoglobin measure minus admission date is ≥1 day, from outpatient lab, and within one year before the procedure itself.
Albumin (g/L)	Shared Health Diagnostic Services	Most recent preoperative outpatient albumin drawn prior to the procedure, i.e. date of albumin measure minus admission date is ≥1 day, from outpatient lab, and within one year before the procedure itself.